Towards Evaluating Narrative Quality In Student Writing

Swapna Somasundaran[1], **Michael Flor**[1], **Martin Chodorow**[2]
Hillary Molloy[3] **Binod Gyawali**[1] **Laura McCulla**[1]
[1]Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
[2]Hunter College and the Graduate Center, CUNY, New York, NY 10065, USA
[3]Educational Testing Service, 90 New Montgomery Street, San Francisco, CA 94105, USA
{ssomasundaran,mflor,hmolloy, bgyawali,LMcCulla}@ets.org
martin.chodorow@hunter.cuny.edu

Abstract

This work lays the foundation for automated assessments of narrative quality in student writing. We first manually score essays for narrative-relevant traits and sub-traits, and measure inter-annotator agreement. We then explore linguistic features that are indicative of good narrative writing and use them to build an automated scoring system. Experiments show that our features are more effective in scoring specific aspects of narrative quality than a state-of-the-art feature set.

1 Introduction

Narrative, which includes personal experiences and stories, real or imagined, is a medium of expression that is used from the very early stages of a child's life. Narratives are also employed in various capacities in school instruction and assessment. For example, the Common Core State Standards, an educational initiative in the United States that details requirements for student knowledge in grades K-12, employs literature/narratives as one of its three language arts genres. With the increased focus on automated evaluation of student writing in educational settings (Adams, 2014), automated methods for evaluating narrative essays at scale are becoming increasingly important.

Automated scoring of narrative essays is a challenging area, and one that has not been explored extensively in NLP research. Previous work on automated essay scoring has focused on informational, argumentative, persuasive and source-based writing constructs (Stab and Gurevych, 2017; Nguyen and Litman, 2016; Farra et al., 2015; Somasundaran et al., 2014; Beigman Klebanov et al., 2014; Shermis and Burstein, 2013). Similarly, operational essay scoring engines (Attali and Burstein, 2006; Elliot, 2003) are geared towards evaluating language proficiency in general. In this work, we lay the groundwork and present the first results for automated scoring of narrative essays, focusing on narrative quality.

One of the challenges in narrative quality analysis is the scarcity of scored essays in this genre. We describe a detailed manual annotation study on scoring student essays along multiple dimensions of narrative quality, such as narrative development and narrative organization. Using a scoring rubric adapted from the U.S. Common Core State Standards, we annotated 942 essays written for 18 different essay-prompts by students from three different grade levels. This data set provides a variety of story types and language proficiency levels. We measured inter-annotator agreement to understand reliability of scoring stories for traits (e.g., development) as well as sub-traits (e.g., plot development and the use of narrative techniques).

A number of techniques for writing good stories are targeted by the scoring rubrics. We implemented a system for automatically scoring different traits of narratives, using linguistic features that capture some of those techniques. We investigated the effectiveness of each feature for scoring narrative traits and analyzed the results to identify sources of errors.

The main contributions of this work are as follows: (1) To the best of our knowledge, this is the first detailed annotation study on scoring narrative essays for different aspects of narrative quality.

Transactions of the Association for Computational Linguistics, vol. 6, pp. 91–106, 2018. Action Editor: Alexander Clark.
Submission batch: 7/2017; Revision batch: 10/2017; Published 2/2018.

(2) We present an automated system for scoring narrative quality, with linguistic features specific to encoding aspects of good story-telling. This system outperforms a state-of-the-art essay-scoring system. (3) We present analyses of trait and overall scoring of narrative essays, which provide insights into the aspects of narratives that are easy/difficult for humans and machines to evaluate.

2 Related Work

2.1 Narrative assessments

Researchers have approached manual assessments of creative writing in a variety of ways. The "consensual assessment technique" (Amabile, 1982; Broekkamp et al., 2009) evaluates students' creative writing on criteria such as creativity, originality and technical quality. Consensus scoring is used, but the genre is considered to be too subjective for close agreement between scorers.

Story-telling in children has been studied and evaluated using a number of techniques. For example, the Test of Narrative Language (Gillam and Pearson, 2004) is a standardized, picture-based, norm-referenced measure of narrative ability, used to identify language disabilities. Stein and Glenn (1979) used a story-schema approach to evaluate story recall in school children. Miller and Chapman (1985) adapted it to score story re-telling, mainly for clinical purposes. Similarly, narrative re-telling is recorded and analyzed for length, syntax, cohesion, and story grammar in the Strong Narrative Assessment Procedure (Strong et al., 1998). The Index of Narrative Complexity (Petersen et al., 2008) scores oral narratives on several dimensions and is used to study the effectiveness of clinical interventions.

Olinghouse and Leaird (2009) used picture-prompts for eliciting narratives from about 200 students at the 2nd and 4th grade levels. The stories were evaluated for organization, development and creative vocabulary, but the study focused on vocabulary characteristics at different grade levels. McKeough et al. (2006) studied 150 student narratives in order to compare talented and average writers.

Halpin and Moore (2006) analyzed students' retelling of exemplar stories. They focused on event extraction, with the final goal of providing advice in an interactive story-telling environment. Passonneau et al. (2007) annotated oral retellings of the same story on three consecutive days in order to study and model children's comprehension.

2.2 Narrative Analysis in Computational Linguistics

Research on narratives in Computational Linguistics has employed fables, fairy tales, and literary texts, aiming at representing, understanding and extracting information, e.g., Charniak (1972). Goyal et al. (2010) analyzed Aesop's fables, producing automatic plot-unit representations (Lehnert, 1981) with a task-specific knowledge base of affect.

Character traits and personas in stories have also been analyzed. For example, Elsner (2012) proposed a rich representation of story-characters for the purpose of summarizing and representing novels. Bamman et al. (2014) automatically inferred latent character types in English novels. Valls-Vargas et al. (2014) extracted characters and roles from Russian folk tales, based on their actions. Chaturvedi et al. (2015) analyzed short stories for characters' desires and built a system to recognize desire fulfillment, using textual entailment.

Researchers have also studied social networks and have modeled relationships in stories (Elson et al., 2010; Celikyilmaz et al., 2010). Agarwal et al. (2013) modeled character interactions from *Alice in Wonderland* for the purpose of social network analysis. Chaturvedi et al. (2016) modeled character relationships in novels, using structured prediction.

Wiebe (1994) proposed a method for tracking psychological points of view in narratives, looking at private states and subjective sentences. Ovesdotter Alm and Sproat (2005) studied emotional sequencing and trajectories in 22 Grimm's fairy tales. Ware et al. (2011) analyzed dimensions of conflict in four simple, constructed stories, with the goal of evaluating story content. Similarly, Swanson et al. (2014) analyzed blog narratives for narrative clause sub-types such as orientation, action and evaluation. Reagan et al. (2016) used sentiment analysis to generate emotional profiles for English novels.

NLP methods have also been used for modeling and understanding narrative structures (Finlayson, 2012; Elson, 2012). See Finlayson (2013) and Mani (2012) for detailed literature surveys.

One important aspect of a narrative is that it

conveys a sequence of events (Fludernik, 2009; Almeida, 1995). Chambers and Jurafsky (2009; 2008) presented techniques for the automatic acquisition of event chains and event schemas (Chambers, 2013), which are related to earlier notions of scripts as prepackaged chunks of knowledge (Schank and Abelson, 1977). This line of research has received a great deal of attention (Nguyen et al., 2015; Balasubramanian et al., 2013; Jans et al., 2012; McIntyre and Lapata, 2010). For narratives, Ouyang and McKeown (2015) focused on automatic detection of compelling events. Bogel et al. (2014) worked on extraction and temporal ordering of events in narratives.

Based on the 'Narrative Cloze Test' (Chambers and Jurafsky, 2008), Mostafazadeh et al. (2016) presented a framework for evaluating story understanding algorithms, the 'Story Cloze Test', whose goal is to predict a held-out continuation of a short story.

Our research differs significantly from previous work. We aim to evaluate, on an integer scale, the *quality* of narratives in student-generated essays. Insights from previous work on narrative analysis can be useful for our purposes if they capture narrative techniques employed by student writers, and if they correlate with scores representing narrative quality. It is still an open question whether an elaborate representation and understanding of the story is needed for evaluating student writing, or whether encoding features that capture different narrative aspects might be sufficient. Further, depending on the type of story, not all aspects of narrative analysis may come into play. For example, plot construction and narrative elements such as conflict may be central to creating a hypothetical story about an antique trunk, but not so much in a personal story about a travel experience. To the best of our knowledge, this work makes a first attempt at investigating the evaluation of narrative quality using automated methods.

2.3 Automated essay scoring

There are a number of automated essay scoring (AES) systems, many of which are used operationally, such as e-rater® (Attali and Burstein, 2006), Intellimetric (Elliot, 2003), the Intelligent Essay Assessor (Landauer et al., 2003) and Project Essay Grade (Page, 1994). However, these previous studies have not been focused on narratives.

In a somewhat related study to this one, Somasundaran et al. (2015) scored oral narratives that were generated by international students in response to a series of pictures. Some of the features used in that study overlap with our work due to the overlap in the genre; however, their focus was on scoring the response for language proficiency. Graph features, which we have used in this work, have been shown to be effective in capturing idea development in essays (Somasundaran et al., 2016). This work also employs graph features, but it is one of the many we explore for encoding the various linguistic phenomena that characterize good narratives.

3 Data

Our data comprises narrative essays written by school students in the Criterion® program[1], an online writing evaluation service from Educational Testing Service. It is a web-based, instructor-led writing tool that helps students plan, write and revise their essays. Narrative essays were obtained from grade levels 7, 10 and 12. Each essay was written in response to one of 18 story-telling prompts related to personal experiences, hypothetical situations, or fictional stories. Below are some example prompts:

[Personal Experience] There are moments in everyone's lives when they feel pride and accomplishment after completing a challenging task. Write a story about your proudest moment.

[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] Throughout the years, many have placed messages in sealed bottles and dropped the bottles into the ocean where they eventually washed up on foreign shores. Occasionally the finder has even contacted the sender. Write a story about finding your own message in a bottle.

The average essay length in our data is 320 words, with a range of 3 to 1310 words and a standard deviation of 195. A sample essay, "Message in a bottle", in response to the fiction story prompt above is presented below:

> *Last year, I went back to my hometown. There was a big beautiful beach on which I had often played as a child. Nevertheless, when I went to the beach, it changed. I looked a great deal of trash, and*

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

many animal disappeared. Without original breathtaking scene, there had been destroyed very well.

All of a sudden, I watched a bottle When I walked on the beach. I opened the bottle with my curiosity. There was a message in the bottle. The message was "Whoever you are, please help this beach. We need more clean beach to survive." I was surprised that this message should be from the sea creature. They need humans' help, or they would die.

Therefore, I persuaded the other people who live there to clean the beach immediately. They all agreed to come and to help those animals. Finally, with a lot of people's help, the beach became beautiful as before. I thought that those who under the sea were very comfortable and happy to live a clean surroundings.

4 Scoring Narrative Essays

Our work focuses on automatically evaluating and scoring the proficiency of narrative construction in student essays.

Therefore, we use a rubric[2] created by education experts and teachers, and presented by Smarter Balanced, an assessment aligned to U.S. State Standards for grades K-12.

4.1 Trait Scoring

The scoring rubric provides guidelines for scoring essays along three traits (dimensions): Purpose/Organization (hereafter, referred to as Organization or Org.), Development/Elaboration (Development or Dev.) and Conventions (or Conv.). Each of the dimensions is described below.

4.1.1 Organization

Organization is concerned with the way a story is arranged in general. It focuses on event coherence, on whether the story has a coherent start and ending, and whether there is a plot to hold all the pieces of the story together. This dimension is judged on a scale of 1-4 integer score points, with 4 being the highest score. The rubric provides the following criteria for an essay of score point 4 in terms of five aspects or sub-traits: *"The organization of the narrative is fully sustained and the focus is clear and maintained throughout: 1. an effective Plot; 2. effectively establishes Character/Setting/POV; 3. consistent use of a variety of Transitioning strategies; 4. natural, logical Sequencing of events; 5. effective Opening/Closing."*

An essay is judged non-scorable if it is insufficient, written in a language other than English, off-topic, or off-purpose. Such essays are assigned a score of 0 in our scheme.

4.1.2 Development

Development focuses on how the story is developed. It evaluates whether the story provides vivid descriptions, and whether there is character development. This dimension is also judged on a scale of 1-4 integer score points, with 4 being the highest score. As in the scoring of Organization, in our scheme, non-scorable essays are assigned a 0 score for Development. The rubric provides the following criteria for an essay of score point 4 in terms of five aspects or sub-traits: *"The narrative provides thorough, effective elaboration using relevant details, dialogue, and/or description: 1. clearly developed Character/Setting/Events; 2. connections made to Source Materials; 3. effective use of a variety of Narrative Techniques; 4. effective use of sensory, concrete, and figurative Language; 5. effective, appropriate Style."*

4.1.3 Conventions

This dimension evaluates the language proficiency, judged on a scale of 1-3 integer score points, with 3 being the highest score. According to the rubrics, the following characterizes an essay of score point 3: *"The response demonstrates an adequate command of conventions: adequate use of correct sentence formation, punctuation, capitalization, grammar usage, and spelling."*

4.2 Sub-trait scoring

As noted above, Organization and Development are each composed of 5 sub-traits. We scored these sub-traits manually using the same 4-point scale as the main trait scores. This yields 10 sub-trait scores in addition to the 3 main trait scores, for a total of 13 manually assigned scores per essay. We produced guidelines and selected a small set of benchmark essays for training two scorers.

[2] https://portal.smarterbalanced.org/library/en/performance-task-writing-rubric-narrative.pdf

4.3 Narrative and Total Scores

Based on the human-assigned trait scores, we derive Narrative and Total composite scores for each essay. The Narrative score for each essay is calculated by summing the Organization and Development trait scores. This gives the essay a Narrative score on an integer scale from 0 to 8. We sum up the three trait scores (Organization + Development + Conventions) to get a Total score on an integer scale from 0 to 11. Even though Narrative and Total composites are not defined separately/independently from their components, they provide us with an estimate of how manual and automated scoring will perform on these data for scenarios where, for example, a single overall score has to be assigned.

5 Annotation and Data Statistics

Two research assistants, both co-authors on the paper but not involved in system development, performed the scoring. Both annotators are native speakers of English with more than four years of linguistic annotation experience. Using the scoring rubric described above, the lead annotator created a guideline and benchmark dataset of 20 essays spanning all score points. This was used for training a second annotator and three researchers (all co-authors on the paper), and the resulting feedback was used to refine the guidelines. Two rounds of training were conducted, with 10 and 20 essays respectively. A score discrepancy of more than one point for any of the traits triggered a discussion in order to bring the scores closer (that is, the scores should only differ by one point). Exact agreement was not sought due to the very subjective nature of judging stories. One of the researchers served as adjudicator for the discussions. No specific training was performed for the sub-traits; disagreements on sub-traits were discussed only within trait-level discussions.

Once the training was completed, a total of 942 essays[3] were scored. Of these, 598 essays were singly scored and 344 essays were double-scored to measure agreement. Scoring of each essay thus involved assigning 13 scores (3 traits + 10 sub-traits) and took approximately 10 to 20 minutes. Table 1 shows the distribution of scores across the score-points for the three traits.[4]

Score	0	1	2	3	4
Org.	40	63	217	381	241
Dev.	40	84	270	319	229
Conv.	-	115	365	462	-

Table 1: Score distributions for traits

5.1 Inter-annotator Agreement

To calculate agreement, we use Quadratic Weighted Kappa (QWK) (Cohen, 1968), a well-established metric in assessment that takes into account agreement due to chance. It is equivalent to a form of intra-class correlation and, in most cases, is comparable to Pearson's r. The QWKs calculated over 344 doubly annotated essays are reported in Table 2. The three main traits are shown in **bold**, the sub-traits are prefixed with a ":", and the composite traits (Narrative and Total) are shown in italics.

Trait:Sub-trait	**QWK**
Organization	0.71
:Plot	0.62
:Characters/Setting/POV	0.65
:Transitioning	0.57
:Sequencing	0.63
:Opening/Closing	0.66
Development	0.73
:Characters/Setting/Events	0.68
:Narrative Techniques	0.64
:Language	0.59
:Source Materials	0.52
:Style	0.58
Convention	0.46
Narrative (Org. + Dev.)	0.76
Total (Org. + Dev. + Conv.)	0.76

Table 2: Inter-annotator agreement

For the Organization and Development traits, which capture the narrative aspects of writing,

[3]For data requests see https://www.ets.org/research/contact/data_requests/.

[4]The "Message in a Bottle" sample essay in Section 3 received scores of Org.:3, Dev.:4, and Conv.:3. The high score for Conventions reflects the rubric's requirement of *adequate* (but not stellar) command of language usage.

scoring agreement is quite high: Organization (QWK=0.71) and Development (QWK=0.73). This result is promising as it indicates that Organization and Development of story-telling can be reliably scored by humans. Surprisingly, the agreement for the non-narrative dimension, Conventions, is only rather moderate (QWK=0.46). Discussion among the two annotators revealed that the criteria for the score points in Conventions were very subjective. For example, they had difficulty deciding on when a Conventions violation, such as a specific grammatical error, was severe, and how much variety among the error types was needed to move the Conventions score from one score point to another.

Table 2 shows that agreement for all sub-traits is lower than agreement for the corresponding trait. Sub-trait agreement results also show that some story traits are more reliably scored than others. For example, it is easier to evaluate good openings and closings in stories (QWK=0.66) than to evaluate the quality of story style (QWK=0.58). Evaluation of stylistic devices and whether they indeed enhance the story is rather subjective.

Agreement for the Narrative and Total scores is also quite good. Narrative achieves a higher QWK than its individual components. The high agreement of the Total scores is interesting, as it incorporates the Conventions scores, on which substantial agreement was not achieved.

5.2 Inter-trait correlations

Previous research on writing has shown that traits are usually correlated (Lee et al., 2010; Bacha, 2001; Klein et al., 1998). We also observed this in our data. Inter-trait correlations (Pearson's r) are shown in Table 3. Scores for Organization and Development, are highly correlated (r = 0.88), and each is also correlated with Conventions (r = 0.40 and 0.42, respectively), albeit not as strongly. Not surprisingly, the composite scores, Narrative and Total, are highly correlated to their components.

	Org.	**Dev.**	**Conv.**	**Nar.**	**Tot.**
Org.	1.00	0.88	0.40	0.97	0.93
Dev.		1.00	0.42	0.97	0.94
Conv.			1.00	0.42	0.64
Nar.				1.00	0.97
Total					1.00

Table 3: Score correlations for traits, Narrative and Total.

6 Linguistic Features

We used the scoring rubric as a guideline for exploring construct-relevant features with a view towards automated analysis. We developed sets of features for the different narrative characteristics. Each set is described in detail in the following sections.

6.1 Transition Feature Set

Effective organization of ideas and events is typically achieved with the use of discourse markers. In order to encode effective transitioning, we compiled a transition-cue lexicon, and constructed features based on it.

We compiled a list of 234 discourse cues from the Penn Discourse Treebank (PDTB) manual (Prasad et al., 2008), and we manually collected a list of transition cues from the web by mining websites that provide tips on good essay/narrative writing. The latter, with a total of 484 unigrams and multi-word expressions, is more focused on cues that are used commonly to write stories (e.g., cues that provide locational or temporal connections) than the former.

Using the lexicon, we extracted two features from each essay: the number of cues in the essay and that number divided by the essay length. These two features form the *Transition* feature set.

6.2 Event-oriented Feature Set

Events are the building blocks of narratives, and good story-telling involves skillfully stringing events together. We construct an event-based feature set, *Events*, to capture event cohesion and coherence. Following the methodology proposed by Chambers and Jurafsky (2008), we built a database of event pairs from the GigaWord Fifth Edition corpus (Parker et al., 2011). Specifically, we used the Annotated Gigaword distribution (Napoles et al., 2012), which has been automatically annotated with typed dependency information (de Marneffe and Manning, 2008). Following Chambers and Jurafsky (2008), we define events as verbs in a text (excluding *be/have/do*) and pairs of events are defined as those verbs that share arguments in the text.

In the present work we limit our scope to the following set of (typed dependency) arguments: nsubj, dobj, nsubjpass, xsubj, csubj, csubjpass.

To estimate event cohesion, we extract all event pairs from an essay after pre-processing it with the Stanford Core NLP toolkit (Manning et al., 2014). Event tokens from the essay are linked into pairs when they share a filler in their arguments. For essays, we use Stanford co-reference resolution for matching fillers of verb-argument slots. For all event pairs extracted from an essay, we query the events database to retrieve the pair association value (we use the point-wise mutual information (Church and Hanks, 1990)). We define three quantitative measures to encode event cohesion: (1) total count of event pairs in the essay; (2) proportion of in-essay event-pairs that are actually found in the events database; (3) proportion of in-essay event-pairs that have substantial association (we use $PMI \geq 2$).

We also capture aspects of coherent event sequencing. For this, we compute event chains, which are defined as sequences of events that share the same actor or object, in subject or direct object role (Chambers and Jurafsky, 2008). Specifically, we encode the following additional features in the *Events* feature set: (4) the length of the longest chain found in the essay (i.e., number of event pairs in the chain); (5) the score of the longest chain (computed as the sum of PMI values for all links (event pairs) of the chain); (6) the length of the second longest chain found in the essay; (7) the score of the highest scoring chain in the essay; (8) the score of the second highest scoring chain in the essay; (9) the score of the lowest scoring chain is the essay; (10) the sum of scores for all chains in the essay. For each of the features 4-10, we also produce a feature that is normalized by the log of the essay length (log word-count).

6.3 Subjectivity-based Feature Set

Evaluative and subjective language is used to describe characters (e.g., *foolish, smart*), situations (e.g., *grand, impoverished*) and characters' private states (e.g., thoughts, beliefs, happiness, sadness) (Wiebe, 1994). These are evidenced when characters are described and story-lines are developed.

We use two lexicons for detecting sentiment and subjective words: the MPQA subjectivity lexicon (Wilson et al., 2005) and a sentiment lexicon, ASSESS, developed for essay scoring (Beigman Klebanov et al., 2012). MPQA associates a positive/negative/neutral polarity category to its entries, while ASSESS assigns a positive/negative/neutral polarity probability to its entries. We consider a term from ASSESS to be polar if the sum of positive and negative probabilities is greater than 0.65 (based on manual inspection of the lexicon). The neutral category in MPQA comprises subjective terms that indicate speech acts and private states (e.g., *view, assess, believe*), which is valuable for our purposes. The neutral category in ASSESS consists of non-subjective words (e.g., *woman, technologies*), which we ignore. The polar entries of the two lexicons differ too. ASSESS provides polarity for words based on the *emotions they evoke*. For example, *alive, awakened* and *birth* are highly positive, while *crash, bombings* and *cyclone* are strongly negative.

We construct a *Subjectivity* feature set comprised of 6 features encoding, for each essay, the presence (a binary feature) and the count of MPQA and ASSESS polar words and MPQA neutral words.

6.4 Detailing Feature Set

Providing specific details, such as names to characters, and describing the story elements, helps in developing the narrative and providing depth to the story. Proper nouns, adjectives and adverbs come into play when a writer provides descriptions. Thus, we create a *Details* feature set comprised of a total of 6 features encoding, separately, the presence (a binary feature) and the count of proper nouns, adjectives and adverbs.

6.5 Graph Feature Set

Graph statistics have been reported to be effective for capturing development and coherence in essays (Mesgar and Strube, 2016; Somasundaran et al., 2016). We closely follow the implementation and features described in Somasundaran et al. (2016) for capturing narrative development (due to space constraints we refer the reader to the original paper). Graphs were constructed from essays by representing each content word (word type) in a sentence as a node in the graph. Links were drawn between words belonging to adjacent sentences. Features based on connectivity, shape and PageRank were extracted,

giving a total of 19 *Graph* features. Specifically, the features used were: percentage of nodes with degrees one, two and three; the highest, second-highest and median degree in the graph; the highest degree divided by the total number of links; the top three PageRank values in the graph, their respective negative logarithms, and their essay length-normalized versions; the median PageRank value in the graph, its negative log and essay length-normalized version.

6.6 Content word usage

Content word usage, also known as *lexical density* (Ure, 1971), refers to the amount of open-class (content words) used in an essay. The greater proportion of content words in a text, the more difficult or advanced it is (Yu, 2010; O'Loughlin, 1995), and it has been suggested that, for academic discourse, too much lexical density is detrimental to clarity (Halliday and Martin, 1993). The *Content* feature is the inverse of the proportion of content words (POS tagged noun/verb/adjective/ adverb) to all words of an essay.

6.7 Pronoun Usage

The use of pronouns in story-writing has several important aspects. On one hand, pronouns can indicate the point of view (perspective) in which the story is written (Fludernik, 2009; Rimmon-Kenan, 2002). Perspective is important in both construction and comprehension of narrative (Rimmon-Kenan, 2002). The use of pronouns is also related to reader engagement (Mentzell et al., 1999) and immersion (Oatley, 1999). Stories with first person pronouns lead to stronger reader immersion, while stories written in third person lead to stronger reader arousal (Hartung et al., 2016). In our data, we counted personal pronouns (e.g., *I, he, it*), including contractions (e.g., *he's*), and possessive pronouns (e.g., *my, his*). For each story, the counts were normalized by essay length. A single feature, *Pronoun*, was encoded using the proportion of first and third person singular pronouns in the essay.

6.8 Modal Feature

As an account of connected events, a narrative typically uses the past tense. By contrast, modals appear before untensed verbs and generally refer to the present or the future. They express the degree of ability (*can, could*), probability (*shall, will, would, may, might*), or obligation/necessity (*should, must*). An overabundance of modals in an essay might be an indication that it is not a narrative or is only marginally so. This idea is captured in the *Modal* feature, which is the proportion of modals to all words of an essay.

6.9 Stative Verbs

Stative verbs are verbs that describe states, and are typically contrasted with dynamic verbs, which describe events (actions and activities) (Vendler, 1967). In narrative texts, stative verbs are often used in descriptive passages (Smith, 2005), but they do not contribute to the progression of events in a story (Almeida, 1995; Prince, 1973). Our conjecture is that if a text contains too many stative verbs, then it may not have enough of an event sequence, which is a hallmark of a narrative. We compiled a list of 62 English stative verbs (e.g., *know, own, resemble, prefer*) from various linguistic resources on the web. During processing of an essay, we identify verbs by POS tags, and stative verbs via list-lookup. Separately, we identify copular uses of "to be" and count them as statives. Our feature, *Statives*, is the proportion of stative verbs out of all verbs in an essay.

7 Experiments

Our experiments investigate the following questions: (1) Is it possible to score narrative quality traits in essays using automated methods? (2) Which of our feature sets are effective for scoring narrative quality traits? (3) How do our narrative-inspired features perform as compared to a baseline that is competitive but does not specifically address the narrative construct? (4) How does overall scoring of narrative essays differ from trait scoring? (5) What are the best feature combinations for narrative scoring?

To answer these questions, we built and evaluated scoring systems for each trait, overall Narrative and Total scores. In each case, we performed detailed ablation studies at the feature-set level. We have 10 features sets (9 feature sets described above plus a baseline feature set); thus 1024 feature set combinations were investigated. As our traits are highly correlated, we used all of our features for building

systems for each trait, leaving it to the ablation process to reveal the most promising feature set combination.

7.1 Baseline

E-rater (Attali and Burstein, 2006), a state-of-the-art commercial system for automatic essay scoring, uses a comprehensive suite of features covering many aspects of writing quality, such as grammar, language use, mechanics, fluency, style, organization, and development. We use all of the features from e-rater, a total of 10 features, as the *Baseline* feature set. While e-rater is not designed for trait scoring, it incorporates features that address the traits of interest in this work. Development and Organization are captured by features that, among other things, count and encode the number and length of discourse elements such as thesis, main points, supporting ideas, and conclusion (Burstein et al., 2003).

7.2 Results

We experimented with Linear Regression, Support Vector Regression (RBF kernel), Random Forests, and Elastic Net learners from the scikit-learn toolkit (Pedregosa et al., 2011), with 10-fold cross-validation on 942 essays. As Linear Regression results were consistently better, both for Baseline and for our features, we only report results from this learner. Trimming of the predicted linear regression output was performed; that is, if the predicted score was above the max score, or below the min score, it was 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 1000 bootstrap samples).

For each trait-scoring experiment, we extracted all the features (described in Section 6) from the essays and used the corresponding human trait scores for training and testing. Thus, the input essays and their features are the same across all experiments. What varies is the trait to be predicted and, consequently, the performance of feature sets as well as the best feature combination.

Table 4 shows the performance of Baseline, the individual features, all features, and the best feature combination, for all three traits, overall Narrative and Total scoring. Performance of individual features that exhibit some predictive power is also shown in the table. The single-measure features Modal, Pronoun, Content, and Stative show no predictive power individually (QWKs = 0) and are omitted from the table for space reasons.

Organization Understandably, Baseline performs poorly for scoring Organization in narratives, as its focus is evaluating overall writing proficiency. Individual feature sets, Details, Transition, Events and Subjectivity, have some predictive capability, but it is not very high. This is not surprising as they each encode only a specific aspect of narrative quality. The Graph feature set outperforms the Baseline feature set, but the difference is not statistically significant. When all features are put together (*All features*), the QWK obtained is 0.56, which is substantially higher than Baseline ($p < 0.001$), but as not as good as the best performing feature set.

The best combination of our proposed features (Details+ Modal+ Pronoun+ Content+ Graph+ Subjectivity+ Transition) achieves a QWK of 0.60, substantially better performance than Baseline ($p < 0.001$), reflecting an improvement of 13 percentage points. This result indicates that developing features to encode narrative quality is important for evaluating Organization in narrative essays. Most of our explored feature sets, even those that do not individually perform well, are part of the best system. Two feature sets that are not present in the best feature combination are Statives and Events. The exclusion of the former is reasonable – stative verbs are related to story development. The exclusion of Events is surprising, as it intuitively encodes the coherence of events, impacting the organization of the essay. The best feature combination that includes Events achieves a QWK of 0.58. The Baseline features are not part of the best system, confirming our intuition that features that specifically encode narrative quality are needed for this narrative trait.

From our ablation results, we inspected the top 10 best-performing feature set combinations in order to determine which features consistently produce good systems. Pronoun, Content, Graph and Subjectivity were a part of all 10 of the 10 top systems, Transition was in 9, Details was in 7 and Modal was in 6 feature sets. This suggests that singleton features such as

Feature set	Organization	Development	Conventions	Narrative	Total
Baseline	0.47	0.51	0.44	0.53	0.60
Details	0.36	0.41	0.19	0.39	0.41
Transition	0.39	0.50	0.23	0.49	0.48
Events	0.39	0.43	0.26	0.45	0.45
Subjectivity	0.41	0.47	0.20	0.47	0.46
Graph	0.49	0.54	0.17	0.56	0.54
All features	0.56	0.63	0.46	0.65	0.67
Best feature combination*	0.60	0.66	0.50	0.67	0.70

Table 4: Performance (QWK) on predicting traits and Narrative and Total scores; Best feature combinations:
*For **Organization**: Details+ Modal+ Pronoun+ Content+ Graph+ Subjectivity+ Transition;
*For **Development**: Details+ Modal+ Content+ Graph+ Statives+ Transition;
*For **Conventions**: Baseline + Details + Graph;
*For **Narrative**: Baseline+ Details+ Modal+ Pronoun+ Content+ Graph+ Statives+ Subjectivity+ Transition;
*For **Total**: Details+ Baseline+ Modal+ Content+ Graph+ Subjectivity+ Transition

Pronoun and Content are indeed useful, even though they cannot be used in isolation.

Development We observe similar trends seen with the Organization trait – the Baseline feature set does not capture Development very effectively, and some individual feature sets have predictive power for this trait but perform poorly. Graph outperforms Baseline, but this is not statistically significant. Using all of the available features produces QWK=0.63, a significant improvement over Baseline, ($p < 0.001$). The best system achieves a performance of QWK=0.66, outperforming Baseline by 15 percentage points ($p < 0.001$). The best feature combination contains 6 of the 9 proposed features and differs from the best features for Organization by the inclusion of Statives and the exclusion of Pronoun and Subjectivity. Content, Graph and Transition also occur in all of the top 10 best-performing systems.

Conventions Even though scoring language conventions is not the focus of this work, we were curious how well our features evaluate this dimension. We observe that overall performance is lower than for the other two traits, which is to be expected as we do not have high human inter-rater agreement to start with. The Baseline e-rater feature set is the best performing individual feature set, and the narrative-specific features perform rather poorly. Using all features (QWK=0.46) only produces a 2 point improvement over Baseline, which is not statistically significant. Adding Details and Graph to Baseline produces the best system, an improvement of 6 percentage points, QWK=0.50, ($p < 0.001$). All three features are also the only feature sets that consistently occur in all the 10 top-performing systems.

Narrative In general, the results for Narrative scoring follow the same trends as the results for Organization. Graph features outperform the Baseline significantly ($p < 0.05$). Using all available features produces a significant improvement in performance (0.65 QWK; $p < 0.001$). Baseline features are now a part of the best feature set combination (Baseline+ Details+ Modal+ Pronoun+ Content+ Graph+ Statives+ Subjectivity+ Transition), which achieves a QWK of 0.67, an improvement of 14 percentage points ($p < 0.001$). The best feature combination without the Baseline features achieves QWK = 0.66, and this is not statistically different from the performance of the best system. Modal, Content, and Graph occur in all 10, and Subjectivity and Transition occur in nine of the top 10 feature combinations.

Total For Total scoring, the Baseline feature set is the best performing individual feature set, with QWK = 0.60. Using all features produces a significant ($p < 0.001$) performance boost at 0.67 QWK. The best feature combination (Details+ Baseline+ Modal+ Content+ Graph+ Subjectivity+ Transition) improves over Baseline by 10 percentage points,

with a QWK of 0.70 ($p < 0.001$). The best result obtained by a feature combination without Baseline (Details+ Modal+ Content+ Graph+ Subjectivity+ Transition) is QWK = 0.68, which is significantly higher than the Baseline performance ($p < 0.001$), indicating that our features are able to effectively score essays by themselves, as well as in combination with the Baseline features to get an improved system. Except for Details and Transition, all features of the best system also occur in all the top-10 systems.

8 Analysis and Discussion

The results show that our proposed features vary in effectiveness. Graph features proved to be more effective than Transition, Subjectivity and Details. The effectiveness of single-measure features (Pronoun, Statives, Content and Modal) was evident by their inclusion in the best combination models.

Although Events was reasonably predictive on its own for Organization and Development, it was not found in the best performing combinations, nor did it participate in the top 10 feature sets for any of the traits. This surprising result suggests that other features, which are correlated with Events, must be stronger indicators of narrative competence.

Our results also show no clear segregation of features by trait, as most of the features appearing in the best models for Organization and Development were the same. We attribute this to the high correlation between the human scores for the two traits; a model that is good for one will be good for the other.

8.1 Correlation Study

We performed correlation analysis to test if our intuitions regarding the feature sets, as discussed in Section 6, are supported by the data, and to study the effect of length. Length is a well-known confounding factor in essay scoring as longer essays tend to get higher scores (Chodorow and Burstein, 2004). This also applies to narratives, as it is difficult to tell a good story without using a sufficient amount of words. In our data, Pearson correlations of essay length with human scores are: Conv.: 0.35, Dev.: 0.58, Org.: 0.54. However, it is important that our encoded features capture more than just the length of the narrative. In order to test this, we conducted correlation analysis between each feature and human trait score by partialling out length.

Feat	Org	Dev	Conv
Base.	0.19 (0.28)	0.19 (0.41)	0.39 (0.43)
Detl.	0.17 (0.21)	0.16 (0.20)	0.08 (0.18)
Trans.	-0.10 (0.23)	-0.15 (0.22)	-0.05 (0.27)
Event	0.20 (0.27)	0.19 (0.26)	0.14 (0.19)
Subj.	0.17 (0.48)	0.19 (0.52)	0.07 (0.12)
Graph	0.36 (0.61)	0.39 (0.65)	0.06 (0.28)
Cont.	-0.19 (-0.30)	-0.20 (-0.31)	-0.20 (-0.28)
Pron.	0.19 (0.18)	0.17 (0.17)	0.12 (0.10)
Modal	-0.17 (-0.17)	-0.21 (-0.17)	-0.01 (-0.18)
Statv.	-0.10 (-0.18)	-0.10 (-0.18)	-0.05 (-0.11)

Table 5: Maximal partial correlations with scores, controlling for length (simple correlations in parentheses).

Table 5 shows the *maximal* partial correlation of each feature set with the human scores. For feature sets that contain only a single feature (e.g., Modal), we directly report the partial correlation for that feature. For feature sets that contain multiple features, due to space constraints, we report the maximal partial correlation achieved by any feature within that set[5]. The value in the parentheses indicates the corresponding feature's simple correlation with score.

We observe that for all features except Pronoun and Modal, the correlation with score drops when length is accounted for, indicating the influence of essay length on scores. This effect is more pronounced in features that employ counts (e.g., counts of adverbs), as more support is found in longer essays. The baseline is correlated more with conventions than the two narrative traits. An opposite effect is seen for our narrative-specific features. The negative sign for Statives, Content and Modal supports our intuitions regarding these features – more use of these reduces story quality.

8.2 Error Analysis

Table 6 shows the human-machine confusion matrix for Development trait scores. Confusion matrices for other traits also show a similar trend. We observe that most of the errors in score prediction are at adjacent score points. This is perhaps in part due to our human-human agreement criterion during data an-

[5]Note that, within a set, different features might have maximum values for different traits

Human	Machine					
	0	*1*	*2*	*3*	*4*	**total**
0	8	9	18	5	0	**40**
1	8	28	43	5	0	**84**
2	1	8	159	101	1	**270**
3	0	0	83	205	31	**319**
4	0	0	9	125	95	**229**

Table 6: Human-machine confusion matrix for Development traits scores

notation – disagreement of one score point did not trigger adjudication.

The system encounters more difficulty predicting the correct scores at the ends of the scale (score points 0-1 and score point 4). The difficulty with scores 0 and 1 is partially attributable to the small amount of training data for these scores.

In a more detailed analysis of the human-machine discrepancies, we first focus on the forty essays that were rated 0 by the annotators (Table 6, row 1). The machine and human agreed on only eight of these. All eight are non-narratives, and seven of them are extremely short (3 to 51 words). Twenty seven of the remaining 32 were well-written, long, non-narrative essays (and thus off-purpose according to our rubric). For example, one of the essays, which was written for a "describe a travel experience" prompt, presented a discussion about the educational advantages of travel in general.

Next, we consider the 84 essays (all narratives) that were rated 1 by the annotators (row 2 of Table 6). Of these, the eight that were scored 0 by the machine were rather short (length 15 to 69 words) and poorly written. The human and the machine agreed on 28 essays, whose average length was somewhat longer (93 words). For the 43 essays that the machine over-scored by 1 point, the average length was 154 words. All five essays that the machine over-scored by 2 points were long, ranging from 200 to 421 words, but were either expository essays or were very poorly written. This scoring pattern suggests that human-machine disagreement is at least partially rooted in essay length.

For the essays that were rated 4 by the human annotators (Table 6, last row), the machine underestimated nine essays by 2 points. These essays were relatively short (from 135 to 383 words). For comparison, in the 125 essays where the machine underestimated the human score by only one point, the average length was 418 words. For the 95 essays that were scored 4 by both the human and machine, the average length was 653 words. A similar effect of length was seen among the essays scored 2 and 3 by the human annotators.

The error analysis at the lowest range of human scores demonstrates that an accurate system must be able to properly handle non-narrative essays. One possible solution is to consider coupling our system with a binary narrative classifier that would flag non-narrative essays. Further research is also clearly needed to reduce the influence of essay length on automated scoring. This was particularly demonstrated for essays where writers managed to produce well written, but very short, stories that were underscored by the machine.

9 Conclusions and Future Work

In this article, we have presented evidence that humans can reliably score development and organization traits and their sub-traits in narratives, and that some sub-traits can be more reliably scored than others. We have also presented evidence that automated systems with narrative-specific features can reliably score narrative quality traits and can do so significantly better than a state-of-the-art system designed to assess general writing proficiency.

Scoring narrative essays is challenging because typically there is no right answer, nor any limit to the creative possibilities in effective story-telling. In this work, we have explored only the proverbial tip of the iceberg in terms of features and methods for scoring narrative essays. While we are encouraged by our results, we believe that further improvement will require more elaborate representations of story content and meaning. Accordingly, we plan to explore automated evaluation of narrative sub-traits, including plot, point of view and character development, and of the relationships among them.

References

Caralee J. Adams. 2014. Essay-grading software seen as time-saving tool. *Education Week*, 33(25):13–15.

Apoorv Agarwal, Anup Kotalwar, and Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on Alice in Wonderland. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1202–1208.

Michael J. Almeida. 1995. Time in narratives. *Deixis in Narrative: A Cognitive Science Perspective*, pages 159–189.

Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997 – 1013.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment*, 4:3.

Nahla Bacha. 2001. Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29(3):371–383.

Niranjan Balasubramanian, Stephen Soderland, Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, Seattle, WA, October.

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, June.

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.

Beata Beigman Klebanov, Nitin Madnani, Jill Burstein, and Swapna Somasundaran. 2014. Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 247–252.

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.

Thomas Bogel, Jannik Strotgen, and Michael Gertz. 2014. Computational narratology: Extracting tense clusters from narrative texts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Hein Broekkamp, Tanja Janssen, and Huub van den Bergh. 2009. Is there a relationship between literature reading and creative writing? *Journal of Creative Behavior*, 43(4):281 – 296.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

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 *In Proceedings of the 24th Annual Conference on Neural Information Processing Systems*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 602–610.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1798–1807, Seattle, WA, October.

Eugene Charniak. 1972. Toward a model of children's story comprehension. Technical report, MIT, Cambridge, MA, USA.

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 Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, pages 2704–2710. Association for the Advancement of Artificial Intelligence Press.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater's performance on TOEFL essays. TOEFL research report 73, Educational Testing Service, Princeton, NJ, USA.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March.

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

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

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

Scott Elliot. 2003. Intellimetric: From here to validity. *Automated essay scoring: A cross-disciplinary perspective*, pages 71–86.

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.

David K. Elson. 2012. *Modeling Narrative Discourse*. Ph.D. thesis, Columbia University.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Tenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Mark Alan Finlayson. 2012. *Learning narrative structure from annotated folktales*. Ph.D. thesis, Massachusetts Institute of Technology.

Mark A. Finlayson. 2013. A survey of corpora in computational and cognitive narrative science. *Sprache Und Datenverarbeitung (International Journal for Language Data Processing)*, 37(1–2).

Monika Fludernik. 2009. *An Introduction to Narratology*. Routledge, London.

Ronald B. Gillam and Nils A. Pearson. 2004. TNL: Test of Narrative Language. *Austin, TX: Pro-Ed*.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empircal Methods in Natural Language Processing*, Boston, MA.

Michael A. K. Halliday and James R. Martin. 1993. *Writing Science: Literacy and Discursive Power*. The Falmer Press, London.

Harry Halpin and Johanna D. Moore. 2006. Event extraction in a plot advice agent. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 857–864, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franziska Hartung, Michael Burke, Peter Hagoort, and Roel M. Willems. 2016. Taking perspective: Personal pronouns affect experiential aspects of literary reading. *PLoS ONE*, 5(11).

Bram Jans, Steven Bethard, Ivan Vulic, and Marie Francine Moens. 2012. Skip N-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344, Avignon, France, April.

Stephen P. Klein, Brian M. Stecher, Richard J. Shavelson, Daniel McCaffrey, Tor Ormseth, Robert M. Bell, Kathy Comfort, and Abdul R. Othman. 1998. Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education*, 11(2):121–137.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2010. Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3):391–417.

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

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies*, 5(3):1–142.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics System Demonstrations*, pages 55–60.

Neil McIntyre and Mirella Lapata. 2010. Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden.

Anne McKeough, Randy Genereux, and Joan Jeary. 2006. Structure, content, and language usage: How do exceptional and average storywriters differ? *High Ability Studies*, 17(2):203 – 223.

Phyllis Mentzell, Elizabeth Vander Lei, and Duane H. Roen. 1999. Audience Considerations for Evaluating Writing. *Evaluating Writing: The Role of Teacher's Knowledge about Text, Learning, and Culture*.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423.

Jon Miller and Robin Chapman. 1985. Systematic Analysis of Language Transcripts. *Madison, WI: Language Analysis Laboratory*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Pushmeet Kohli Lucy Vanderwende, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 12-17. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction*, pages 95–100.

Huy Nguyen and Diane J. Litman. 2016. Improving Argument Mining in Student Essays by Learning and Exploiting Argument Indicators versus Essay Topics. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pages 485–490.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besancon. 2015. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 188–197, Beijing, China, July.

Keith Oatley. 1999. Meetings of minds: Dialogue, sympathy, and identification, in reading fiction. *Poetics*, 26:439–454.

Natalie G. Olinghouse and Jacqueline T. Leaird. 2009. The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading & Writing*, 22(5):545 – 565.

Kieran O'Loughlin. 1995. Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, 12:217–237.

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, Lisbon, Portugal.

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.

Ellis Batten Page. 1994. Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, 62(2):127–142.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. *Philadelphia: Linguistic Data Consortium*.

Rebecca J. Passonneau, Adam Goodkind, and Elena T. Levy. 2007. Annotation of Children's Oral Narrations: Modeling Emergent Narrative Skills for Computational Applications. In *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*, pages 253–258.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Douglas B. Petersen, Sandra Laing Gillam, and Ronald B. Gillam. 2008. Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders*, 28(2):115 – 130.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Gerald Prince. 1973. *A Grammar of Stories: An Introduction*. Mouton, The Hague.

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. *The European Physical Journal Data Science*, 5(1):31.

Shlomith Rimmon-Kenan. 2002. *Narrative Fiction: Contemporary Poetics*. Routledge, London.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Mark D. Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

Carlota S. Smith. 2005. Aspectual entities and tense in discourse. In Paula Kempchinsky and Roumyana Slabakova, editors, *Aspectual Inquiries*, pages 223–237. Springer Netherlands, Dordrecht.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings*

of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578, Osaka, Japan, December.

Christian Stab and Iryna Gurevych. 2017. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Long Papers*, volume 1.

Nancy L. Stein and Christine G. Glenn. 1979. An Analysis of Story Comprehension in Elementary School Children: A test of a schema. *New Directions in Discourse Processing*.

Carol J. Strong, Mercer Mayer, and Marianna Mayer. 1998. *The Strong Narrative Assessment Procedure*. Thinking Publications.

Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn A. Walker. 2014. Identifying narrative clause types in personal stories. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 171.

Jean Ure. 1971. Lexical density and register differentiation. In Perren G. E. and Trim J. L. M., editors, *Applications of linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969*, pages 443–452. Cambridge University Press, Cambridge, UK.

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

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

Stephen G. Ware, Brent E. Harrison, Robert Michael Young, and David L. Roberts. 2011. Initial results for measuring four dimensions of narrative conflict. In *The Fourth Workshop on Intelligent Narrative Technologies at the 2011 AI and Interactive Digital Entertainment Conference*.

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.

Guoxing Yu. 2010. Lexical diversity in writing and speaking task performances. *Applied Linguistics*, 31(2):236–259.