Genre-Based Paragraph Classification for Sentiment Analysis

Maite Taboada
Department of Linguistics
Simon Fraser University
Burnaby, BC, Canada
mtaboada@sfu.ca

Julian Brooke
Department of Computer Science
University of Toronto
Toronto, ON, Canada
jbrooke@cs.toronto.edu

Manfred Stede
Institute of Linguistics
University of Potsdam
Potsdam, Germany
stede@ling.uni-potsdam.de

Abstract

We present a taxonomy and classification system for distinguishing between different types of paragraphs in movie reviews: formal vs. functional paragraphs and, within the latter, between description and comment. The classification is used for sentiment extraction, achieving improvement over a baseline without paragraph classification.

1 Introduction

Much of the recent explosion in sentiment-related research has focused on finding low-level features that will help predict the polarity of a phrase, sentence or text. Features, widely understood, may be individual words that tend to express sentiment, or other features that indicate not only sentiment, but also polarity. The two main approaches to sentiment extraction, the semantic or lexicon-based, and the machine learning or corpus-based approach, both attempt to identify low-level features that convey opinion. In the semantic approach, the features are lists of words and their prior polarity, (e.g., the adjective *terrible* will have a negative polarity, and maybe intensity, represented as -4; the noun *masterpiece* may be a 5). Our approach is lexicon-based, but we make use of information derived from machine learning classifiers.

Beyond the prior polarity of a word, its local context obviously plays an important role in conveying sentiment. Polanyi and Zaenen (2006) use the term 'contextual valence shifters' to refer to expressions in the local context that may change a word's polarity, such as intensifiers, modal verbs, connectives, and of course negation. Further beyond the local context, the overall structure and organization of the text, influenced by its genre, can help the reader determine how the evaluation is expressed, and where it lies. Polanyi and Zaenen (2006) also cite genre constraints as relevant factors in calculating sentiment.

Among the many definitions of genre, we take the view of Systemic Functional Linguistics that genres are purposeful activities that develop in stages, or parts (Eggins and Martin, 1997), which can be identified by lexicogrammatical properties (Eggins and Slade, 1997). Our proposal is that, once we have identified different stages in a text, the stages can be factored in the calculation of sentiment, by weighing more heavily those that are more likely to contain evaluation, an approach also pursued in automatic summarization (Seki et al., 2006).

To test this hypothesis, we created a taxonomy of stages specific to the genre of movie reviews, and annotated a set of texts. We then trained various classifiers to differentiate the stages. Having identified the stages, we lowered the weight of those that contained mostly description. Our results show that we can achieve improvement over a baseline when classifying the polarity of texts, even with a classifier that can stand to improve (at 71.1% accuracy). The best performance comes from weights derived from the output of a linear regression classifier.

We first describe our inventory of stages and the manual annotation (Section 2), and in Section 3 turn to automatic stage classification. After describing our approach to sentiment classification of texts in Section 4, we describe experiments to improve its performance with the information on stages in Section 5. Section 6 dis-

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 62–70,
Queen Mary University of London, September 2009.

cusses related work, and Section 7 provides conclusions.

2 Stages in movie reviews

Within the larger *review* genre, we focus on movie reviews. Movie reviews are particularly difficult to classify (Turney, 2002), because large portions of the review contain description of the plot, the characters, actors, director, etc., or background information about the film.

Our approach is based on the work of Bieler et al. (2007), who identify formal and functional zones (stages) within German movie reviews. Formal zones are parts of the text that contribute factual information about the cast and the credits, and also about the review itself (author, date of publication and the reviewer's rating of the movie). Functional zones contain the main gist of the review, and can be divided roughly into *description* and *comment*. Bieler et al. showed that functional zones could be identified using 5-gram SVM classifiers built from an annotated German corpus.

2.1 Taxonomy

In addition to the basic Describe/Comment distinction in Bieler et al., we use a Describe+Comment label, as in our data it is often the case that both description and comment are present in the same paragraph. We decided that a paragraph could be labeled as Describe+Comment when it contained at least a clause of each, and when the comment part could be assigned a polarity (i.e., it was not only subjective, but also clearly positive or negative).

Each of the three high-level tags has a subtag, a feature also present in Bieler et al.'s manual annotation. The five subtags are: overall, plot, actors/characters, specific and general. 'Specific' refers to one particular aspect of the movie (not plot or characters), whereas 'general' refers to multiple topics in the same stage (special effects and cinematography at the same time). Outside the Comment/Describe scale, we also include tags such as Background (discussion of other movies or events outside the movie being reviewed), Interpretation (subjective but not opinionated or polar), and Quotes. Altogether, the annotation system includes 40 tags, with 22 formal and 18 functional zones. Full lists of zone/stage labels are provided in Appendix A.

2.2 Manual annotation

We collected 100 texts from rottentomatoes.com, trying to include one positive and one negative review for the same movie. The reviews are part of the "Top Critics" section of the site, all of them published in newspapers or on-line magazines. We restricted the texts to "Top Critics" because we wanted well-structured, polished texts, unlike those found in some on-line review sites. Future work will address those more informal reviews.

The 100 reviews contain 83,275 words and 1,542 paragraphs. The annotation was performed at the paragraph level. Although stages may span across paragraphs, and paragraphs may contain more than one stage, there is a close relationship between paragraphs and stages. The restriction also resulted in a more reliable annotation, performed with the PALinkA annotation tool (Orasan, 2003).

The annotation was performed by one of the authors, and we carried out reliability tests with two other annotators, one another one of the authors, who helped develop the taxonomy, and the third one a project member who read the annotation guidelines[1], and received a few hours' training in the labels and software. We used Fleiss' kappa (Fleiss, 1971), which extends easily to the case of multiple raters (Di Eugenio and Glass, 2004). We all annotated four texts. The results of the reliability tests show a reasonable agreement level for the distinction between formal and functional zones (.84 for the 3-rater kappa). The lowest reliability was for the 3-way distinction in the functional zones (.68 for the first two raters, and .54 for the three raters). The full kappa values for all the distinctions are provided in Appendix B. After the reliability test, one of the authors performed the full annotation for all 100 texts. Table 1 shows the breakdown of high-level stages for the 100 texts.

Stage	Count
Describe	347
Comment	237
Describe+Comment	237
Background	51
Interpretation	22
Quote	2
Formal	646

Table 1. Stages in 100 text RT corpus

[1]Available from http://www.sfu.ca/~mtaboada/nserc-project.html

3 Classifying stages

Our first classification task aims at distinguishing the two main types of functional zones, Comment and Describe, vs. Formal zones.

3.1 Features

We test two different sets of features. The first, following Bieler et al. (2007), consists of 5-grams (including unigrams, bigrams, 3-grams and 4-grams), although we note in our case that there was essentially no performance benefit beyond 3-grams. We limited the size of our feature set to n-grams that appeared at least 4 times in our training corpus. For the 2 class task (no formal zones), this resulted in 8,092 binary features, and for the 3 and 4 class task there were 9,357 binary n-gram features.

The second set of features captures different aspects of genre and evaluation, and can in turn be divided into four different types, according to source. With two exceptions (features indicating whether a paragraph was the first or last paragraph in text), the features were numerical (frequency) and normalized to the length of the paragraph.

The first group of genre features comes from Biber (1988), who attempted to characterize dimensions of genre. The features here include frequency of first, second and third person pronouns; demonstrative pronouns; place and time adverbials; intensifiers; and modals, among a number of others.

The second category of genre features includes discourse markers, primarily from Knott (1996), that indicate contrast, comparison, causation, evidence, condition, and similar relations.

The third type of genre features was a list of 500 adjectives classified in terms of Appraisal (Martin and White, 2005) as indicating Appreciation, Judgment or Affect. Appraisal categories have been shown to be useful in improving the performance of polarity classifiers (Whitelaw et al., 2005).

Finally, we also include text statistics as features, such as average length of words and sentences and position of paragraphs in the text.

3.2 Classifiers

To classify paragraphs in the text, we use the WEKA suite (Witten and Frank, 2005), testing three popular machine learning algorithms: Naïve Bayes, Support Vector Machine, and Linear Regression (preliminary testing with Decision Trees suggests that it is not appropriate for this task). Training parameters were set to default values.

In order to use Linear Regression, which provides a numerical output based on feature values and derived feature weights, we have to conceive of Comment/Describe/Describe+Comment not as nominal (or ordinal) classes, but rather as corresponding to a Comment/Describe ratio, with "pure" Describe at one end and "pure" Comment at the other. For training, we assign a 0 value (a Comment ratio) to all paragraphs tagged Describe and a 1 to all Comment paragraphs; for Describe+Comment, various options (including omission of this data) were tested. The time required to train a linear regression classifier on a large feature set proved to be prohibitive, and performance with smaller sets of features generally quite poor, so for the linear regression classifier we present results only for our compact set of genre features.

3.3 Performance

Table 2 shows the performance of classifier/feature-set combinations for the 2-, 3-, and 4-class tasks on the 100-text training set, with 10-fold cross-validation, in terms of precision (P), recall (R) and F-measure [2]. SVM and Naïve Bayes provide comparable performance, although there is considerable variation, particularly with respect to the feature set; the SVM is a significantly ($p<0.05$) better choice for our genre features [3], while for the n-gram features the Bayes classification is generally preferred. The SVM-genre classifier significantly outperforms the other classifiers in the 2-class task; these genre features, however, are not as useful as 5-grams at identifying Formal zones (the n-gram classifier, by contrast, can make use of words such as *cast*). In general, formal zone classification is fairly straightforward, whereas identification of Describe+Comment is quite difficult, and the SVM-genre classifier, which is more sensitive to frequency bias, elects to (essentially) ignore this category in order to boost overall accuracy.

To evaluate a linear regression (LR) classifier, we calculate correlation coefficient ρ, which reflects the goodness of fit of the line to the data. Table 3 shows values for the classifiers built from the corpus, with various Comment ratios

[2] For the 2- and 3-way classifiers, Describe+Comment paragraphs are treated as Comment. This balances the numbers of each class, ultimately improving performance.

[3] All significance tests use chi-square (χ^2).

Classifier	Comment			Describe			Formal			Desc+Comm			Overall Accuracy
	P	R	F	P	R	F	P	R	F	P	R	F	
2-class-5-gram-Bayes	.66	**.79**	.72	.70	.55	.62	-	-	-	-	-	-	68.0
2-class-5-gram-SVM	.53	.63	.64	.68	.69	**.69**	-	-	-	-	-	-	66.8
2-class-genre-Bayes	.66	.75	.70	.67	.57	.61	-	-	-	-	-	-	66.2
2-class-genre-SVM	**.71**	.76	**.74**	**.71**	.65	.68	-	-	-	-	-	-	**71.1**
3-class-5-gram-Bayes	.69	.49	.57	.66	**.78**	**.71**	**.92**	**.97**	**.95**	-	-	-	**78.1**
3-class-5-gram-SVM	.64	.63	.63	**.68**	.65	.65	.91	**.97**	.94	-	-	-	77.2
3-class-genre-Bayes	**.68**	.68	.66	.67	.46	.55	.84	.96	.90	-	-	-	74.0
3-class-genre-SVM	.66	**.71**	**.68**	.67	.56	.61	.90	.94	.92	-	-	-	76.8
4-class-5-gram-Bayes	**.46**	.35	.38	**.69**	.47	.56	**.92**	**.97**	**.95**	**.42**	**.64**	**.51**	69.0
4-class-5-gram-SVM	.43	**.41**	.44	.59	.62	.60	.91	**.97**	.94	.45	.41	.42	**69.6**
4-class-genre-Bayes	.38	.31	.34	.66	.30	.41	.86	**.97**	.90	.33	.60	.42	62.3
4-class-genre-SVM	.46	.32	.38	.53	**.82**	.65	.87	.94	.90	.26	.03	.06	67.4

Table 2. Stage identification performance of various categorical classifiers

(C) assigned to paragraphs with the Describe+Comment tag, and with Describe+Comment paragraphs removed from consideration.

Classifier	ρ
LR, Des+Com C = 0	.37
LR, Des+Com C = 0.25	.44
LR, Des+Com C = 0.5	.47
LR, Des+Com C = 0.75	.46
LR, Des+Com C = 1	.43
LR, No Des+Com	.50

Table 3. Correlation coefficients for LR classifiers

The drop in correlation when more extreme values are assigned to Describe+Comment suggests that Describe+Comment paragraphs do indeed belong in the middle of the Comment spectrum. Since there is a good deal of variation in the amount of comment across Describe+Comment paragraphs, the best correlation comes with complete removal of these somewhat unreliable paragraphs. Overall, these numbers indicate that variations in relevant features are able to predict roughly 50% of the variation in Comment ratio, which is fairly good considering the small number and simplistic nature of the features involved.

4 Sentiment detection: SO-CAL

In this section, we outline our semantic orientation calculator, SO-CAL. SO-CAL extracts words from a text, and aggregates their semantic orientation value, which is in turn extracted from a set of dictionaries. SO-CAL uses five dictionaries: four lexical dictionaries with 2,257 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs, and a fifth dictionary containing 177 intensifying expressions. Although the majority of the entries are single words, the calculator also allows for multiword entries written in regular expression-like language.

The SO-carrying words in these dictionaries were taken from a variety of sources, the three largest a corpus of 400 reviews from Epinions.com, first used by Taboada and Grieve (2004), a 100 text subset of the 2,000 movie reviews in the Polarity Dataset (Pang and Lee, 2004), and words from the General Inquirer dictionary (Stone, 1997). Each of the open-class words were given a hand-ranked SO value between 5 and -5 (neutral or zero-value words are not included in the dictionary) by a native English speaker. The numerical values were chosen to reflect both the prior polarity and strength of the word, averaged across likely interpretations. For example, the word *phenomenal* is a 5, *nicely* a 2, *disgust* a -3, and *monstrosity* a -5. The dictionary was later reviewed by a committee of three other researchers in order to minimize the subjectivity of ranking SO by hand.

Our calculator moves beyond simple averaging of each word's semantic orientation value, and implements and expands on the insights of Polanyi and Zaenen (2006) with respect to contextual valence shifters. We implement negation by shifting the SO value of a word towards the opposite polarity (*not terrible*, for instance, is calculated as -5+4 = -1). Intensification is modeled using percentage modifiers (*very engaging*: 4x125% = 5). We also ignore words appearing within the scope of *irrealis* markers such as certain verbs, modals, and punctuation, and decrease the weight of words which appear often in the text. In order to counter positive linguistic

bias (Boucher and Osgood, 1969), a problem for lexicon-based sentiment classifiers (Kennedy and Inkpen, 2006), we increase the final SO of any negative expression appearing in the text.

The performance of SO-CAL tends to be in the 76-81% range. We have tested on informal movie, book and product reviews and on the Polarity Dataset (Pang and Lee, 2004). The performance on movie reviews tends to be on the lower end of the scale. Our baseline for movies, described in Section 5, is 77.7%. We believe that we have reached a ceiling in terms of word- and phrase-level performance, and most future improvements need to come from discourse features. The stage classification described in this paper is one of them.

5 Results

The final goal of a stage classifier is to use the information about different stages in sentiment classification. Our assumption is that descriptive paragraphs contain less evaluative content about the movie being reviewed, and they may include noise, such as evaluative words describing the plot or the characters. Once the paragraph classifier had assigned labels we used those labels to weigh paragraphs.

5.1 Classification with manual tags

Before moving on to automatic paragraph classification, we used the 100 annotated texts to see the general effect of weighting paragraphs with the "perfect" human annotated tags on sentiment detection, in order to show the potential improvements that can be gained from this approach.

Our baseline polarity detection performance on the 100 annotated texts is 65%, which is very low, even for movie reviews. We posit that formal movie reviews might be particularly difficult because full plot descriptions are more common and the language used to express opinion less straightforward (metaphors are common). However, if we lower the weight on non-Comment and mixed Comment paragraphs (to 0, except for Describe+Comment, which is maximized by a 0.1 weight), we are able to boost performance to 77%, an improvement which is significant at the $p<0.05$ level. Most of the improvement (7%) is due to disregarding Describe paragraphs, but 2% comes from Describe+Comment, and 1% each from Background, Interpretation, and (all) Formal tags. There is no performance gain, however, from the use of aspect tags (e.g., by increasing the weight on Overall paragraphs), justifying our decision to ignore subtags for text-level polarity classification.

5.2 Categorical classification

We evaluated all the classifiers from Table 2, but we omit discussion of the worst performing. The evaluation was performed on the Polarity Dataset (Pang and Lee, 2004), a collection of 2,000 online movie reviews, balanced for polarity. The SO performance for the categorical classifiers is given in Figure 1. When applicable, we always gave Formal Zones (which Table 2 indicates are fairly easy to identify) a weight of 0, however for Describe paragraphs we tested at 0.1 intervals between 0 and 1. Testing all possible values of Describe+Comment was not feasible, so we set the weights of those to a value halfway between the weight of Comment paragraphs (1) and the weight of the Describe paragraph.

Most of the classifiers were able to improve performance beyond the 77.7% (unweighted) baseline. The best performing model (the 2-class-genre-SVM) reached a polarity identification accuracy of 79.05%, while the second best (the 3-class 5-gram-SVM) topped out at 78.9%. Many of the classifiers showed a similar pattern with respect to the weight on Describe, increasing linearly as weight on Describe was decreased before hitting a maximum in the 0.4-0.1 range, and then dropping afterwards (often precipitously). Only the classifiers which were more conservative with respect to Describe, such as the 4-class-5-gram-Bayes, avoided the drop, which can be attributed to low precision Describe identification: At some point, the cost associated with disregarding paragraphs which have been mistagged as Describe becomes greater that the benefit of disregarding correctly-labeled ones. Indeed, the best performing classifier for each class option is exactly the one that has the highest precision for identification of Describe, regardless of other factors. This suggests that improving precision is key, and, in lieu of that, weighting is a better strategy than simply removing parts of the text.

In general, increasing the complexity of the task (increasing the number of classes) decreases performance. One clear problem is that the identification of Formal zones, which are much more common in our training corpus than our test corpus, does not add important information, since most Formal zones have no SO valued words. The delineation of an independent Describe+Comment class is mostly ineffective,

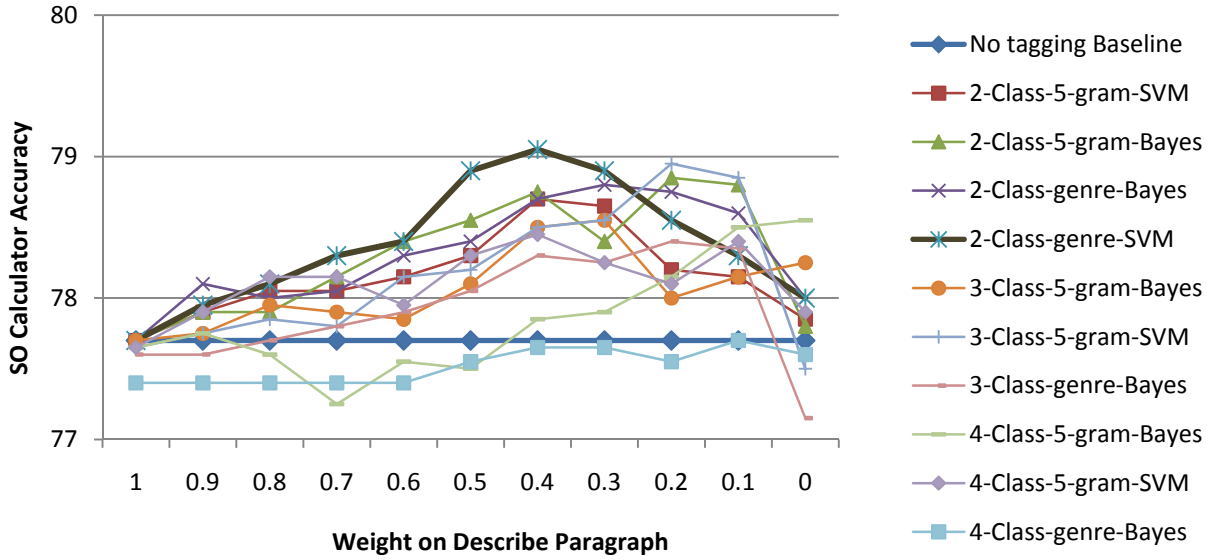

Figure 1. SO Performance with various paragraph tagging classifiers, by weight on Describe

probably because this class is not easily distinguishable from Describe and Comment (nor in fact should it be).

We can further confirm that our classifier is properly distinguishing Describe and Comment by discounting Comment paragraphs rather than Describe paragraphs (following Pang and Lee 2004). When Comment paragraphs tagged by the best performing classifier are ignored, SO-CAL's accuracy drops to 56.65%, just barely above chance.

5.3 Continuous classification

Table 4 gives the results for the linear regression classifier, which assigns a Comment ratio to each paragraph used for weighting.

Model	Accuracy
LR, Des+Com C = 0	78.75
LR, Des+Com C = 0.25	**79.35**
LR, Des+Com C = 0.5	79.00
LR, Des+Com C = 0.75	78.90
LR, Des+Com C = 1	78.95
LR, No Des+Com	79.05

Table 4. SO Performance with linear regression

The linear regression model trained with a 0.25 comment ratio on Describe+Comment paragraphs provides the best performance of all classifiers we tested (an improvement of 1.65% from baseline). The correlation coefficients noted in Table 4 are reflected in these results, but the spike at C = 0.25 is most likely related to a general preference for low (but non-zero) weights on Describe+Comment paragraphs also noted when weights were applied using the manual tags; these paragraphs are unreliable (as compared to pure Comment), but cannot be completely discounted. There were some texts which had only Describe+Comment paragraphs.

Almost a third of the tags assigned by the 2-class genre feature classifier were different than the corresponding n-gram classifier, suggesting the two classifiers might have different strengths. However, initial attempts to integrate the various high performing classifiers—including collapsing of feature sets, metaclassifiers, and double tagging of paragraphs—resulted in similar or worse performance. We have not tested all possible options (there are simply too many), but we think it unlikely that additional gains will be made with these simple, surface feature sets. Although our testing with human annotated texts and the large performance gap between movie reviews and other consumer reviews both suggest there is more potential for improvement, it will probably require more sophisticated and precise models.

6 Related work

The bulk of the work in sentiment analysis has focused on classification at either the sentence level, e.g., the subjectivity/polarity detection of Wiebe and Riloff (2005), or alternatively at the level of the entire text. With regards to the latter, two major approaches have emerged: the use of machine learning classifiers trained on n-grams

or similar features (Pang et al., 2002), and the use of sentiment dictionaries (Esuli and Sebastiani, 2006; Taboada et al., 2006). Support Vector Machine (SVM) classifiers have been shown to out-perform lexicon-based models within a single domain (Kennedy and Inkpen, 2006); however they have trouble with cross-domain tasks (Aue and Gamon, 2005), and some researchers have argued for hybrid classifiers (Andreevskaia and Bergler, 2008).

Pang and Lee (2004) attempted to improve the performance of an SVM classifier by identifying and removing objective sentences from the texts. Results were mixed: The improvement was minimal for the SVM classifier (though the performance of a naïve Bayes classifier was significantly boosted), however testing with parts of the text classified as subjective showed that the eliminated parts were indeed irrelevant. In contrast to our findings, they reported a drop in performance when paragraphs were taken as the only possible boundary between subjective and objective text spans.

Other research that has dealt with identifying more or less relevant parts of the text for the purposes of sentiment analysis include Taboada and Grieve (2004), who improved the performance of a lexicon-based model by weighing words towards the end of the text; Nigam and Hurst (2006), who detect polar expressions in topic sentences; and Voll and Taboada (2007), who used a topic classifier and discourse parser to eliminate potentially off-topic or less important sentences.

7 Conclusions

We have described a genre-based taxonomy for classifying paragraphs in movie reviews, with the main classification being a distinction between formal and functional stages, and, within those, between mainly descriptive vs. comment stages. The taxonomy was used to annotate 100 movie reviews, as the basis for building classifiers.

We tested a number of different classifiers. Our results suggest that a simple, two-way or continuous classification using a small set of linguistically-motivated features is the best for our purposes; a more complex system is feasible, but comes at the cost of precision, which seems to be the key variable in improving sentiment analysis.

Ultimately, the goal of the classification was to improve the accuracy of SO-CAL, our semantic orientation calculator. Using the manual annotations, we manage to boost performance by 12% over the baseline. With the best automatic classifier, we still show consistent improvement over the baseline. Given the relatively low accuracy of the classifiers, the crucial factor involves using fine-grained weights on paragraphs, rather than simply ignoring Describe-labeled paragraphs, as Pang and Lee (2004) did for objective sentences.

An obvious expansion to this work would involve a larger dataset on which to train, to improve the performance of the classifier(s). We would also like to focus on the syntactic patterns and verb class properties of narration, aspects that are not captured with simply using words and POS labels. Connectives in particular are good indicators of the difference between narration (temporal connectives) and opinion (contrastive connectives). There may also be benefit to combining paragraph- and sentence-based approaches. Finally, we would like to identify common sequences of stages, such as plot and character descriptions appearing together, and before evaluation stages. This generic structure has been extensively studied for many genres (Eggins and Slade, 1997).

Beyond sentiment extraction, our taxonomy and classifiers can be used for searching and information retrieval. One could, for instance, extract paragraphs that include mostly comment or description. Using the more fine-grained labels, searches for comment/description on actors, directors, or other aspects of the movie are possible.

Acknowledgements

This work was supported by SSHRC (410-2006-1009) and NSERC (261104-2008) grants to Maite Taboada.

References

Andreevskaia, Alina & Sabine Bergler. 2008. When specialists and generalists work together: Domain dependence in sentiment tagging. Proceedings of 46th Annual Meeting of the Association for Computational Linguistics (pp. 290-298). Columbus, OH.

Aue, Anthony & Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. Proceedings of the International Conference on Recent Advances in Natural Language Processing. Borovets, Bulgaria.

Biber, Douglas. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Bieler, Heike, Stefanie Dipper & Manfred Stede. 2007. Identifying formal and functional zones in film reviews. Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue (pp. 75-78). Antwerp, Belgium.

Boucher, Jerry D. & Charles E. Osgood. 1969. The Pollyanna hypothesis. Journal of Verbal Learning and Verbal Behaviour, 8: 1-8.

Di Eugenio, Barbara & Michael Glass. 2004. The kappa statistic: A second look. Computational Linguistics, 30(1): 95-101.

Eggins, Suzanne & James R. Martin. 1997. Genres and registers of discourse. In Teun A. van Dijk (ed.), Discourse as Structure and Process. Discourse Studies: A Multidisciplinary Introduction (pp. 230-256). London: Sage.

Eggins, Suzanne & Diana Slade. 1997. Analysing Casual Conversation. London: Cassell.

Esuli, Andrea & Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. Proceedings of 5th International Conference on Language Resources and Evaluation (LREC) (pp. 417-422). Genoa, Italy.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76: 378-382.

Kennedy, Alistair & Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence, 22(2): 110-125.

Knott, Alistair. 1996. A Data-Driven Methodology for Motivating a Set of Coherence Relations. Edinburgh, UK: University of EdinburghThesis Type.

Martin, James R. & Peter White. 2005. The Language of Evaluation. New York: Palgrave.

Nigam, Kamal & Matthew Hurst. 2006. Towards a robust metric of polarity. In Janyce Wiebe (ed.), Computing Attitude and Affect in Text: Theory and Applications (pp. 265-279). Dordrecht: Springer.

Orasan, Constantin. 2003. PALinkA: A highly customizable tool for discourse annotation. Proceedings of 4th SIGdial Workshop on Discourse and Dialog (pp. 39 – 43). Sapporo, Japan.

Pang, Bo & Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of 42nd Meeting of the Association for Computational Linguistics (pp. 271-278). Barcelona, Spain.

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

Polanyi, Livia & Annie Zaenen. 2006. Contextual valence shifters. In James G. Shanahan, Yan Qu & Janyce Wiebe (eds.), Computing Attitude and Affect in Text: Theory and Applications (pp. 1-10). Dordrecht: Springer.

Seki, Yohei, Koji Eguchi & Noriko Kando. 2006. Multi-document viewpoint summarization focused on facts, opinion and knowledge. In Janyce Wiebe (ed.), Computing Attitude and Affect in Text: Theory and Applications (pp. 317-336). Dordrecht: Springer.

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

Taboada, Maite, Caroline Anthony & Kimberly Voll. 2006. Creating semantic orientation dictionaries. Proceedings of 5th International Conference on Language Resources and Evaluation (LREC) (pp. 427-432). Genoa, Italy.

Taboada, Maite & Jack Grieve. 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.

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 & 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 (pp. 337-346). Gold Coast, Australia.

Whitelaw, Casey, Navendu Garg & Shlomo Argamon. 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, Janyce & Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005). Mexico City, Mexico.

Witten, Ian H. & Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques (2nd edn.). San Francisco: Morgan Kaufmann.

Appendix A: Full lists of formal and functional zones

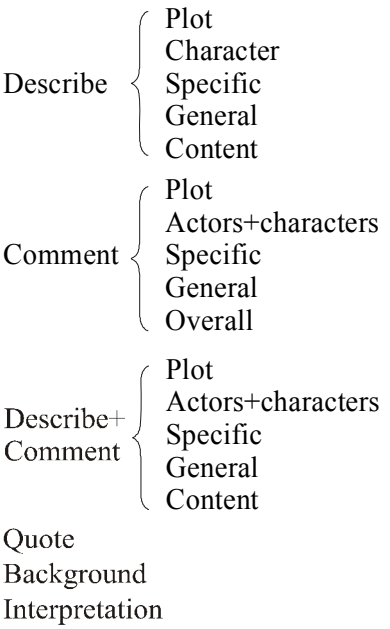

Figure A1. Functional zones

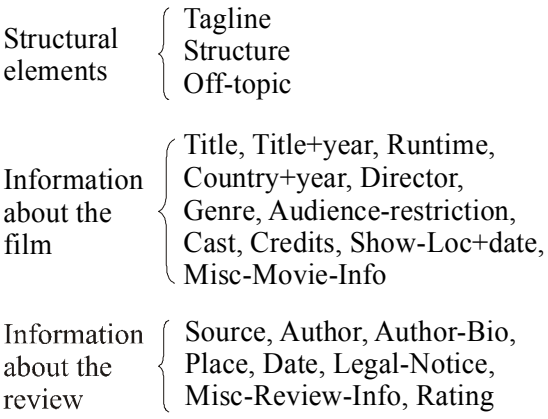

Figure A2. Formal zones

Appendix B: Kappa values for annotation task

Classes	**2-rater kappa**	**3-rater kappa**
Describe/Comment/Describe+Comment/Formal	.82	.73
Describe/Comment/Formal	.92	.84
Describe/Comment/Describe+Comment	.68	.54
Describe/Comment	.84	.69

Table B1. Kappa values for stage annotations