Patterns in the Stream: Exploring the Interaction of Polarity, Topic, and Discourse in a Large Opinion Corpus

Julian Brooke
Department of Computer Science
University of Toronto*
27 King's College Circle
Toronto, ON, Canada M5S 1A1
jbrooke@cs.toronto.edu

Matthew Hurst
Microsoft Corporation
1 Microsoft Way
Redmond, WA, 98052 USA
mhurst@microsoft.com

ABSTRACT
A qualitative examination of review texts suggests that there are consistent patterns to how topic and polarity are expressed in discourse. These patterns are visible in the text and paragraph structure, topic depth, and polarity flow. In this paper, we employ sentence-level sentiment classifiers and a hand-built tree ontology to investigate whether these patterns can be quantitatively identified in a large corpus of video game reviews. Our results indicate that the beginning and the end of major textual units (e.g. paragraphs) stand out in the flow of texts, showing a concentration of reliable opinion and key topic aspects, and that there are other important regularities in the expression of opinion and topic relevant to their ordering and the discourse markers with which they appear.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Measurement, Experimentation.

Keywords: Opinion mining, sentiment analysis, topic detection, discourse analysis

1. INTRODUCTION
The expression of opinion relies on the establishment of topic, its association with positive or negative sentiment and often the justification of the speaker's point of view. From a linguistic stance, various syntactic and discourse structures are employed. In this paper, we use simple empirical techniques and some extant sentiment analysis tools to explore the relationships between sentiment, topic and discourse. The results of this investigation will prove valuable in expanding the scope of sentiment analysis beyond the simplistic (e.g. bag-of-words) methods that are often employed. We begin with examples taken from our video game corpus (see Section 2) which illustrate some of the key themes we will be exploring:

The graphics in the game are highly detailed, and character models look very nice. The problem is that the environments, while sometimes vast, are dull and uninteresting. Pair this with the fact that there aren't even that many NPCs in the game's world, and you've got a land which you don't even care about saving. Battle animations are the high point here, with some cool looking summons and spells. Overall I think the graphics weren't a big problem, but the style was certainly lacking.

The paragraph as a whole is about graphics, but there is a distinct structure to the way the topic is discussed. The author begins with the general term *graphics*, but quickly moves to specific aspects (*models*, *environments*, *animation*), returning to *graphics* at the end. Both *environments* and *animations* involve a further regression, in the former case the mention of another fact to make a general point, in the latter some specific examples. There are patterns to the change in polarity as well. The two clauses in the first sentence are joined by *and*, showing the ideas are of compatible polarity, but then *the problem* signals a turn to negative polarity that is maintained until the next change in topic, where the use of *the high point* is an indicator that another switch has occurred. *Overall* signals an end to the details, and an offering of the author's primary opinion on the topic.

Here is another example (from a different text):

Sound- 10/10. The sound is actually very impressive in oblivion. While wandering around aimlessly in the countryside, a nice, calm, melodic theme will be accompanying you in your travels, but if you happen to come across a non-friendly character of any kind, a more up-beat music will kick in, to let you know to be on high alert. Also, you might notice the subtle changes in the sound that your feet make as you transition from dirt, to mud, to grass and so on. This is fairly impressive, because there is a countless amount of terrain put into Oblivion, and it all has high quality sound.

This paragraph is consistently positive, signaled by the *also* in the third section. Again, the main topic (*sound*) is mentioned at both the beginning and the end of the paragraph, using simple predicative language. The rest of the paragraph is devoted to explanation to two aspects that directly support that opinion,

*Research conducted at Microsoft Live Labs and MSN.

TSA'09, November 6, 2009, Hong Kong, China.

music and *sound effects*, and in fact much of the actual text is not direct opinion, but rather neutral and matter-of-fact; words like *aimlessly*, *non-friendly*, and *alert* are not to be understood as reflecting negatively on the topic of music (nor *up-beat* positively, despite the fact it modifies *music* directly), but in the context of an example that illustrates why, ultimately, the sound is so impressive.

These examples highlight an interesting relationship that exists between topic and sentiment at the level of discourse. Similar patterns, i.e. simpler, more general, more polar language at the beginning and end of textual units and internal polarity and topic shifts that are predictable and often marked, are also apparent at higher levels of discourse. To explore these relationships across large number of texts, we collected a sizable corpus of video game reviews, and carried out an analysis using two different sentence-level sentiment classifiers (one lexical, one machine-learning based), a hand-built video game ontology, and various cues to the structure of the discourse at the text, paragraph, and sentential levels. Not only were we able to confirm most of our intuitions, we also gathered useful information about the strengths and weakness of the tools that were used in the analysis.

This paper is structured as follows: Sections 2 and 3 introduce the dataset and our tools for sentence polarity detection, respectively. Section 4 is concerned with topic, which we equate with traversal of a tree ontology. Section 5 brings text-level structure and local discourse markers into the discussion, and it is in this section that we present our key findings. Section 6 is review of relevant work, and in Section 7 we offer conclusions and directions for future research.

2. CORPUS

In the field of text-level sentiment analysis, movie, camera, and restaurant reviews seem to have received the most attention [8,18,25]. For this project, we chose to collect a new corpus of video game reviews partially as an effort to expand our understanding of different domains, and also because video games have a particularly interesting set of features; they have much in common with other cultural products like movies and books (for instance, they often have a storyline, setting, characters), but are grounded in technology and involve an interactive, on-going experience.

We collected 48,050 unique reviews from the epinions.com website, including reviews on 2,822 games by 19,530 different authors, representing 14 different gaming platforms and 10 different genres (though there were only 390 Puzzle game reviews and 83 Educational game reviews). From the html of each review we extracted the title of the game, the genre, the platform, the star rating, the recommendation, the pros and cons, the helpfulness rating, the number of comments, the author, the number of users who trust the author, the number of reviews written by the author, and of course the text itself.

Though not explicitly required for our analysis, the extra information provided in the review proved valuable. For instance, the titles and common derivations thereof (e.g. acronyms) were used as indicators for the game node in our ontology (Section 3). Importantly, we found that the corpus was strongly biased towards positive reviews, with 21,910 5-star reviews, 14,868 4-star reviews, 5,815 3-star reviews, 3,371 2-star reviews, and 2,080 1 star reviews (5 reviews had no rating). The corpus was also strongly biased towards helpful reviews, with 21,961 rated *very helpful*, 11,627 rated *helpful*, 10,405 rated *somewhat helpful*, and 4,056 with no helpfulness ratings; there is an option to indicate off-topic reviews that was never used. We looked at the correlations between the numerical aspects, and found weak correlations between *number of comments*, *trusted count*, and *helpfulness*, and a strong correlation between *trusted count* and *reviews written*. The vast majority of authors wrote only a single review, but there were a small number who wrote tens or even hundreds; a preliminary analysis indicated that reviews written by a single author showed far less variation in areas such as helpfulness, length and even the mentioning of specific aspects in our ontology.

The average text had 27.9 sentences (the entire corpus has about 1.3 million sentences) in 7.9 paragraphs, 73% of the reviews had more than one paragraph. About 15% of the reviews also had section headings, identified based on capitalization, punctuation, and length (though we did not detect headings embedded at the beginning of a paragraph, as in second example above). In tokens, the average text was 529 words long, with a standard deviation of 596.99. About 1000 reviews were excluded from our main analysis because they contained sentences of extreme length (which in turn caused problems with our sentiment classifiers); this seemed to be the result of non-standard spacing. Manual inspection of the corpus was mostly limited to a small sample of about 200 reviews.

3. POLARITY DETECTION

Our corpus provides us with recommendations and star ratings at the level of the text, however for our investigation, we require low-level sentiment information; in particular, we would like to be able to distinguish between positive, negative, and neutral sentences. Sentence polarity detection is of course a well studied problem [27], but we are not interested here in the details of sentiment detection, we just need something that will give us an idea about polarity distributions. One option would be to train a classifier using a portion of the corpus, however we wanted a domain-independent classifier which would not confuse topic with polarity (see the discussion in 4.2). We instead made use of two separate classifiers that were available to us. The first is a lexical classifier (SO-CAL) which has shown good cross-domain performance [6]; it uses a hand-ranked, multi-POS dictionaries, contextual valence shifters [20], and negative weighting to achieve approximately 75-80% 2-way text polarity classification accuracy in unfamiliar product domains. The other classifier is a maximum entropy machine learning (ML) model which has been trained on unigrams and bigrams from a mixed corpus of movie reviews[19], books, DVDs, electronics, and kitchen appliances [4]; both classifiers are intended for cross-domain use, but neither had been built using data from the video game domain. The classifiers do not have a specific neutral class, but both provide numerical values (the lexical classifier a semantic orientation value, the machine classifier a confidence measure) for which thresholds between positive/negative/neutral were selected based on inspection of output; an attempt was made to have to have comparable counts of each class. The percentage of positive/negative/neutral sentences in our corpus

were .38/.23/.39 for the lexical classifier and .38/.21/.40 for the ML classifier.

To evaluate the performance of these two classifiers at the task, we randomly extracted 300 sentences from our corpus, and had two human annotators manually assign polarity ratings (positive, negative, or neutral). We then calculated the kappa agreement [9] among human (H1, H2) and computer annotators (ML, LEX), provided in Table 1.

Table 1: Kappa Agreement for Classifier Evaluation

	H2	LEX	ML
H1	0.69	0.44	0.25
H2	-	0.42	0.24
LEX	-	-	0.20

Although below the human agreement, both classifiers are well above chance. In general, the lexical classifier correlates best with human judgment, however the fact that their human agreements are higher than their mutual agreement suggests that they each get some correct that the other gets wrong (a manual inspection confirms this). In light of this, we decided to use both classifiers, preferring results where the classifiers were in agreement. Table 2 provides the classification statistics (precision, recall, f-score) for the classifiers using H1 as the gold standard:

Table 2: Classification Statistics for Polarity Classifiers

Polarity	ML Classifier			Lexical Classifier		
	P	R	F	P	R	F
Positive	.57	.62	.59	.63	.70	.66
Negative	.31	.37	.34	.55	.60	.58
Neutral	.59	.50	.54	.70	.60	.65

We also found that we could boost the precision of positive and negative detection by an average of .12 if we classify sentences as positive and negative only when the two classifiers agree. One potential confounding factor: both classifiers tended to classify short sentences as neutral, and longer sentences as either positive or (especially very long sentences) negative, however although there were consistent sentence length variations (sentences got longer towards the end of a paragraph), none of the patterns we saw would be explained by this fact alone.

4. TOPIC

4.1 A Tree Ontology for Video Games

In the introduction, we noted that topics in a video game review seem to be ordered hierarchically, in fairly predictable patterns that correspond roughly to the meronymy relation (as understood in this very specific context). In order to explore this in more detail, we constructed a domain-specific ontology using direct observation of patterns in the data as well as general knowledge about the domain. Since the relationships between aspects were hierarchical, with a few major nodes and many minor ones, a tree structure seemed the most logical choice, allowing us to indicate these relationships without the complexity of a full graph. One section of the current tree is given in Figure 1.

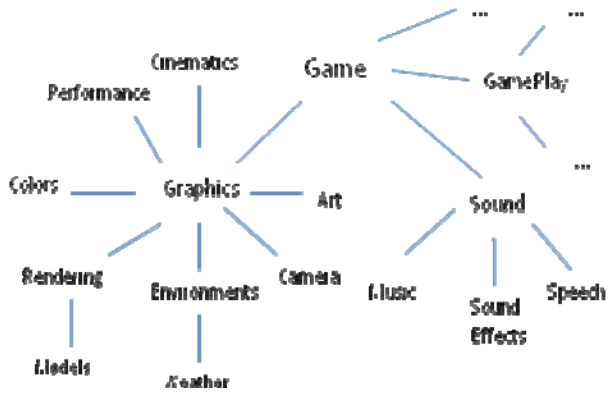

Figure 1: Section of the Video Game Tree Ontology

In all, there are 61 nodes in the ontology, including 12 non-terminal nodes. A small subset of the nodes (6) were genre specific, meaning we only extracted these features when a review was of a particular genre[1].

Each node is associated with a number of regex indicators (there are 890 indicators in all) that allow us to identify mentions of these topics in the text. After initial construction of the ontology was complete, we randomly extracted 10 sentences from the corpus for each node in the ontology and manually judged whether they had correctly identified a mention of the topic, yielding a precision of 84.6%. Recall is harder to judge, though one easy indication is the percentage of sentences that were found to contain at least some topic in the ontology, 67.3%.[2]

Besides the effort involved, there are a number of drawbacks to this sort of ontology. First, though it gets many of the high-level patterns across a broad spectrum of reviews in this domain, there is a clear inability to capture the "true" ontology of the individual game, which often leaves large sections of the ontology untouched while relying on genre and game-specific terminology. Though the relationship among certain nodes is clear, others feel arbitrary and unsatisfying, in particular nebulous concepts like *options* and *modes*. Also, the tree architecture fails to capture the distinction between various facets of an element (a *weapon* as a *gameplay* object versus a *weapon* as a source of *sound*), requiring an arbitrary decision.

Nevertheless, the ontology clearly reflects some important aspects of the domain. For example, we found a strong correlation (.47) between the reader-assigned helpfulness of the review and the overall ontological coverage (the total number of nodes which are indicated in the text), the best correlation of

[1] For instance, the node *races* is a common aspect of role-playing games, e.g. elves and dwarves, but this word means something completely different in the context of a driving game

[2] Though a manual inspection of the topic-less sentences suggests that some of them truly have no clear indication of their topic and cannot be reliably interpreted outside of their discourse context, e.g. *A second chance, I like that*.

helpfulness with any other feature (including user trustedness) and significantly better than simple token length (.41); in general, shallow nodes in the ontology (such as *gameplay*) did not strongly predict helpfulness while deep nodes (such as *model*, the best predictor) did. An investigation of section headings finds non-terminal nodes overwhelmingly represented, with other nodes relatively rare. Other experiments which provide additional support for the utility of the model are described later in this paper.

In order to derive useful statistics from our ontology, we derived two simple metrics to be applied to sets of features (instances of nodes) found within a text: feature depth and feature breadth. Feature depth is simply the average depth of relevant nodes in the tree (*game* is at depth 0) while feature breadth is calculated using *c*, the number of nodes represented (directly or through inheritance) at tree depth 1 and *n*, the total number of features in the tree, according to the formula c^2/n, the result being that a tree whose *n* features are well spread out across the nodes of the tree will have a breadth > 1, whereas a tree whose features are concentrated on a single node will have breadth < 1. Note that this metric only deals which breadth at the highest level of the tree. The advantage of this is that the metric is also entirely independent of depth; for the purposes of the calculation, it is not relevant whether, say, the feature is *graphics* or *models*, only that there is some feature under the *graphics* node. We will use these two metrics later when discussing topic and discourse structure.

4.2 Topic and Polarity

Looking at the intersection of automatically-labeled topic and polarity features (the percentage of sentences with each polarity and topic), we first note that many of the nodes in the ontology show strong preferences for one polarity or another, even beyond the biases already present in the corpus. One reason for this is encoded in the ontology itself: many of the words which indicate topic simultaneously indicate polarity. For instance, the indicators of the *graphics* node include (seemingly) neutral words like *visuals*, but also sentiment-laden words like *beautiful* and *ugly*. To ignore these words in the interests of separability is one option, but we have chosen to include them and instead attempt to balance positive and negative indicators. However, this has obviously been unsuccessful in some cases, several nodes with a number of positive-connotation indicators are extremely positive (*gameplay*, *graphics*, *achievements, and extras*) whereas other nodes have been obviously been dragged down by negative-connotation indicators (*difficulty*, *enemies*, *fighting*). This is a primarily a problem with the polarity provided by the lexical classifier, as the machine learning classifier trained on consumer reviews is unlikely to learn, for instance, that *enemy* is a negative polarity word.

Another aspect of the bias has to do with elements of the topic that, by their very mention, suggest a positive or negative stance. There seem to be a number of these in the video game domain: positive topics include *originality*, *realism*, *physics*, and *genre*, whereas *bugs*, *performance*, and *camera* are negative. This last example highlights exactly why machine classifiers are so domain-dependent [2]; if a classifier trained in this video game corpus were used in the popular domain of camera reviews, reviews would be deemed negative simply based on the use of the word *camera*, a dubious indicator to say the least. Even in the video game domain it is questionable, since it is a tendency, not a categorical distinction (Even *bugs* might be used positively, when noting, say, that bugs from the previous installment in the series have been fixed). In short, where topic meets polarity either kind of polarity classifier can be led astray.

One potential application of a fine-grained topic model is learning which aspects are more central to overall opinion. To investigate this, we counted positive and negative sentences containing nodes in our ontology (we only used sentences where the two classifiers agreed), and calculated to what extent the polarity of individual nodes were able to predict the overall recommendation. In order to control for the overwhelming positive bias of our corpus, we split the corpus into recommended and not recommended texts, looking separately at each. Table 3 contains lists of nodes that were consistently good (above average) predictors or consistently poor (below average predictors) for both parts of the corpus. Note that in general, all but the most negatively-biased nodes were better predictors of positive recommendations, which suggest additional positive bias in either the classifiers or the text.

Table 3: Predictors of Review Recommendation

Good Predictors	Poor Predictors
game, gameplay, graphics, sound, story, controls	cinematics, weather, saving, customization, goals, instruction, characters, setting, interfaces, release, developer

All of the best predictors were key high-level nodes, i.e. those that tend to appear in section headings, whereas the worse predictors tended to be lower in the tree and/or somewhat tangential to the main gameplay experience. Overall, there was a -0.32/-0.46 correlation between predictive power and tree depth of the node in the positive/negative corpus. This provides some preliminary support for the tree architecture of our ontology and the efficacy of the classifiers. That said, there is a lot of relevant opinion that is not to being connected to a particular node in the ontology; for negative texts this "floating" opinion is actually more predictive than the *game* node itself. Figuring out where this opinion "attaches" will, we believe, require a better understanding of how the flow of topic and polarity fit into the structure of discourse.

5. DISCOURSE

5.1 Indicators of Discourse Structure

As noted earlier, the majority of the texts in our corpus are organized into paragraphs. There is considerable debate as to exactly what a paragraph break means [22], however the structure provided by paragraphs plays an important role in, for instance, features that identify the discourse structure in automatic essay-scoring [7]. Here, we examine two aspects: the position of paragraphs in the text and the position of sentences within the paragraph. Since the paragraph length of texts and the sentence length of paragraphs vary widely, we use a set number of length "buckets" based on the rounded-off average lengths (for paragraphs, 10, for sentences, 5), scaling appropriately for texts/paragraphs and that are longer or shorter than average.

First and last paragraphs/sentences are always placed in the first and last buckets, respectively, and single paragraph texts and single sentence paragraphs were not included.

Another way to view the discourse is as a graph, with transition probabilities between various nodes (in this case polarities or topic aspects). We can calculate these probabilities directly from the data using co-occurrence counts. This kind of information gives us a sense of how topic/polarity flows throughout a text.

Our other indication of discourse structure comes from sets of discourse cues included in Knott [11]. These cues indicate the logical and functional relationships within text, including causation, evidence, condition, continuation, comparison contrast, restatement, alternatives, initiation, and conclusion. These groupings are somewhat analogous to what is available in theories of discourse such as RST [14], however some categories are collapsed to avoid ambiguity. We used lists of words relevant to categories in the Appraisal Hierarchy of Martin and White [15], including Appreciation, Judgement, and Affect as well as indicators of Engagement and Graduation, and some syntactic patterns that can be derived directly from tags, including predicative and attributive adjectives.

5.2 Discourse and Polarity

We begin our intersection of discourse and polarity at the textual level, looking at the distribution of positive, negative, and neutral sentences in paragraphs from the beginning to the end of the text. Based on the patterns we noted in Section 1, our expectations are increased neutrality in the middle of the text, and, since our corpus is positive-biased, strong positive sentiment in the beginning and the end. Figure 2 shows the average polarities across the text, for both the lexical and machine learning classifier.

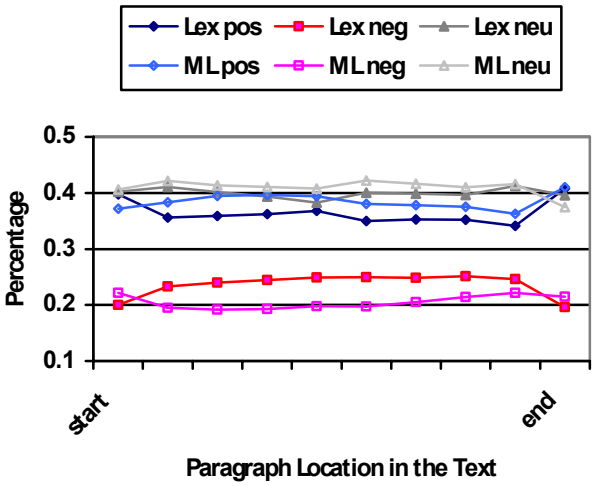

Figure 2: Average Polarity of Paragraphs by Relative Position in the Text

Both classifiers show a strong positive upturn at the end of the texts and a corresponding drop in neutrality, consistent with our predictions. A positive preference is also visible at the beginning of texts (though the two classifiers disagree somewhat to its shape), and the changes at both beginning and end are significant at the P<0.0001 level. What is not immediately clear is why the lexical classifier has (relatively) elevated negative polarity for the middle of the text. In previous work by one of the authors [5], it was noted that movie reviews, which have a great deal of plot/character description, are on average much more negative (in terms of individual lexical items) than other reviews. As mentioned in Section 3, SO-CAL has weighting of negative expressions to counteract the general positive bias of text, weighting which leads to statistically significant improvement in classification at the text level; however, when dealing with movie reviews the default weighting tends to overshoot the mark, resulting in increased negative classification. That is the most plausible explanation for what is happening here, i.e. descriptive passages being classified preferentially as negative rather than neutral (the relatively low negative precision in Table 2 supports this). Tellingly, when only negative (not recommend) texts are considered, the ML classifier shows significantly increased negativity both at the beginning and the end of the text, while for the Lexical classifier the negative polarity curve is high but nearly flat.

Within paragraphs, the pattern is similar. Figure 3 shows the average polarity of sentences by their location in the paragraph.

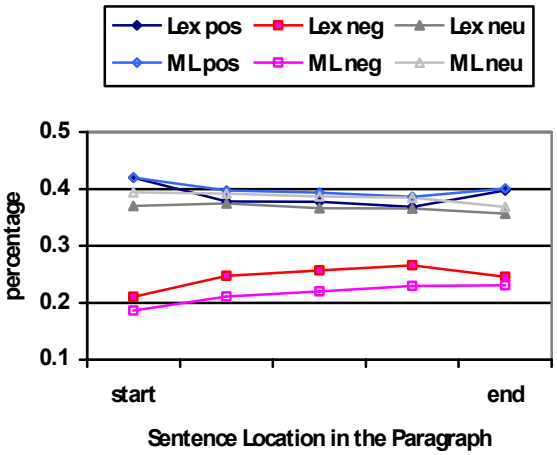

Figure 3: Average Polarity of Sentences by Relative Position in the Paragraph

Again, positive sentiment is significantly ($P < 0.0001$) higher at the beginning and end of paragraphs, and the middle is associated with elevated negative sentiment. The most likely explanation is that the increasing negative sentiment is a result of misclassified neutral sentences.

Next we look at the (paragraph-internal) transition probabilities between sentences of various polarities.

Table 4: Polarity Transition Probabilities

Initial Polarity	Subsequent Polarity					
	Lexical Classifier			ML Classifier		
	Pos	Neu	Neg	Pos	Neu	Neg
Positive	.44	.34	.21	.45	.37	.17
Neutral	.36	.41	.22	.38	.41	.20
Negative	.34	.34	.32	.34	38	.28

Table 4 suggests that two adjacent sentences are more likely to have the same polarity than would otherwise be expected (compared to the base probabilities, see Section 3), a desirable result. This phenomenon could be attributed to either a single topic and/or the organization of text into pros and cons.

Finally, we examine which discourse-relevant markers tend to appear in sentences of each polarity, ignoring those instances when the two classifiers disagreed. One of the associations follows directly from the results in Figure 2: discourse markers that indicate conclusion are highly positive, as are markers of continuation (recall that our corpus is biased towards the positive; conclusion and continuation, as markers of polarity flow, would tend to amplify this bias). The discourse markers we grouped into initiators, on the other hand, were more negative, perhaps because some of them indicate an initial state that eventually underwent change (*at first..*) or seem to have some inherent negativity in tone (*for starters*). Also intriguing is the strong polarity (and, in particular, negative polarity) in sentences with causative markers; the simplest explanation involves mistagging of more complex, mixed opinion/fact sentences, though it might also involve a tendency to provide explicit reasons for strong emotional (subjective) content.

We took special interest in the sentences at the (relatively rare) polarity flip boundary. Regardless of the direction of the switch, and even after sentence length was controlled for, there were clear patterns in the data; several of our discourse markers were more common than usual in the sentence after the switch, including contrast, hedges, conditionals, even modals. Though the appearance of contrast was expected (and the effect was strong), the other markers give us pause; it seems likely that some of the effect here is due, once again, to mistagging of complex sentences. Nevertheless, finding an independent way to identify points in the text where a polarity transition occurs seems a promising application of discourse information.

5.3 Discourse and Topic

In Section 4.2 we introduced two metrics for measuring the distribution of features within our tree ontology: average depth and surface breadth. Here we begin by applying those metrics at two levels of discourse within the text. Figure 4 shows the average feature breadth for paragraphs at various (relativized) locations within the text, while Figure 5 shows the same information for average feature depth.

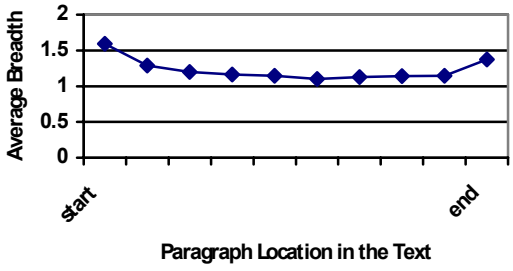

Figure 4: Average Feature Breadth of Paragraphs by Relative Position in the Text

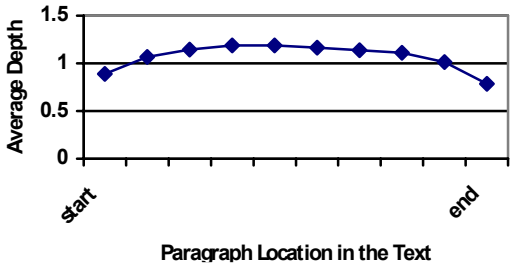

Figure 5: Average Feature Depth of Paragraphs by Relative Position in the Text

As expected, the two metrics, though independent, show almost opposite patterns; both the beginning and the end of the text are broad and shallow, consistent with introductions and conclusions, while the text gets deeper and more focused towards the middle. Note, however, that the curves are not mirror images of one another; the beginning of the text is broader than the end, but also deeper. The depth and breadth differences between the beginning, the middle, and the end of the text are significant at the $P<0.0001$ level (t-test). Interestingly, single paragraph texts (excluded from the above analysis) on average show much more breadth than paragraphs in multi-paragraph texts (breadth = 1.9) and in terms of depth they fall in the middle of the range between the middle and edges of the text (1.2), suggesting they contain elements of both.

Figures 6 and 7 show the spread for the same metrics at the sentence level, by relative position in the paragraph.

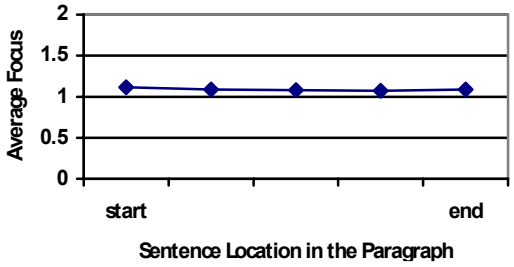

Figure 6: Average Feature Breadth of Sentences by Relative Position in the Paragraph

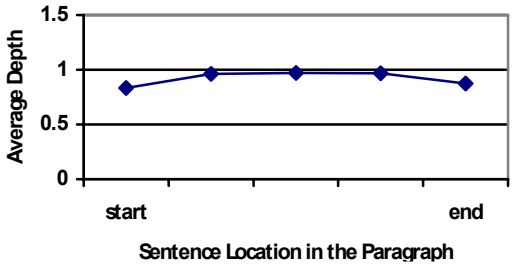

Figure 7: Average Feature Depth of Sentences by Relative Position in the Paragraph

Though less pronounced, the basic character of both curves is directly comparable to those that we saw with the text, suggesting paragraph-internal organization; in both cases, the difference between the lowest and highest points on the curve is statistically significant at the $P < 0.0001$ level. We would not anyway expect breadth to be a particularly useful metric at the sentence level, as it would vary from 1 only when two topics were indicated in the same sentence, and it only indicates breadth at the highest level of the ontology and thus would not be relevant to much of the within paragraph variation.

Despite a rebound toward the end of a paragraph, the dominant tendency at the paragraph/sentence level is increasing ontological depth. This tendency is magnified, and other interesting patterns emerge, when we look at sentence-level transitions between individual nodes. With 61 nodes of widely varying frequency, the transition matrix is too complex to effectively eyeball for patterns (except to confirm that ontologically-related nodes often appear together), however one investigation that proved fruitful was looking for major imbalances (a ratio greater than 1.5) in direction of the association, i.e. when one node appeared first in the discourse much more often. The *graphics* node, for instance, demonstrated this "unidirectional" property with all of its (ontological) children and grandchildren as well as the *sound* node and most of its children—obviously graphics and sound go together, and in exactly that order. Looking at the children of *graphics*, we see a few (typically) one-way sibling connections among them (e.g. rendering->camera, models->performance), one other parent-child connection (environments->weather), additional connections to *sound*, and then a few mostly predicable connections with other nodes. The *models* node, for instance, associates unidirectionally with *weapons* and *enemies*, game objects which usually have 3D models; in many of these cases, sequential ordering of the discourse as well as the hierarchical ordering of topics should allow us to understand the mention of game objects as occurring in a graphical context, though our ontology as it stands does not provide that flexibility.

One rather surprising fact emerged when we compared transition probabilities within sentences, across sentences boundaries, and across paragraph boundaries. The average entropy (the unpredictability of transitions) was slightly higher for across sentence transitions (4.66) as for within sentences (4.55), but markedly lower for paragraph transitions (4.32). This is unintuitive, since we would expect that paragraphs boundaries more often involve a topic shift, and thus a wider range of possible nodes. The likely explanation has to do with the depth of nodes at the paragraph boundaries; as we have seen, paragraphs tend to be shallower at the edges, and so the transition probabilities are concentrated in a small number of nodes higher in the ontology.

Examining the relationship between discourse markers and topic within the sentence (using again a 1.5 ratio above expected), we first note the influence of polarity, with some of aspects that we identified as inherently biased patterning after their most common polarity. In general, there seems to be a set of more neutral, descriptive nodes in the tree (e.g. *items*, *vehicles*, *achievements*) which show a richer collection of discourse and syntactic associations, with discourse markers related to alternative, evidence, and comparison appearing frequently (also the ambiguous semi-colon); attributive adjectives appear with these kinds of nodes more frequently than predicative adjectives. This pattern is reversed for other, generally more opinionated nodes, which are often associated with our Appraisal features but little else; *graphics*, for instance, is associated strongly only with predicative adjectives, Appreciation, Judgment, and downplayers, suggesting simple predicative statements (e.g. *the graphics were pretty good*).

6. RELATED WORK

The field of Sentiment Analysis is typically subdivided into work on sentence-level sentiment detection [27] and text-level sentiment detection [18,26], with another important distinction arising from the difference between lexical (dictionary) classification [26] and machine-learning classification [18]. Our work is relevant to all these areas, since we are using sentence-level polarity for text-level analysis, and employ both kinds of classifiers. There is other recent work has compared [10] or combined [1] the two types of classifiers.

With respect to topic in the sentiment domain, most of the work has been focused on deriving topic aspects automatically from text [21], and, most recently, mixed topic/sentiment models [16,24]. In this area, the work closest to ours is probably Carenini et al. [8], who automatically derive connections between a user-defined taxonomy of features (similar to our tree ontology) and words in the text.

With respect to discourse, there has been some recent interest in using paragraphs as a unit of analysis in sentiment detection [3,24]. Our effort to identify more and less sentiment-relevant spans at the text level is similar to work by Pang and Lee [18] and Tabaoda and Grieve [23]. Perhaps most similar is work by Mao and Lebanon [13], who model local sentiment flow using conditional random fields. There is also a clear connection to research on lexical chains [16].

Though we have dealt with the text quality only tangentially here, in the area of automated essay grading the organization of discourse has been used as a feature to the determine the quality of a text [7], and there has been a lot of recent work on review helpfulness and its relation to polarity and topic [12].

7. DISCUSSION AND FUTURE WORK

Here, we have adopted a different approach from most research in this area, using a set of tools to quantify some qualitative patterns in our data relevant to the organization of polarity and topic in opinion discourse. For the most part, we saw what we expected: concentration of opinion and key aspects at the beginning and end of textual units, and fairly predictable transitions of opinion and topic that are reflected in the discourse cues. The next step would be to move from showing that such patterns exist to identifying specific instances. Some of our results suggest that we might need to improve our tools, including a sentence classifier which is better adapted to 3-way classification and which is neither too domain specific (biased) nor too general (cannot adapt to within domain word-senses), as well as an ontology that can capture the hierarchical organization of topic in a more flexible, less arbitrary way; the patterns we see here could also be applied to identify new nodes that should be placed into the ontology.

One parallel goal is to move beyond reviews, applying this kind of analysis to blogs or other social media: are the patterns of topic and sentiment due to the constraints of (text) genre, or are they reflected in a wider variety of documents? We are also interested in patterns related to individual authors: does an author tend to organize multiple reviews/post in the same way? Do certain authors avoid "standard" organization patterns? Do some show a consistent positive or negative bias? What kind of author is well-regarded in the community? Finally, it would be interesting to investigate how much neutral support tends to be provided for various aspects, and whether more discussed aspects tend to be more central to overall opinion (our preliminary results here suggest that they are), and whether some kind of weighting is appropriate when an aspect is discussed more than expected.

8. ACKNOWLEDGEMENTS

We would like to thank Michael Gamon and Maite Taboada for the use of their sentiment classifiers.

9. REFERENCES

[1] Alina Andreevskaia and Sabine Bergler. When specialists and generalists work together: Domain dependence in sentiment tagging. *In Proceedings of 46th Annual Meeting of the ACL*, pages 290-298, 2008.

[2] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of the RANLP*, 2005.

[3] Heike Bieler, Stefanie Dipper and Manfred Stede. Identifying formal and functional zones in film reviews. In *Proceedings of the 8th SIGDIAL*, pages 75-78, 2007.

[4] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaption for Sentiment Classification. In *Proceedings of the ACL*, 2007.

[5] Julian Brooke. *A Semantic Approach to Automated Text Sentiment Analysis*. Unpublished Master's Thesis, Simon Fraser University, 2009.

[6] Julian Brooke, Milan Tofiloski, and Maite Taboada. Cross-linguistic sentiment analysis: from English to Spanish. In *Proceedings of the RANLP*, 2009.

[7] Jill Burnstien, Daniel Marcu, and Kevin Knight. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing*, 18(1), pages 32-39, 2003.

[8] Guiseppe Carenini, Raymond Ng, and Ed Zwart. Extracting knowledge from evaluative text. In *Proceedings of the Third International Conference on Knowledge Capture*, 2005.

[9] Joseph L Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, pages 378-382, 1971.

[10] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2), pages 110-125, 2006.

[11] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Edinburgh, UK: University of Edinburgh Thesis, 1996.

[12] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, Ming Zhou. Quality product review detection in opinion summarization. In *Proceedings of the EMNLP*, pages 334-342, 2007.

[13] Yi Mao and Guy Lebanon. Isotonic Conditional Random Fields and Local Sentiment Flow. *In Advances in Neural Information Processing Systems*, 2007

[14] William C. Mann and Sandra A. Thompson. Towards a functional theory of text organization. *Text*, 8, pages 243-281.

[15] James R. Martin and Peter White. *The Language of Evaluation*. New York: Palgrave, 2005.

[16] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and Chengxiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[17] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), pages 21-48, 1991.

[18] Bo Pang, Lillian Lee, and Shvakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the EMNLP*, 2002.

[19] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimal cuts. In *Proceedings of the ACL*, 2004.

[20] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*. J.G. Shanahan, Y. Qu, and J. Wiebe, Eds., Springer: Dordecht, pages 1-10, 2006.

[21] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the EMNLP*, 2005.

[22] Heather Stark. What do paragraph markings do? *Discourse Processes*, 11, pages 275-303, 1998.

[23] Maite Taboada and Jack Grieve. Analyzing appraisal automatically. AAAI Technical Report SS-04-07, pages 158-161, 2004.

[24] Maite Taboada, Julian Brooke, and Manifred Stede. Genre-based paragraph classification for sentiment analysis. In *Proceedings of 10th SIGDIAL*, 2009.

[25] Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the ACL*, 2008.

[26] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.In *Proceedings of the ACL*, 2002.

[27] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*, 2005.