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 expression 0 Twv4t0.0386 Tw T[coa147fic oot exS sha)1t exl2he pogatw147 Td[a 17 hangeof a wo4(

nlevopn 0 ( )]TJ0.0009 Tc 011475 Tw 0 -1.147 TDstagisedriorar it(Eggins(i an)4(Mar inse)

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

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	.38	[illegible]		

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 classi-

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

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 kaatat2a &eiloo(a),s. C tl Li aandksguiceeor, 3(s, 3(I)63(h75)5(163(h7):)11p.)6(9(I563(h-)5(16 I)1C /P7<</MCID 2 BDC -0

Appendix A: Full lists of formal and functional zones

Describe

Comment

Figure A2. Formal zones

Figure A1. Functional zones