Emotion Arcs of Student Narratives

Swapna Somasundaran, Xianyang Chen, and Michael Flor
Educational Testing Service
660 Rosedale Road, Princeton
NJ 08541, USA
{ssomasundaran,xchen,mflor}@ets.org

Abstract

This paper studies emotion arcs in student narratives. We construct emotion arcs based on event affect and implied sentiments, which correspond to plot elements in the story. We show that student narratives can show elements of plot structure in their emotion arcs and that properties of these arcs can be useful indicators of narrative quality. We build a system and perform analysis to show that our arc-based features are complementary to previously studied sentiment features in this area.

1 Introduction

This work deals with the study of emotion arcs in student narratives. Plots (Lehnert, 1981) and emotion arcs (Vonnegut, 1995; Reagan et al., 2016; Chu and Roy, 2017) form the foundations of story-telling.

Story-telling is an important literacy skill. Children are taught to understand and write narratives in school, and literacy standards[1] require students to write increasingly competent narratives.

While researchers have already introduced analysis of plots in well-written stories and emotion arcs in novels, not much attention has been paid to these phenomena in narratives written by novices. In this work we study student narratives along the dimension of emotion arcs. We show that even novice writing can show plot elements. We then investigate if the quality of narratives can be determined by measuring properties of the emotion arcs.

There has been work investigating scoring of student narratives (Somasundaran et al., 2018). However, previous focus has been on other aspects of the narrative, such as event progression, organization, vividness, detailing and subjectivity. We investigate if emotion arc characteristics can help to improve automated narrative scoring systems.

Plot analyses and research on constructing shapes of stories have considered the general sentiment or affect present in the text. In our work we focus on a specific type of sentiment/affect that we believe is closer to plot structure: events that produce good/bad effect, affective events and sentiment connotations. We show that while there is overlap with subjectivity and sentiment, our approach captures a different dimension of narrative quality.

We believe that our work advances the understanding of plots elements in narratives written by novices. Our work connects evidence of plots elements to the quality of narratives as judged by human raters using standard scoring rubrics. Specifically, the contributions of our work can be summarized as follows: 1. We show that emotion arcs can be seen even in simple narratives written by novice writers. 2. Our experiments show that emotion arc properties are indicative of the quality of narratives. They are related to other sentiment factors in narratives, but are distinct in what they capture about the narrative quality. 3. We show that encoding emotion arc characteristics help to improve narrative essay scoring systems.

2 Narrative Data

As our focus is student narratives, we use the annotated narrative dataset from Somasundaran et al. (2018). The data comprises of 942 narrative essays written by school students from the Criterion® program[2]. Criterion is an online writing evaluation service from Educational Testing Service[3]. It is a web-based, instructor-led writing tool that helps students plan, write and revise their essays. Nar-

[1] http://www.corestandards.org/ELA-Literacy/W/11-12

[2] https://criterion.ets.org/criterion

[3] www.ets.org

Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events, pages 97–107

rative essays in this dataset belong to writers from three grade levels: grades 7, 10 and 12. Each essay is in response to one of 18 story-telling prompts; prompts belong to topics related to personal experiences, hypothetical situations, and fictional stories. Given below are example prompts:

[Personal Experience] There are moments in everyone's lives when they feel pride and accomplishment after completing a challenging task. These moments can happen in the classroom, on the field, or in their personal lives. Write a story about one of your proudest moments.

[Hypothetical Situation] Pretend that one morning you wake up and find out that you've become your teacher for a day! What happened? What do you do? Do you learn anything? Write a story about what happens. Use your imagination!

[Fictional Story] While some well-loved films feature sequels, many do not. These movies can leave the audience wanting to know more about the plot and characters they've enjoyed. Is there a film you've wanted to continue past the ending? Write a synopsis of your own "sequel" to a beloved movie using the same characters and settings as the real film. Remember to include a summary of the previous title and plot, as well as specific new details to draw the reader into your continuation of the movie.

The average essay length in the data is 320 words, with a range of 3 to 1310 words and a standard deviation of 195. The rubric used for scoring the essays was created by education experts and teachers. It defines a separate score (0-4) each for essay organization and essay development. The dataset also provides a *Narrative Score* for each essay, which is the sum of the organization and development scores. The score is an integer value from 0 to 8, with 8 corresponding to perfect organization and development of the narrative. The human inter-annotator agreement for the narrative quality score is 0.76 QWK[4]. We use this score for our work. We refer the reader to the original paper for details on the data, rubrics and annotation.

3 Emotion Arc

An emotion arc involves the plotting or tracking of sentiment valence of some form along the time axis (Vonnegut, 1995; Reagan et al., 2016; Chu and Roy, 2017; Del Vecchio et al., 2018). However, we observed that sentiment words and phrases occurring in narratives serve different *purposes*, such as describing character and settings, embellishing the story, advancing the plot, etc. For example, sentiment words may be used to describe a scene (e.g. "beautiful house"), a character (e.g. "smart girl"), a character's private state (e.g. "Peter thought that was foolish") or emotions (e.g. "Sally was furious").

In our work, we are primarily interested in sentiments and emotions as they relate to the plot. Thus our focus is mainly on events and implicit sentiments. Events are the core building blocks of narratives, and positive and negative events are closely tied to plot progression. This intuition is in line with Lehnert's work on plot units (Lehnert, 1981), which also focuses on positive/negative events (called *events that please*, and *events that displease*). Additionally, much of the plot movement is brought about by elements that have implicit sentiment value. For example, if "A kills B" in a story, it indicates an objective event on the surface, but denotes a conflict (or resolution, depending on whether B is an antagonist) in the story. Given this focus, our emotion arcs are constructed based on the following phenomena that have been previously developed for other purposes in computational linguistics:

Good-for and bad-for events: Good-for and Bad-for events, also known as benefactive and malefactive events, positively/negatively affect the entities on which they act (Deng et al., 2013). These events indicate someone (or something) doing something that affects someone (or something) in a positive or negative manner. In the context of stories, we hypothesize that such events can indicate elements of a plot, such as conflict, resolution and goal achievement.

Affective events: These are events that affect an experiencer in positive or negative ways (Ding and Riloff, 2016) even though they do not, in their surface form, hold a valence. The events are implicitly affective based on the human knowledge of the event itself, such as going on a vacation or breaking a record.

Sentiment Connotation : These are words that imply a positive or negative sentiment even though they appear objective on the surface (Feng et al., 2013). For example, a gun-shot invariably indicates a conflict in the plot, even though it is objective on the surface.

3.1 Constructing Emotion Arcs

In order to construct the emotion arcs, we first extract the elements of interest described above. For

[4]Quadratic Weighted Kappa (Cohen, 1968) is a standard metric in essay evaluation

this we use the EffectWordNet (Choi and Wiebe, 2014) for extracting good-for/bad-for events, event polarity lexicon (Ding and Riloff, 2018) for extracting affective events, and a connotation lexicon (Feng et al., 2013) for extracting sentiment connotation words. Once the elements are extracted for each sentence, they are aggregated to obtain a valence-token offset plot with a sliding window. This process is detailed below:

Preprocessing For a given student essay, we first get the tokenization, part-of-speech tags and dependency parse of each sentence with ZPar[5]. Then we lemmatize the words with NLTK[6].

Good-for/Bad-for Event extraction The EffectWordnet lexicon is a subset of WordNet, with an extra effect polarity annotation for every synset. The effects are either positive, negative or neutral. We pick out all the verbs by POS tags and exclude the stopwords. Then for each verb, we look up its synset(s) in WordNet, and if the synsets are covered in EffectWordnet, we look up its effect polarity. One verb can have multiple senses, and thus multiple synsets in WordNet, with potentially contradicting effect polarities. Here, we take the majority voting approach. For example, if a verb has 3 positive senses, 1 negative sense and 2 neutral senses, we treat it as having a positive effect.

Affective Event extraction The Affective Event lexicon is a mapping from event templates to their polarities. An event template is a verb frame, with optional subject/object/prepositional phrase contexts. For example, *@I@,love,@my@ partner*. We pick out verbs from the sentences by their POS tag, then find out their subject/object by dependency parse, and match with the lexicon.

Sentiment Connotation extraction The Connotation lexicon is a mapping from verbs/nouns/adjectives to their connotation polarities. We simply traverse through all tokens with relevant POS tag, lemmatize them and look up their connotation polarity from the lexicon.

Arc Generation After the above extraction steps we associate every token with a set of extracted polarities. We quantify each token by the following rules: a positive polarity equals +1, a negative polarity equals -1, and a neutral polarity equals 0. The score of a token is the sum of all its associated polarities from different sources. If a token has no associated polarity, it's score is 0. Once the sentiment score for each token is determined, we use a sliding window to slide over the whole narrative, moving by one token at a time, and aggregate the scores within the window. The scores are weighted with Gaussian distribution, with the center of the sliding window being the mean of distribution, and 1/4 of window size as standard derivation. We use a fixed window size of 50, and essays shorter than that are dropped (25 out of 942 total essays). We plot the aggregated scores against the sliding window position, and smooth it with the Savitsky-Golay filter[7] to fit a smooth curve over the narrative. As will be detailed in Section 4, this smoothing is very important for feature extraction on the arcs.

3.2 Emotion Arcs in Student Narratives

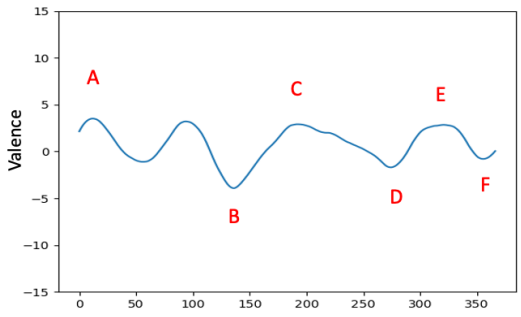

Figure 1: Emotion arc for a narrative on *Proudest Moment*. The Y-axis represents positive/negative valence while the story time-line is along the X-axis

Figure 1 shows the emotion arc constructed for a first person narrative describing The Proudest Moment (an essay written in response to the prompt "Write a story about one of your proudest moments."). In this narrative, the writer talks about her tryout for a marching band performance. The narrator begins with a statement that qualifying for the marching band was her proudest moment. She describes "flagline", the marching band ("Flagline is a group of 10 to 30 girls and they perform in costumes that show school colors."). This story setup and the writer's aspiration is seen in the region (A) of Figure 1. The narrator then goes on to describe how her friends and family thought she could not do it ("My family and friends didn't take me seriously.") and how that created self-doubt ("I started

[5] https://www.sutd.edu.sg/cmsresource/faculty/yuezhang/zpar.html

[6] https://www.nltk.org/

[7] https://plotly.com/python/smoothing/

doubting myself."). This conflict in the plot is evidenced as a dip into negative valence in region (B) in the figure. Then she went to her grandmother for advice ("She gave me the best advice.", "...god will always answer your prayers.). This is evidenced in part (C) where the emotion arc peaks on the positive side of the graph. When the narrator finally goes to the tryout, she is extremely nervous ("It felt like my knees were going to fall off", "It felt like I was going to faint ."). Corresponding to this suspense in the plot, the arc dips again at (D). Finally her name gets called by the judges, and she is extremely elated ("... hugging everyone around me ", "It was the proudest day of my life."). This happy resolution is corroborated by the arc at region (E). The narrative concludes with a reflective note that she remembers this day later on in life when she faces a tough situation (F).

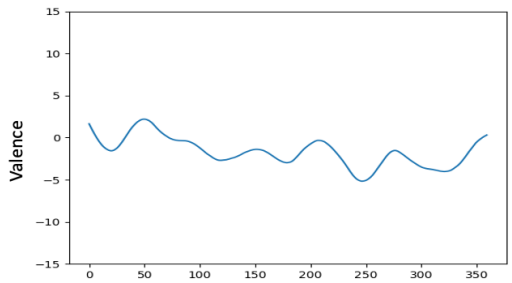

Figure 2: Emotion arc for a narrative on *Movie Sequel* The Y-axis represents positive/negative valence while the story time-line is along the X-axis

Figure 2 shows the emotion arc for a third person fictional narrative. This essay was written in response to the prompt *Write a synopsis of your own "sequel" to a beloved movie using the same characters and settings as the real film.* The student chose to write a sequel to the movie "The Grey" starring Liam Neeson. Similar to the original movie, the sequel too is a survival thriller and follows Ottway, the character from the original. The story is full of adversities, such as vicious wolves, harsh climate of the Alaskan wilderness and starvation. The emotion arc correspondingly, remains in the negative valence area, with the small wave-like fluctuations to the neutral (relatively less negative) side for small victories. Towards the end of the climax, the protagonist is critically wounded. In the last sentence, the narrative says that a rescue team is coming his way, indicating a positive ending.

We noticed that short narrative essays written by school students show variations in the emotion arc corresponding to elements of a plot such as setting, conflict, suspense, resolution and reflection. Overall, emotion arcs vary across narrative genres, topics, and even within a topic due to creative variety. Nevertheless, the basic elements of a good story can still be found across all narrative types. For example, a plot almost always requires an emotional variation and an effective narrative will have some form of conflict or dip in emotional valence.

4 Relationship to Narrative Quality

We saw in the previous section that well-written student narratives generally tend to have emotion arcs corresponding to plot elements. The next question is whether these elements can be indicators of narrative quality score as defined by standardized essay scoring rubrics. In order to answer this, we study the relationship of narrative quality scores to properties of the emotion arcs. Note that the narrative quality score is a function of a number of factors, such as organization of the story, effective use of transitioning, clear opening and closing, vivid description, character development, use of dialog, event sequencing, effective use of figurative language and other narrative techniques. Hence we expect presence of plot element in a story to be just a component that would contribute to the determination of the score.

While individual creativity makes it difficult to directly equate emotion arcs to scores, we can extract features that represent arc characteristics. We extract the following features from the arcs. For the ease of explanation, we denote the arc value at position i as $d(i)$. We define *local maximum* as positions where $d(i-1) < d(i) > d(i+1)$, and *local minimum* as points where $d(i-1) > d(i) < d(i+1)$. The *slope* at each point is calculated as $d(i) - d(i-1)$.

1. **Max Peak**: We find the "peaks" in the arc by looking for local maximums. And we choose the maximum among the local maximums as Max Peak.

2. **Second Max Peak**: Similar to Max Peak, but the second greatest one of the local maximums.

3. **Min Valley**: The minimum of local minimums.

4. **Second Min Valley**: The second least local minimum.

5. **Number of Peaks**: Number of identified local maximums.

6. **Number of Valleys** : Number of identified local minimums.

7. **Positive Slope**: The slope where the arc is most steep and going upward, i.e. $\max_i d(i) - d(i-1)$.

8. **Negative slope**: The slope where the arc is most steep and going downward, i.e. $\min_i d(i) - d(i-1)$.

The plots generated by sliding windows are noisy and contain lots of spurious local maximums/minimums. So we apply Savitsky-Golay filter to smooth the high-frequency variations in the plots, as we are not interested in the fine-grained perturbations, but in how valence emerges or drops in at a coarser granularity.

The maximum peak and minimum valley capture the height of happiness and depth of despair in the story. The second max peak (and min valley) capture the second highest points. Presumably, narratives with non-trivial story-lines will show multiple significant peaks and valleys. The number of peaks and valleys try to capture the emotional variance in the story. The positive and negative slopes try to capture the emotional pace of the story.

4.1 Correlation with Narrative Quality

Using the scored essays, we compute correlation (Pearson's r) for each of our features to the Narrative Quality score. Previous studies on essay scoring (Chodorow and Burstein, 2004) have found that essay length is strongly correlated with its score. Thus, for each of our features, we calculate correlation with score after accounting for length, in order to see its effect on the narrative quality independent of essay length.

Table 1 presents the Pearson's correlation r (sorted in ascending order) and partial correlation with narrative score. Observe that the number of peaks and valleys are strongly correlated with score – having more peaks and valleys is related to higher the score. However, such stories are also relatively longer, and hence correlation drops dramatically when the length factor is removed. The slope-based features (Positive Slope and Negative Slope) show moderate correlation with score. The features related to negative dips in the story (Min Valley and Second Min Valley) show moderately strong correlation with score, and have a relatively smaller drop after accounting for length. This indicates that elements corresponding to strong adversities are effective narrative techniques even in short stories.

Feature	**Pearson's r**	r After controlling length
Max Peak	0.151	0.006
Second Max Peak	0.184	-0.003
Positive Slope	0.307	0.155
Negative Slope	0.312	0.136
Min Valley	0.412	0.202
Second Min Valley	0.431	0.193
Num of Peaks	0.538	-0.016
Num of Valleys	0.541	-0.011

Table 1: Correlation (Pearson's r) of each feature with score.

4.2 Narrative Quality Prediction

The next question we explore is if and by how much the emotion arc features, individually or as a group, are useful for predicting narrative quality. Given that there has been previous work on developing narrative quality features, our focus is on how much the emotion arc features can help to improve a system based on previous narrative features. For this, we closely follow the procedure from (Somasundaran et al., 2018): we build a Linear regression model using scikit-learn toolkit (Pedregosa et al., 2011), with 10-fold cross-validation. Trimming of the predicted output is performed; that is, if the predicted score was above the max score (8), or below the min score (0), it is assigned the max or the min score, respectively. Bootstrapping experiments (Berg-Kirkpatrick et al., 2012; Efron and Tibshirani, 1994) were performed to test for statistical significance. We used 10,000 bootstrap samples.

The system using the *best narrative features* from previous work is our baseline. Thus the baseline comprises of the following features: Details+ Modal+ Pronoun+ Content+ Graph+ Statives+ Subjectivity+ Transition + Quote[8]. We build prediction models by (1) adding one emotion arc-based fea-

[8]This feature, capturing the presence of dialog or air quotes, was added by the authors after the publication of their paper. It produces a small improvement in performance.

ture at a time to the baseline (2) adding all of our features to baseline

Feature	QWK
Baseline	0.656
+ Negative Slope	0.652
+ Max Peak	0.657
+ Second Max Peak	0.657
+ Number of Valleys	0.661
+ Second Min Valley	0.663
+ Number of Peaks	0.666 *
+ Positive Slope	0.667 **
+ Min Valley	0.669 **
+ All	0.668

Table 2: Performance of the system when new features are added to the baseline.
$*$ indicates $p < 0.1$; $**$ = $p < 0.05$

Table 2 reports the performance of the resulting systems sorted in ascending order (for individual feature additions). Features corresponding to the positive emotional peaks in the story (Max Peak and Second Max Peak) add only minor improvements. Features corresponding to negative valence (Min Valley, Second Min Valley) help to improve the performance, indicating that detecting negative dips can improve the reliability of scoring. With respect to pacing, moving from a negative point to a positive point (Positive slope) seems to be indicative of narrative quality. However, adding all features together seems to produce no improvement indicating that while some individual features show promise, others tend to bring the performance down.

We performed a detailed ablation study (8 features resulted to 256 experiments) to find the subset of features that can be used together. The resulting feature combination that gave the best performance was [Max Peak + Min Valley + Second Max Peak + Number of Peaks + Positive Slope] and had a QWK value of 0.676.

4.3 Correlation with Other Sentiment Features

Somasundaran et al. (2018) have explored subjectivity-based features for predicting narrative score. Their motivation was to capture evaluative and subjective language that is used to describe characters, situations, and characters' private states (Wiebe, 1994). While our features are also sentiment-based, we believe that our arc-based features capture a different dimension of the narrative and are complementary in nature.

In order to investigate this, we compared our features with the following subjectivity-based features in the baseline system: count of MPQA (Wilson et al., 2005) polar words (CMP), count of MPQA neutral words (CMN), presence of MPQA neutral words (PMN), presence of MPQA polar words (PMP), count of ASSESS (Beigman Klebanov et al., 2012) polar words (CAP), count of unique ASSESS polar words (UAP).

Table 3 presents the Pearson's correlation r between our features and subjectivity features from the baseline system. Values of r greater than 0.5 are shown in **bold**.

As expected, there is strong correlation between the emotion arc features and subjectivity/sentiment. We believe that this is because (1) There are events that are *also* clearly sentiment-bearing words (e.g., "failing"), (2) Good/bad events and feeling about them would co-occur in the story. For example, if something adverse happens to a character, he might feel bad about it. (3) It is very likely that there is overlap between the lexical resources we use for constructing emotion arcs and the subjectivity/sentiment features.

However, the correlation values also indicate that there is some separation between our plot-motivated features and the subjectivity features – except for the high correlation between count-based subjectivity features and number-based arc features (all of which also correlate with length), the rest have $r < 0.5$.

5 Related Work

Narratives can be analyzed along many different dimensions, such as sentiment, emotion, plot, characters, engagement, creativity, and success of stories. Computational linguistic analyses started with shorter texts, concentrating mostly on fables, folk stories and fairy tales. In the last decade they fully embraced analysis of full length novels and movie scripts.

Several studies focused on character traits and personas in stories. Elsner (2012) proposed a rich representation of story-characters for summarizing and representing novels. Bamman et al. (2014) automatically inferred character types in English novels.Valls-Vargas et al. (2014) extracted character roles from Russian folk tales, based on character actions. Chaturvedi et al. (2015) analyzed short sto-

Emotion arc feature	CMP	CMN	PMN	PMP	CAP	UAP
Max Peak	0.34	0.19	0.07	0.05	0.33	0.34
Second Max Peak	0.40	0.23	0.06	0.03	0.39	0.40
Min Valley	0.38	0.31	0.13	0.00	0.41	0.43
Second Min Valley	0.43	0.35	0.08	0.00	0.46	0.47
Num of peaks	**0.73**	**0.61**	0.19	0.04	**0.78**	**0.73**
Num of Valleys	**0.73**	**0.60**	0.19	0.05	**0.78**	**0.73**
Positive Slope	0.38	0.23	0.14	0.08	0.39	0.42
Negative Slope	0.42	0.27	0.12	0.10	0.41	0.45

Table 3: Correlation (Pearson's r) of each feature with previously explored subjectivity features:
count of MPQA polar words (CMP), count of MPQA neutral words (CMN), presence of MPQA neutral words (PMN), presence of MPQA polar words (PMP), count of ASSESS polar words (CAP), count of unique ASSESS polar words (UAP)

ries for characters' desires and desire fulfillment.

Researchers have also studied social networks and have modeled relationships in stories (Elson et al., 2010; Celikyilmaz et al., 2010; Agarwal et al., 2013). Iyyer et al. (2016); Chaturvedi et al. (2016) modeled character inter-relations and their development in novels. Evolving relations were represented as *relationship sequences/trajectories* and learned using structured prediction techniques.

Ouyang and McKeown (2015) analyzed personal narratives from a blogging platform for automatic detection of *turning points* in stories. Papalampidi et al. (2019) presented the task of detecting turning points in movie scripts, as a particular way for analyzing narrative structure. They used neural network models for automatically detecting sequences of major explicit events in stories.

Sentiment analysis has been employed for narrative analysis in many studies. Goyal et al. (2010a,b) analyzed Aesop's fables, producing automatic plot-unit representations (Lehnert, 1981), using task-specific knowledge base of affect. Several studies focused on annotation of folk stories and fairy tales for emotions (Francisco et al., 2012; Volkova et al., 2010; Alm and Sproat, 2005). Alm et al. (2005) described a machine learning approach for multi-class classification of sentences for their emotional content.

Piper and So (2015) used a sentiment lexicon and compared the proportion of sentiment words between several groups of novels. They found that, on average, 19th century novels have a larger proportion of sentiment words (7%) than modern novels (about 5.5%). Bostan and Klinger (2018) presented a survey of recent corpora annotated for emotion classification in text, with a variety of classification schemata. Liu et al. (2019) present a new dataset of classic and modern novels, with passages (sections of 40 to 200 words) manually annotated for emotion classes based on Plutchik's eight basic emotions. The data was used for training and evaluating Deep Learning architectures for emotion classification. Kim and Klinger (2019, 2018) presented a novel challenge and datasets for affect and emotion detection in text, calling it *emotional relationships* classification: which character feels which emotion to which character.

The idea that, in stories, emotion and sentiment is not static, but changes dynamically, is an old one. It is often presented as practical advice in writing guides for aspiring screenwriters and novelists (McKee, 1997). The idea of actually charting the emotional progression of scenes and stories (*emotion arcs*) is attributed to Kurt Vonnegut (see e.g. Vonnegut (1995, 2004), described by Jockers (2014) and Del Vecchio et al. (2018)).

The annotation study of Alm and Sproat (2005) was one of the earliest computational studies that considered a notion of *emotional trajectory* - plotting the emotional values for sequences of text segments in 22 Grimms' fairy tales. Reagan et al. (2016) used sentiment analysis to generate emotional profiles for full-length English novels. They implemented the notion of emotion arcs, tracking the level of emotion-laden content through consecutive segments (10K-long word segments) of literary works (from Project Gutenberg), using a lexicon of words with ratings on a single positive-negative scale - a sentiment polarity lexicon. They found that large-scale arcs cluster into six common arc shapes. A similar approach was described by Jockers (2014) and Gao et al. (2016).

Del Vecchio et al. (2018) used a similar technique to track emotion arcs in about 6000 movie scripts. That study also found that large-scale arcs cluster into six common shapes. The study went further, relating the arc shapes to movie success, using movie gross revenue as success indicator.

Chu and Roy (2017) performed a multi-modal analysis of emotion arcs. They used neural network models to construct emotional arc representations from movie clips, audio clips, images and two-word image captions (the latter analyzed with SentiWordNet lexicon, (Baccianella et al., 2010)). They applied clustering to find major groupings of emotional arcs. They evaluated their models by predicting viewer engagement with short online videos (measured as number of comments the videos received).

Kim et al. (2017) used lexical expressions of emotion for genre-classification of whole novels. They extended the *emotion arc* approach, by using eight fundamental emotions defined by Plutchik (2001), instead of a single sentiment valence dimension. Thus, their tracking reflects the development of each of the eight emotions throughout the time course of the narrative. Overall, this method contributes to significant improvement of genre classification over a strong lexical baseline.

6 Discussion

To the best of our knowledge, our work is a first attempt in exploring emotion arcs and plot elements in student writing via event affect and implied sentiment. The exploration is by no means complete. In this section we discuss the rationale of our choices, some limitations of the current study and challenges in this task.

In order to construct emotion arcs in Section 3 we relied on a number of resources and made a number of choices, which has influenced the precision of the resulting arcs. First, we used (semi) automatically created lexicons, and these have issues with both noise and coverage. We did not combine the lexicons or remove duplicates between lexicons. This could have led to a single instance of an easily recognizable emotional element in a sentence being counted more strongly than it otherwise should have. Contextual polarity resolution was not performed, which could have influenced the determination of the sentence-level valence. Finally, when employing the EffectWordnet, we simply used the most frequent sense instead of performing complete word sense disambiguation. Our curve-fitting function might have dampened some of the spurious errors, but there is obviously scope for improvement.

In Section 4 we sampled arc properties and employed them as features. Our choices were driven by the requirements that the properties should be efficiently extracted, and be generalizable across narrative sub-genres and writing proficiency. More constrained environments/applications could afford closer modeling of arcs (e.g. finding components of the curve equation).

In our work, we have used the simplifying assumption that there are emotional correlates to plot elements, and by tracking the emotion arc, we are able to capture some elements of the plot. While this assumption may hold for simple short stories, it is likely to collapse for stories from creative authors. Accomplished authors can create tension and resolution without emotional accompaniment.

Finally, it is important to note that plot and components of a plot, such as rising action, conflict, falling action, resolution, etc., while easily discernible to human readers, pose significant challenges for a machine. World knowledge, human experiences, interpretation of motivations, inferences of human actions and context of the story, play an important part in the recognition. Consequently, automated methods too will need to look beyond words, sentences and paragraphs.

7 Conclusion

In this work, we studied emotion arcs in student narratives, and explored ways to harness them for automatically determining narrative quality.

We showed that emotion arcs are manifest in student writing and intuitively correspond to plot elements. Our analyses showed that simple arc properties correlate with narrative quality score, and that the features derived from the arcs are similar to, yet distinguished from previously explored subjectivity features. We built scoring systems and showed that adding our arc-based features have the potential to improve the scoring performance.

Our future work will include addressing the limitations and challenges discussed in Section 6.

References

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from

literary text: A case study on Alice in Wonderland. In *IJCNLP*, pages 1202–1208.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, pages 668–674. Springer.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 370–379, Baltimore, MA, USA.

Beata Beigman Klebanov, Jill Burstein, Nitin Madnani, Adam Faulkner, and Joel Tetreault. 2012. Building subjectivity lexicon (s) from scratch for essay data. *Computational Linguistics and Intelligent Text Processing*, pages 591–602.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics.

Asli Celikyilmaz, Dilek Hakkani-Tur, Hua He, Greg Kondrak, and Denilson Barbosa. 2010. The actor-topic model for extracting social networks in literary narrative. In *NIPS Workshop: Machine Learning for Social Computing*.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daume III. 2015. Ask, and shall you receive?: Understanding desire fulfillment in natural language text. *arXiv preprint arXiv:1511.09460*.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daumé III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2704–2710. Association for the Advancement of Artificial Intelligence.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191.

Eric Chu and Deb Roy. 2017. Audio-visual sentiment analysis for learning emotional arcs in movies. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 829–834. IEEE.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213.

Marco Del Vecchio, Alexander Kharlamov, Glenn Parry, and Ganna Pogrebna. 2018. The data science of hollywood: Using emotional arcs of movies to drive business model innovation in entertainment industries. *Available at SSRN 3198315*.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125.

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Haibo Ding and Ellen Riloff. 2018. Weakly Supervised Induction of Affective Events by Optimizing Semantic Consistency. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Micha Elsner. 2012. Character-based kernels for novelistic plot structure. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 634–644. Association for Computational Linguistics.

David K. Elson, Nicholas Dames, and Kathleen R. McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147. Association for Computational Linguistics.

Song Feng, Jun Sak Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers), Sofia, Bulgaria. Association for Computational Linguistics.

Virginia Francisco, Raquel Hervás, Federico Peinado, and Pablo Gervás. 2012. Emotales: creating a corpus of folk tales with emotional annotations. *Language Resources and Evaluation*, 46(3):341–381.

Jianbo Gao, Matthew L. Jockers, John Laudun, and Timothy Tangherlini. 2016. A multiscale theory for the dynamical evolution of sentiment in novels. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–4.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010a. Automatically producing plot unit representations for narrative text. In *Empirical Methods in Natural Language Processing (EMNLP)*, Boston, MA.

Amit Goyal, Ellen Riloff, Hal Daumé III, and Nathan Gilbert. 2010b. Toward plot units: Automatic affect state analysis. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 17–25. Association for Computational Linguistics.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Matthew L. Jockers. 2014. A novel method for detecting plot.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. Frowning Frodo, wincing Leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653, Minneapolis, Minnesota. Association for Computational Linguistics.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada. Association for Computational Linguistics.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science*, 5(4):293–331.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019. DENS: A dataset for multi-class emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6293–6298, Hong Kong, China. Association for Computational Linguistics.

Robert McKee. 1997. *Story: Substance, Structure, Style and the Principles of Screenwriting*. Regan Books, Harper-Collins Publishers.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Andrew Piper and Richard Jean So. 2015. Quantifying the weepy bestseller. *New Republic*.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontanón. 2014. Toward automatic role identification in unannotated folk tales. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 188–194. AAAI Press.

Ekaterina P. Volkova, Betty J. Mohler, Detmar Meurers, Dale Gerdemann1 Heinrich H, and Bülthoff. 2010. Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In

Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 98–106. Association for Computational Linguistics.

Kurt Vonnegut. 1995. Kurt Vonnegut on the Shapes of Stories. YouTube.

Kurt Vonnegut. 2004. Kurt Vonnegut Lecture. YouTube.

Janyce M. Wiebe. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.