Literary Event Detection

Matthew Sims
School of Information
UC Berkeley
msims@berkeley.edu

Jong Ho Park
Computer Science Division
UC Berkeley
jjhpark@berkeley.edu

David Bamman
School of Information
UC Berkeley
dbamman@berkeley.edu

Abstract

In this work we present a new dataset of *literary events*—events that are depicted as taking place within the imagined space of a novel. While previous work has focused on event detection in the domain of contemporary news, literature poses a number of complications for existing systems, including complex narration, the depiction of a broad array of mental states, and a strong emphasis on figurative language. We outline the annotation decisions of this new dataset and compare several models for predicting events; the best performing model, a bidirectional LSTM with BERT token representations, achieves an F1 score of 73.9. We then apply this model to a corpus of novels split across two dimensions—prestige and popularity—and demonstrate that there are statistically significant differences in the distribution of events for prestige.

1 Introduction

Do events determine the shape of literary narratives? This question reaches back at least as far as the 1920s, when literary theorists from the Russian Formalist school began making distinctions between *syuzhet* (the way in which events are presented in a narrative) and *fabula* (the chronological sequence of events, distinct from the way they're represented) (Shklovsky, 1990; Propp, 2010). Even on a far more localized scale, events are often considered to play a fundamental role in how literary narratives progress. Moretti (2013), for instance, describes the inherent productivity of events in Daniel Defoe's novel *Robinson Crusoe*, where one event invokes another in a chain of occurrences that seem to flow in "micro-narrative sequences." Such localized sequences in turn relate to the larger architecture of plot, which has its own distinct modes of organization and generation (Forster, 1927; Genette, 1983; Brooks, 1992). The status of events in literature thus inevitably engages larger questions about scale and narrative technique.

At the same time, the representation and identification of events and their participants in NLP have historically focused on the domain of news, including early evaluation campaigns like MUC (Sundheim, 1991), seminal datasets like ACE 2005 (Walker et al., 2006) and the DEFT ERE framework (Aguilar et al., 2014; Bies et al., 2016), as well as other resources that require the identification of events as a precondition for other activities, such as temporal ordering (Pustejovsky et al., 2003b) or factuality judgments (Saurí and Pustejovsky, 2009; de Marneffe et al., 2012; Werner et al., 2015; Lee et al., 2015; Rudinger et al., 2018).

The role of events in literary fiction, however, is very different from their role in fact-based reporting of events in the real world, including historical texts (Sprugnoli and Tonelli, 2017). Novels and even most short stories tend to be much longer than news articles, and tend to have more complex narrative structures both locally (individual scenes) and globally (plot) than works of nonfiction. Furthermore, literature is a creative enterprise. Journalistic discourse typically reports what actually happened in the real world and depicts definite causal chains connecting events; this causality is not hard coded into literary event sequences.

We present in this work a new dataset of event annotations in the domain of literature that aims to bridge this gap between the rich landscape of existing work in event representation in NLP for news—including contemporary neural methods (Orr et al., 2018; Sha et al., 2018; Nguyen and Grishman, 2018)—and the needs of literature scholars for models of events in their domain. To develop a common thread with fact-based rep-

resentations of real-world events while also laying the foundation for models to faithfully track the unique movements of narrative plot, we focus solely on events in literary texts that are depicted as *actually happening*—i.e., those with asserted *realis* (discussed in more detail in §4). As distinct from other epistemic modalities (such as future events, hypotheticals, and extradiegetic summaries by a narrator), realis events are depicted as existing within the imagined world of the literary text, and take place at a specific place and a specific time.

In this work we make the following contributions:

- We present a new annotated dataset of literary events in 210,532 tokens from 100 books and describe some of the key annotation guidelines that we have tailored to the unique challenges posed by novelistic discourse. The dataset is freely available for download under a Creative Commons ShareAlike 4.0 license as a part of LitBank at `https://github.com/dbamman/litbank`.

- We compare multiple models for realis event detection in literary texts, including both featurized and neural approaches, with the best performing model achieving an F1 score of 73.9.

- We apply this model to a corpus of novels and demonstrate that there are statistically significant differences in the ratio of realis events between novels written by authors with high prestige—defined by Underwood (2019) as works that have been reviewed by elite literary journals—and those written by authors without such prestige. High prestige authors (such as James Joyce and Virginia Woolf) use fewer realis events depicting concrete actions in their works.

2 Background and Previous Work

We draw on several threads of previous research in designing a dataset and model to support literary event detection. First, while much work at the intersection of NLP and literary analysis has focused on computational approaches to characters and their relationships (Bamman et al., 2014; Vala et al., 2015; Iyyer et al., 2016; Chaturvedi et al., 2017), far less has explored the event structure of literary texts. Plot is often explored through the lens of sentiment (Alm and Sproat, 2005; Mohammad, 2011; Elsner, 2012; Jockers, 2015; Reagan et al., 2016) rather than the concrete events that comprise it.

Second, we draw on the vast literature in NLP for the detection of events, participants, and their structured relationships, from the featurized models of Ahn (2006) and Li et al. (2013) to the variety of neural architectures that have been applied to the task of event detection, such as CNNs (Nguyen and Grishman, 2015), including dynamic multi-pooling CNNs (Chen et al., 2015) and skip-gram CNNs (Nguyen and Grishman, 2018), RNNs (Nguyen et al., 2016), hybrid LSTM-CNN architectures (Feng et al., 2016), and attention (Liu et al., 2017, 2018)

While most approaches use sentence-level information to detect events, we also draw on the work of Liao and Grishman (2010), which instead incorporates document-level information (potentially useful for longer literary narratives).

3 Data

The corpus we have selected to annotate consists of approximately the first 2000 words of 100 literary works currently in the public domain (i.e., published before 1923), previously used by Bamman et al. (2019). The majority of these texts are canonical novels published in the nineteenth and early twentieth centuries (e.g., Jane Austen's *Pride and Prejudice*, Herman Melville's *Moby Dick*, and James Joyce's *Ulysses*). A smaller percentage of this corpus consists of popular genre fiction published within this same time frame (e.g., *King Solomon's Mines*, *Tarzan of the Apes*, and *Desert Gold*). All of these texts have been selected from the Project Gutenberg corpus and collectively exhibit a range of novelistic discourse. This range is particularly useful and necessary for exploring literary event realis, providing examples of novels that are narratively and stylistically complex as well as others that are more declarative and plot-driven.

4 Event annotations

Events remain a contested category across narrative theory, philosophy, and linguistics, with definitions varying depending on discipline, application, and context. Most linguistic event classifications nevertheless trace their lineage back to

Vendler (1957), who proposed four categories to distinguish the different relationships that exist between verbs and time: *activities* (dynamically unfolding processes), *achievements* (occurrences that are completed almost instantaneously), *accomplishments* (occurrences that have some duration but also have a predetermined endpoint), and *states* (persistent conditions that span a period of time and don't have any definite endpoint).

A simpler classification that some scholars have traced back to Aristotle (Sasse, 2002) simply distinguishes between events and states, the latter usually defined as non-dynamic situations that pertain over time. Many event annotation systems including TimeML (Pustejovsky et al., 2003a), ACE (LDC, 2005), and Light ERE (Aguilar et al., 2014) also treat *changes of state* as being events, since such changes indicate a dynamic break from prior conditions.

In our annotation approach, we include activities, achievements, accomplishments, and changes of state as being events. We introduce several more fine-grained distinctions, however, as far as which subsets within each of these categories should be labelled for our specific purposes, as detailed below.

At a high level, our goal is to model only what is depicted as *actually* occurring in a text; in other words, those events with asserted *realis*. The ACE 2005 event annotation guidelines (LDC, 2005) outline four dimensions for tagging events involving determinations for polarity, tense, specificity, and modality. We follow the Light ERE (Aguilar et al., 2014) approach of only selecting aspects that capture realis:

- Polarity: Events must have a positive polarity (i.e., positively asserted as occurring); events with negative polarity are defined as having *not* taken place (such as "he did not understand").

- Tense: Events must be in the past or present tense. Events in the future tense are not tagged.

- Genericity: Generic events describe a category (e.g., *dogs bark*) rather than a specific occurrence involving a specific entity (*my dog barked this morning*). We only tag specific events in our framework; all generic events are ignored. We consider an event to be specific if it is "a singular occurrence at a particular place and time" (LDC, 2005).

- Modality: Only asserted events—those that are indicated to have actually occurred—are tagged. All other modalities (believed, hypothetical, desired, etc.) are not.

We also employ the following standards in our annotation approach:

- Similar to both the ACE and Light ERE guidelines, we tag *event triggers*, defined as the minimum extent of text capable of representing an event. For our purposes, this extent is always a single word. This is in contrast to Light ERE, which allows for multi-word triggers, and to ACE, which mostly restricts triggers to single words but makes an exception for phrasal verbs by including the particle if it immediately follows the main predicate.

- We limit event triggers to the following three parts of speech: verbs, adjectives, and nouns (including nominals). Adverbs and prepositions are not annotated as events.

- In contrast to both the ACE and Light ERE guidelines, which restrict taggable events to those falling within eight specific types (life, movement, transaction, business, conflict, contact, personnel, and justice), we adopt an open approach and make no restrictions on the types of events that are tagged.

Due to the specific domain we are annotating (English language fiction), we have also found it necessary to define several rules that are not explicitly presented in the ACE or Light ERE standards. In particular, since mental states play an especially prominent and complex role in many novels (and noticeably so relative to more fact-based discourses such as the news) we have given particular attention to defining rules for stative events. In our annotations, we tag a state as being an event assuming one or more of the following conditions has been met:

1. An explicit change of state has occurred (whether initiation, termination, or alteration), and this change can be determined solely within the context of the sentence in which the potential event trigger appears.

(1) Stephen Dedalus, displeased and sleepy, **leaned** his arms on the top of the staircase and **looked** coldly at the **shaking gurgling** face that **blessed** him, equine in its length, and at the light untonsured hair, grained and hued like pale oak. (Joyce, *Ulysses*)

(2) My eyes **followed** his trim figure, richly though sombrely clad, then **fell** with a sudden **dissatisfaction** upon my own stained and frayed apparel. (Johnston, *To Have and to Hold*)

(3) He generally arrived in London (like the influenza) from the Continent, only he arrived unheralded by the Press; and his visitations set in with great severity. (Conrad, *The Secret Agent*)

Table 1: Three annotation examples with tagged event triggers in bold and candidate triggers that would not be tagged underlined.

2. The cause of the state can be deduced (again within the context of the sentence), and it is clear that the cause and resulting state have occurred at the same location. For example, the following states (in bold) would be labelled as events: "When he received this appointment he was both **elated** and **appalled**." (Burroughs, *Tarzan of the Apes*)

3. The potential event trigger refers to a mental state that is inherently acute, semantically speaking. For instance, words such as "astonished," "shocked," "aghast," and "stunned" all suggest mental states that are acute responses to some stimulus and are usually only maintained for a limited duration.

Table 1 presents three sample sentences annotated under our guidelines that illustrate important aspects of our framework, including mental states with no evidence of immediate change (*displeased* and *sleepy* in example 1), resultatives (*stained* and *frayed* in example 2), and generic events that describe periodic activities but not a single action grounded at a single moment in time (*arrived* in example 3).

Meanwhile, Table 2 shows the fifteen words with the highest occurrence as events in the annotations, along with the percentage of the time they are labelled as events. For the most part, these words can be broken down into four respective categories: verbs related to **conversation** (*said*, *asked*, *heard*, *answered*, and *cried* when indicating a vocalization); verbs related to **movement** (*came*, *went*, and *turned)*; verbs related to **perception** (*looked* and *saw*); and verbs related to **obtainment** (*took* and *found*). As the event rates make clear, even these words are only labelled as events a portion of the time (in some cases less than half of all occurrences) either due to contextual usage or the broader constraints imposed by realis.

Word	Count	Event Rate
said	465	89%
came	95	52%
looked	92	58%
went	92	60%
asked	69	93%
heard	63	59%
saw	59	55%
cried	59	97%
took	57	60%
turned	55	74%
told	51	56%
found	49	42%
answered	45	96%
put	44	41%
thought	38	32%

Table 2: The fifteen words with the highest overall occurrence as events in the annotations (Count) along with the percentage they are labelled events relative to their overall occurrence in the corpus (Event Rate).

Finally, to highlight why annotating events in novels is a particularly challenging task, we also briefly mention some of the phenomena that frequently arise. There are no taggable events in the examples below; potential triggers that are *not* tagged are underlined.

Figurative events. Often figurative language or an extended metaphor will be used to represent an event: "He had broken a thickness of ice, the formation of many a winter; had had his reasons for a long silence." (James, *The Turn of the Screw*)

Realis events presented in an irrealis mood. Sometimes events that have actually occurred are presented in a different modality for rhetorical purposes: "As to your practice, if a gentleman walks into my rooms smelling of iodoform, with a black mark of nitrate of silver upon his right forefinger, and a bulge on the right side of his top-hat to show where he has secreted his stethoscope, I must be dull, indeed, if I do not pronounce him to be an active member of the medical profession." (Doyle, *The Adventures of Sherlock Holmes*)

Ambiguous assertions. In some instances, events that appear to be clearly asserted based on semantic and syntactic indicators become ambiguous when considered outside of the narrative frame, such as when a narrator directly addresses the reader: "Why upon your first voyage as a passenger, did you yourself feel such a mystical vibration, when first told that you and your ship were now out of sight of land?" (Melville, *Moby Dick*)

4.1 Annotation process

All annotations were carried out by a single co-author after multiple rounds of discussions and the creation of a set of annotation guidelines heavily dependent on the ACE 2005 annotation guidelines for events (LDC, 2005) and adapted for the realis events under consideration here. To calculate the expected inter-annotator agreement rate, a second co-author independently annotated a random sample of five texts at the end of the annotation process, using only the annotation guidelines for reference. We find the agreement rate to be high (82.1 F-score for event identification and a chance-corrected Cohen's κ of 0.813).

The total dataset comprises 7,849 events among 210,532 tokens in the 100 books in our corpus, and is freely available for public use.

5 Event detection

We consider two classes of models for literary event detection in this data: neural models optimized for event trigger detection in past work (Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016; Feng et al., 2016); and featurized models (Ahn, 2006; Li et al., 2013; Yang and Mitchell, 2016).

5.1 Neural

Previous work has demonstrated the strength of neural models for event trigger detection, where models can leverage the distributional information encoded in word embeddings, along with representations of longer sentence context, to achieve high performance. We explore several variants of these models in this work; all models approach literary event detection as a sequence labeling problem, assigning a binary label to each token denoting its status as an event.

To leverage word representations that are suited for this particular literary domain, we train 100-dimensional skipgram (Mikolov et al., 2013) word embeddings on 15,290 books from Project Gutenberg. With the exception of the model incorporating BERT token representations, all models described below use these same embeddings.

LSTM. The simplest model we consider is a single-direction, 100-dimensional LSTM, with each input token represented as a word embedding from Project Gutenberg.

BiLSTM. Since the decision to label each token as an event may rely on information in the right context of the sentence, we consider a bidirectional LSTM (concatenating the outputs of two 100-dimensional LSTMs).

BiLSTM with document context. Most models for event trigger detection consider contextual information only from the *sentence* when making predictions about the event status of any individual token. Drawing on previous work incorporating global context (Liao and Grishman, 2010), we might hypothesize, however, that the accurate prediction of complex realis events may require greater document context—hypotheticals introduced in one sentence may span multiple ensuing paragraphs, while an extradiegetic aside from the narrator may span several pages and contain no concrete events. To test this, we define a sequence to be the entire (ca. 2,000-word) document, rather than an individual sentence.

BiLSTM with sentence CNN. Several previous methods have shown the strength of a sentence-level CNN (Nguyen et al., 2016; Feng et al., 2016). When predicting the event status of a token at position i in a sentence with n words $w = \{w_1, \ldots, w_n\}$, each CNN convolves over the entire sequence w along with position embeddings

$p = \{p_1, \ldots, p_n\}$ that encode the distance between each token position $j \in [1, n]$ and the target token i. We adopt the architecture of Nguyen and Grishman (2015) in particular, where the output of a CNN is then passed to a max-pooling phase to yield a representation c_i for target position i that is concatenated to the BiLSTM output o_i at that time step when making a binary prediction (with learned parameters W).

$$P(\text{event}) = \sigma([c_i; o_i]^\top W)$$

The CNN contains 200 filters (100 each scoped over word bigrams and trigrams). We encode positional information between the target token at position i and the token at position j using signed bucketing ($\pm$1, 2, 3, 4, 5, 6$-$10, 11$-$20, $>$20). Each bucket corresponds to a discrete choice of position with its own learned 5-dimensional embedding (as in past work).

BiLSTM with subword CNN. Subword character CNNs have been useful for a range of problems (Ma and Hovy, 2016; Chiu and Nichols, 2016) as a way of capturing meaningful representations of words that may be out-of-vocabulary for a set of learned embeddings (or whose use in a given domain may be at odds with the data those embeddings are trained on). We consider this design choice here as well. We represent each word as the output of a CNN with 100 filters (25 filters each scoped over character bigrams, trigrams, 4grams and 5grams), with max pooling over the character sequence to yield a 100-dimensional character representation c_i of a word at position i. This representation is then concatenated to the word embedding e_i for the token at that position and fed as input to the LSTM time step.

BiLSTM with BERT contextual representations. In order to take advantage of recent advances in language model pre-training (Howard and Ruder, 2018; Peters et al., 2018; Radford et al., 2019), we also incorporate contextual representations extracted from the pre-trained base BERT model (Devlin et al., 2019). Rather than fine-tuning the model for the supervised task, we instead use the BERT model in a feature-based way, representing each token in a sequence as the concatenation of the model's final four layers (3,072 dimensions in total) in place of pre-trained word embeddings in a BiLSTM. Since BERT uses WordPiece embeddings (Wu et al., 2016) as input, we take the average of any resulting sub-tokens in order to return a single per token representation (potentially beneficial as many of the literary works in our corpus contain long, complex words).

As Orr et al. (2018) have shown, neural models for event identification can exhibit substantial variation simply as a function of their random initialization, and we observe that with our data and models as well. To report expected performance on future data, we average together the predictions made from five random initializations (i.e., the majority class predicted for a token in context by the five models).

5.2 Featurized

The dataset we have created contains 7,849 events among 210,532 tokens. While this size is comparable to other datasets used for event detection in the past, it is unclear whether the scale is large enough to train highly parameterized neural models well; to test this, we design a linguistically informed featurized model, drawing on previous work in event representations (Ahn, 2006; Li et al., 2013; Chen et al., 2009) and noun phrase genericity and specificity (Reiter and Frank, 2010; Friedrich et al., 2015).

For this featurized model, we use ℓ_2-regularized binary logistic regression to make decisions about each token in its immediate context. We featurize the decision using the following information.

- **Word.** The lowercased word form of the token.
- **Lemma.** The lemma of the token.
- **POS.** The token's part of speech (using the Penn Treebank tagset), predicted using the SpaCy library.[1] In addition to providing important information about the core identification of verbs, the Penn Treebank tags also contribute to the determination of verb tense (important for our characterization of realis events).
- **Context.** The immediate context surrounding the word, represented as the following: a.) unigram indicators for the words found within three positions to the left; b.) indicators for words found three words to the right; c.) unigram $\times$ position indicators for those

[1] https://spacy.io

Method	Precision	Recall	F
Verbs only	17.7 [16.6-18.8]	76.2 [74.1-78.3]	28.7 [27.3-30.2]
Featurized	68.9 [66.2-71.7]	50.5 [48.0-52.9]	58.3 [56.1-60.4]
LSTM	66.6 [64.1-69.1]	60.5 [57.9-63.1]	63.4 [61.3-65.5]
BiLSTM	70.4 [67.8-72.9]	60.7 [58.0-63.4]	65.2 [63.1-67.3]
+ document context	74.2 [71.7-76.6]	58.8 [56.0-61.6]	65.6 [63.5-67.8]
+ sentence CNN	71.6 [69.1-74.1]	56.4 [53.8-59.0]	63.1 [61.0-65.1]
+ subword CNN	69.2 [66.6-71.6]	64.8 [62.2-67.3]	66.9 [64.8-68.9]
+ BERT	75.5 [73.3-77.8]	72.3 [69.7-74.8]	73.9 [72.0-75.7]
+ subword CNN	73.6 [71.2-75.8]	73.3 [70.8-75.7]	73.4 [71.5-75.2]

Table 3: Performance on literary event identification. All metrics are reported with 95% bootstrap confidence intervals.

same words (e.g., *not* appearing at position -1 with respect to the word); d.) the trigram appearing to the left; the trigram to the right; e.) the part-of-speech trigram to the left; and f.) to the right. This immediate contextual information captures important factors that affect modality, such as negation (Chen et al., 2009)

- **Syntax.** Syntactic information encoding the word's dependency relation, syntactic head, and part-of-speech of the syntactic head, predicted using SpaCy.

- **Wordnet.** Following Reiter and Frank (2010), we include WordNet synset and hyponymy information, capturing the synset of the word and the identities of its three hypernyms up the WordNet chain.

- **Embeddings.** We also include word embeddings as features; while a simple linear model like logistic regression cannot exploit important non-linearities between the embedding dimensions, they can provide some corpus level-information about the behavior of the word in the 15,290 Gutenberg texts it was trained on (which the neural models described above also have access to).

- **Bare plurals.** Some generic events (such as "pirates sail ships") contain bare plurals as subjects; inspired by Reiter and Frank (2010) on identifying generic noun phrases, we featurize the presence of a bare plural subject by noting whether the noun phrase subject is plural in form and lacking an explicit determiner, numeric count, or possessive pronoun. We also draw on their countability feature, identifying whether a noun phrase subject is countable (e.g., "the boy") or not (e.g. "the water") using CELEX (Baayen et al., 1996).

5.3 Results

To evaluate the performance of these models, we create training (60%), development (10%), and test (30%) partitions of the data at the level of books, with 60 books in train, 10 in development, and 30 in test. We stratify by book to ensure that no information from the same book appears in different partitions.

All models have access to the same development data for hyperparameter tuning; we use this to explore feature engineering and optimize the ℓ_2 regularization strength for the featurized model, and to explore different neural hyperparameter choices (e.g., size of LSTM).

Table 3 illustrates the comparative performance between the different systems. To contextualize these results, we also provide a simple but interpretable baseline of selecting all and only verbs to be events. This naive verb-only baseline yields an F-score of 28.7; while verbs are strong indicators of events, they are neither sufficient (the recall indicates that nearly one quarter of the true events in the test data are not verbs) nor entirely consistent (many verbs may signal events but not realis events).

While the featurized model improves on the baseline with an F-score of 58.3, all of the neural variants perform substantially better, generating a minimum F-score of 63.1. Although all neural models are statistically significantly better than the featurized model (under a bootstrap test), the variants of a subword CNN, sentence CNN and document context show little difference from

each other. In contrast, a BiLSTM with BERT input representations clearly outperforms all other methods with an F-score of 73.9 (an absolute improvement of +7.0 points over the best non-BERT model), attesting again to the value of unsupervised pre-training for supervised tasks (even in cases where the language model itself is not optimized for the task).

6 Analysis

To illustrate the usefulness of event representations for the analysis of literary texts, we consider the distinction between economic and cultural capital originally put forth by Bourdieu (1993) and analyzed from a computational perspective by Algee-Hewitt et al. (2016) and Underwood (2019). Both computational models find strong textual signals predictive of authorial prestige, measured either by inclusion in the Oxford *Dictionary of National Biography* (Algee-Hewitt et al., 2016) or by the number of times their works were reviewed by elite literary journals (Underwood, 2019). Both models also consider authorial popularity, measured either by the number of times a work was reprinted (Algee-Hewitt et al., 2016) or by the number of times their works can be found on historical bestseller lists (Underwood, 2019). While Underwood (2019) finds that high prestige fiction correlates with Harvard General Inquirer categories of KNOWLEDGE AND AWARENESS and NATURAL OBJECTS, we can similarly ask: is there a relationship in the depiction of realis events and literary prestige or popularity?

To test this, we draw on data from Underwood (2019), selecting the 100 authors identified in that work with the highest and lowest prestige, respectively. In total, 44 of the high prestige authors and 29 of the low prestige authors are present in the Project Gutenberg corpus. We select any works of fiction by these authors that are present in Gutenberg, limiting the maximum number of novels per author to 10. This yields 190 novels in the high prestige class and 159 in the low prestige class. Since Project Gutenberg has wider representation of historically popular texts than unpopular ones, we select the 100 most popular authors and 500 least popular authors. 67 of the high popularity authors and 68 of the low popularity authors appear in the Gutenberg corpus. After selecting a sample of the high popularity texts while again limiting per author novel totals to 10, this yields 182 novels in the high popularity class and 173 in the low popularity class.

We run the best-performing literary event detection model identified above (a bidirectional LSTM with BERT token representations) on each novel, and carry out two related analyses on the output. First, to estimate the overall incidence of realis events, we simply calculate the average event ratio in each novel (the number of realis events normalized by the number of tokens); second, to capture the pacing of realis events more concretely in terms of actual tokens, we invert this metric to calculate the event distance (how many tokens one would have to read on average before coming across an event token).

Class	Ratio	Distance
High prestige	4.6 [4.4-4.7]	23.4 [22.4-24.5]
Low prestige	5.5 [5.3-5.6]	19.2 [18.2-20.1]
High popularity	4.6 [4.4-4.8]	23.2 [22.3-24.1]
Low popularity	4.5 [4.3-4.7]	25.0 [21.9-28.1]

Table 4: Mean event ratios (event tokens / total tokens) and mean event distances (total tokens / event tokens) calculated over all novels in each class. All metrics are reported with 95% confidence intervals.

The results of these analyses are shown in Table 4. We would expect that the pulp novels of Edgar Rice Burroughs would contain more physical description and concrete events than the more meditative novels of Henry James, James Joyce, and Kate Chopin, and we find this to be the case: authors with low prestige use 20% *more* concrete events in their works (the difference in both metrics between the two groups is statistically significant at $p < 0.05$). For the popularity dimension, however, the results on both metrics are statistically indistinguishable.

Although it is difficult to draw definitive conclusions based on these results, the outcome for the prestige dimension in particular indicates a compelling line of inquiry. In fact, the results in Table 4 only tell half the story. As Figure 1 demonstrates, the most marked distinction for event ratios in high prestige and low prestige novels is not the mean but rather the spread. High prestige novels appear to have greater variability in the percentage of realis events (particularly skewed to lower ratios), whereas the percentage for low prestige novels, with the exception of a few outliers, remains within a smaller range. This variability suggests that, as one might expect, prestigious au-

thors tend to conform less programmatically to a regular frequency of realis events. Put differently, prestigious novels don't have the same constraints as less prestigious ones in maintaining our attention through something *happening* in the narrative. While many prestigious novels have event ratios in line with novels lacking prestige, prestigious authors appear to have a higher degree of freedom when it comes to the overall eventfulness of their works.

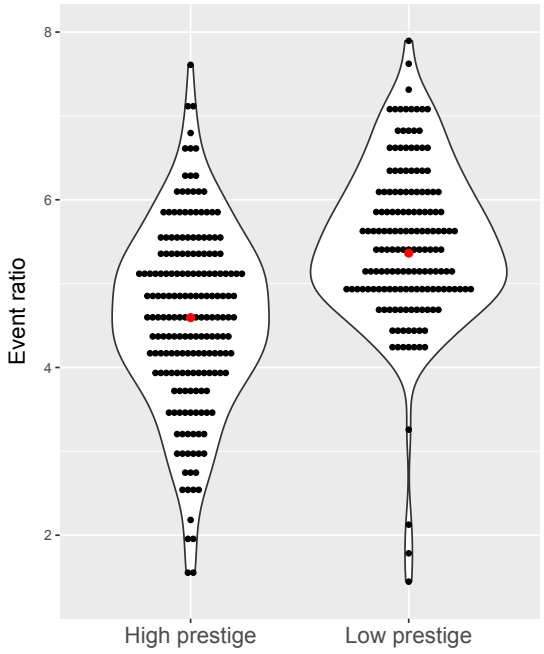

Figure 1: Violin plot of event ratios for novels in the prestige category.

7 Conclusion

We present in this work a new dataset for the representation of events in literary texts in order to bridge the gap between previous efforts to represent fact-based accounts in news (along with contemporary models trained on that data) and the demands of literary scholars for the computational analysis of the micro-narratives that comprise plot.

The relatively straightforward application of our model to the analysis of authorial prestige shows how identifying realis events can help to uncover some important and overlooked aspects of novelistic narrative. To the best of our knowledge, no previous technical or theoretical work has specifically examined the function that events with asserted realis play in the structure of literary fiction. Yet simply by analyzing the ratio of realis events, one can capture a meaningful distinction between novels written by authors whose works are reviewed by elite literary journals and those written by authors whose work is not. We hope this initial application inspires further research by literary scholars and computational humanists in the future.

All event annotations are freely available for public use under a Creative Commons Sharealike license at https://github.com/dbamman/litbank. Code to support this work can be found at: https://github.com/dbamman/ACL2019-literary-events.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback, and Ted Underwood for sharing data to enable our analysis on popularity and prestige. The research reported in this article was supported by an Amazon Research Award and by resources provided by NVIDIA and Berkeley Research Computing.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and Framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53. Association for Computational Linguistics.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning About Time and Events*, ARTE '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. Canon/archive: Large-scale dynamics in the literary field. Literary Lab Pamphlet 11.

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.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. Celex2. LDC.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. NAACL.

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 (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Ann Bies, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, and Tim O'Gorman. 2016. A comparison of event representations in deft. In *Proceedings of the Fourth Workshop on Events*, pages 27–36. Association for Computational Linguistics.

Pierre Bourdieu. 1993. The field of cultural production, or: The economic world reversed. In *The Field of Cultural Production: Essays on Art and Literature*. Columbia University Press.

Peter Brooks. 1992. *Reading for the Plot: Design and Intention in Narrative*. Harvard University Press.

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daumé III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Association for the Advancement of Artificial Intelligence*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176. Association for Computational Linguistics.

Zheng Chen, Heng Ji, and R Haralick. 2009. Event coreference resolution: Algorithm, feature impact and evaluation. In *Proceedings of Events in Emerging Text Types (eETTs) Workshop, in conjunction with RANLP, Bulgaria*.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71. Association for Computational Linguistics.

E. M. Forster. 1927. *Aspects of the Novel*. Edward Arnold.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 21–30.

Gérard Genette. 1983. *Narrative Discourse: An Essay in Method*. Cornell University Press.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

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 *North American Association for Computational Linguistics*.

Matthew Jockers. 2015. Revealing sentiment and plot arcs with the syuzhet package. http://www.matthewjockers.net/2015/02/02/syuzhet/.

LDC. 2005. ACE (Automatic Content Extraction) English annotation guidelines for events. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82. Association for Computational Linguistics.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 789–797, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018. Event detection via gated multilingual attention mechanism. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguist.*, 38(2):301–333.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR*.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114. Association for Computational Linguistics.

Franco Moretti. 2013. *The Bourgeois: Between history and literature*. Verso Books.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *AAAI*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371. Association for Computational Linguistics.

Walker Orr, Prasad Tadepalli, and Xiaoli Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 999–1004. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Vladimir Propp. 2010. *Morphology of the Folktale*. University of Texas Press.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 1–11.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauskas, Marcia Lazo, Andrea Setzer, and Beth Sundheim. 2003b. The TimeBank corpus. *Corpus Linguistics*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 40–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*.

Hans-Jürgen Sasse. 2002. Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state. *Linguistic Typology*, 6(2):199–271.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Viktor Shklovsky. 1990. *Theory of Prose*. Dalkey Archive.

R. Sprugnoli and S. Tonelli. 2017. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485506.

Beth M. Sundheim. 1991. Overview of the third message understanding conference. In *Processing of the Third Message Understanding Conference*.

Ted Underwood. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Hardik Vala, David Jurgens, Andrew Piper, and Derek Ruths. 2015. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 769–774, Lisbon, Portugal. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. LDC.

Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299. Association for Computational Linguistics.