Modeling Reportable Events as Turning Points in Narrative

Jessica Ouyang
Department of Computer Science
Columbia University
New York, NY 10027
ouyangj@cs.columbia.edu

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY 10027
kathy@cs.columbia.edu

Abstract

We present novel experiments in modeling the rise and fall of story characteristics within narrative, leading up to the *Most Reportable Event* (MRE), the compelling event that is the nucleus of the story. We construct a corpus of personal narratives from the bulletin board website Reddit, using the organization of Reddit content into topic-specific communities to automatically identify narratives. Leveraging the structure of Reddit comment threads, we automatically label a large dataset of narratives. We present a change-based model of narrative that tracks changes in formality, affect, and other characteristics over the course of a story, and we use this model in distant supervision and self-training experiments that achieve significant improvements over the baselines at the task of identifying MREs.

1 Introduction

What is a narrative? In one of the early linguistic analyses of storytelling, Prince (1973) defines a story as describing an event that causes a change of state. Prince's minimal story has three parts: the starting state, the ending state, and the event that transforms the stating state into the ending state. An example of a minimal story is as follows:

> A man was unhappy, then he fell in love, then as a result, he was happy.

Polanyi (1976) notes that minimal stories are toy examples that would never hold an audience's interest. So what makes a story interesting?

Labov (1967; 1997) defines a well-formed narrative as a series of actions leading to a *Most Reportable Event* (MRE). The MRE is the point of the story – the most unusual event that has the greatest emotional impact on the narrator and the audience. For a story to be interesting, Prince's change-of-state event should be an MRE.

The following is an example of a narrative from the corpus we create in this work, with the sentence containing the MRE emphasized:

> This isn't exactly creepy, but it's one of the scariest things that's ever happened to me. I was driving down the motorway with my boyfriend in the passenger seat, and my dad in the seat behind my own. My dad is an epileptic and his fits are extremely sporadic. Sometimes he goes extremely stiff and other times he will try to get out of places or grab and punch people. *Mid-conversation I felt his hands wrap around my throat as I was driving, pulling my head back and making it increasingly difficult to drive.* My boyfriend managed to help steer the car into the hard shoulder but it was one of the scariest experiences in my life.

The MRE is the shortest possible summary of a story; it is what we would say about the story if we could only say one thing. If we could identify the MRE of a narrative, we could automatically generate summaries or headline-style titles for online stories. Detecting MREs could also allow us to explore how storytellers build emotional impact as they lead up to the climaxes of their stories.

In this work, we present a novel approach to modeling narrative in order to automatically identify the MRE. The MRE is a real world event underlying the story and thus is difficult to infer; instead, we identify sentences that describe or refer to it. We incorporate Prince's change-of-state formalization as well as Labov's definition of the MRE by modeling changes in story characteristics suggested by Prince, Polanyi, and Labov, such as measures of syntactic complexity and emotional

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2149–2158,
Lisbon, Portugal, 17-21 September 2015.

content. If Prince and Labov are both correct, we should find the MRE at a point of change in the story and in our story characteristics.

We create a corpus of thousands of personal narratives collected from Reddit, a social bulletin board website organized into topic-specific 'subreddit' communities. We automatically label most of this data using heuristics based on the comment-thread structure of Reddit content. Using this corpus, we conduct two experiments in classifying sentences of a story as containing the MRE or not: the first using distant supervision, and the second using self-training.

In Section 2, we discuss prior work on automatically identifying personal narratives, as well as related experiments using Labov's theory of narrative analysis. Section 3 discusses data collection and labeling. Sections 4-5 present our change-based model of narrative and our experiments. Finally, Section 6 discusses our experimental results and proposes directions for future work.

2 Related Work

Prior work using Labov's theory of narrative has focused on classifying clauses by their function.

Rahimtoroghi et al. (2013)worked on 20 of Aesop's fables. The 315 clauses were manually annotated with the three labels of Labov and Waltezky (1967), Orientation (background information), Action (events), and Evaluation (author's perspective), which we discuss in Section 4. Rahimtoroghi et al. used two annotators with high agreement and achieved accuracy and precision around 0.9 on all three labels, as well as recall above 0.9 on all but Orientation. They noted that their data set was very clean: interannotator agreement was nearly perfect, the language was simple, and each clause served a clear narrative purpose.

Ouyang and McKeown (2014) explored identifying the Action chain of the oral narratives in Labov (2013). They used a dataset of 49 narratives (1,277 clauses), transcribed from recordings of speech and annotated by Labov and achieved 0.72 f-score on classifying clauses as Action or not. This task is easier than our proposed task of identifying sentences containing MREs. Actions account for nearly half the clauses in the Labov (2013) dataset, while there are only and average of 2.5 MRE sentences per story. Additionally, identifying Labov's Actions is a problem of detecting causal and temporal relations among events; identifying the MRE is a problem of measuring how impactful and shocking an event is.

Swanson et al. (2014) used 50 stories, which were annotated with an extended label set by three annotators, and each of the 1,602 clauses was assigned the label given by the majority of annotators. The extended label set was then mapped to Labov and Waletzky's three labels. Nearly half of the clauses in this dataset are Evaluations, and Orientations and Actions each make up nearly one quarter of the dataset. Swanson et al. achieved 0.69 overall f-score on three-way classification of clauses. Again, this task is less difficult than our proposed task. The three labels, Orientation, Action, and Evaluation have distinct functions that are reflected in tense, mood, and a clause's position in the narrative. The MRE is not a sentence or clause but an event that may be described or referred to by any sentence in a narrative; it is distinguished from the other events only by its surprisingness and emotional impact, dimensions that are difficult to model computationally without a deep semantic understanding of the story.

The stories that Swanson et al. used were drawn from a corpus drawn from weblog posts (Gordon and Swanson, 2009). Gordon and Swanson used unigram features to classify posts as either stories or not, achieving 75% precision. They note that only about 17% of weblog text consists of stories.

In contrast to the relatively small datasets used by Rahimtoroghi et al., Ouyang and McKeown, and Swanson et al., we use a larger dataset automatically collected from Reddit. Our collection method achieved 94% precision in identifying narratives. A number of researchers have characterized the structure and use of Reddit, currently the 26th most popular website in the world[1]. Weninger et al. (2013) described the structure of Reddit comment threads. Gilbert (2013) measured user participation in the voting process that ranks Reddit content. Singer et al. (2014) conducted a longitudinal study of the Reddit user community, finding a trend favoring original, user-generated content.

3 Data

3.1 Collection

We collected data from the AskReddit subreddit, where users post questions for other members of the community, who reply with comments answering the questions. Table 1 shows some examples

[1] http://www.alexa.com/siteinfo/reddit.com

of these posts, and we can see some of the wide variety of story topics found on AskReddit.

Post Title
Whats your creepiest (REAL LIFE) story?
Your best "Accidentally Racist" story?
What are your stories of petty revenge?

Table 1: Examples of AskReddit posts.

Using PRAW[2], we scraped the top 50 AskReddit posts containing the keyword 'story.' Of these posts, 10 were tagged as NSFW ('not safe for work'), indicating they contained adult content; we did not use these posts in this work, as we felt the language would be too different from that used in posts without the tag. Another 3 posts did not contain personal narratives, and instead were about fictional stories in movies or music.

With the 37 remaining posts, we treated each top-level comment (those that replied directly to the posted question) as a story. The example given in Section 1 is one such story. We collected 6,000 top-level comments and discarded those without comment threads replying to them. As we discuss in Section 3.2, we use comment threads to automatically label our training data. We tokenized the top-level comments by sentence (Bird et al., 2009) and removed all sentences following any variation of the word 'EDIT', as these were usually responses to readers' comments. We discarded texts with fewer than three sentences, based on Prince's definition of a minimal story as consisting of a starting state, an event, and an ending state. We are left with 4,896 stories, with an average length of 16 sentences and a maxiumum of 198.

3.2 Labeling

We partitioned our data into development, seed, and tuning sets of 100 stories each; a testing set of 200 stories; and a training set of 4,178 stories. The development, seed, tuning, and testing sets were manually annotated by a native English speaker (not one of the authors), who was instructed to label all sentences that contained or referred to the MRE. For convenience, from here on, we will use the term 'MRE' to refer to both the Most Reportable Event itself (of which there can only be one per narrative) and to sentences that contain or refer to it (of which there can be more than one).

To measure interannotator agreement, we also had a second annotator (also a native English speaker and not one of the authors) label MREs in the 100 narratives in our development set. We found substantial agreement (Cohen's $\kappa = 0.729$); the two classes, MRE and not-MRE, are highly unbalanced, so percent agreement between the two annotators was extremely high (95%).

In addition to labeling the MREs, our first annotator identified and discarded 31 texts that were not true stories, but rather Reddit-specific inside jokes or comments on how cool the stories in the thread were. From this, we can see that the precision of our story collection method is very high. Gordon et al. (2007) found that stories were 17% of the weblog text that they collected; of the 500 texts given to our annotators, 94% were stories.

Using the development set, we experimented with seven heuristics, defined below, for automatically labeling the training set. Each predicts a sentence index s_h to be the index of an MRE. We measured the performance of each heuristic using root-mean-square error (RMSE), which measures the standard deviation of how far the heuristic's predictions fall from a true MRE.

$$\text{Let } N = \text{number of narratives}$$
$$s_{MRE} = \text{index of a true MRE}$$
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_{MRE_i} - s_{h_i})^2}$$

We used a linear combination of three heuristics with the lowest RMSE to label our training set.

Similarity to comment. The bag-of-words cosine similarity between a sentence and comments replying to the story. We expect comments to refer to the MRE because of its shocking nature and importance to the story. This heuristic achieved RMSE of 5.5 sentences on the development set.

Similarity to tl;dr. The latent semantic similarity between a sentence and the *tl;dr*. The tl;dr (too long; didn't read) is a very short paraphrase of a post given by its author. They are relatively rare – 663 stories, or 14% of our data, had tl;drs. Since the MRE is the central event of the story, we expect it to be included in the tl;dr. We calculated the similarity using the weighted matrix factorization algorithm described by Guo and Diab (2012). This heuristic achieved RMSE of 5.8 sentences.

[2]https://praw.readthedocs.org/en/v2.1.20/, Python Reddit API Wrapper

In contrast, bag-of-words cosine similarity to the tl;dr performed poorly (RMSE of 13.2). This is due to the tl;dr being both short and a paraphrase of its story. There are few words in the tl;dr, and those words are often synonyms of, but not the same as, words in the story. Guo and Diab's latent semantic similarity score addresses this word sparsity problem by modeling words that are not present in the input text. We also experimented with latent semantic similarity for the similarity-to-comment and similarity-to-prompt heuristics, but in these two cases, it did not perform as well as the bag-of-words cosine similarity.

Similarity to prompt. The bag-of-words cosine similarity between a sentence and the AskReddit post that prompted the story. The story should be relevant to the prompt, so we expect the MRE to be similar to the prompt text. This heuristic achieved RMSE of 6.3 sentences.

We used the heuristic with the fourth lowest RMSE as one of the baselines in our experiments:

Last sentence. The last sentence in the story. Since the events of a story build up to the MRE, the MRE should occur near the end of the story. This heuristic achieved RMSE of 6.9 sentences.

Other heuristics. We also tried the following:

- Single-sentence paragraph (RMSE of 8.7). This heuristic was meant to capture emphasis, as an MRE might be placed in its own, separate paragraph to draw attention to it.
- First sentence (RMSE of 13.7). Narratives occasionally open with a brief introductory paragraph that summarizes the events to come. This heuristic was meant to capture a reference to the MRE in this introduction.

The training set was automatically labeled using a linear combination of the three best-performing heuristics: similarity to comment, similarity to tl;dr, and similarity to prompt.

$$h_{\text{label}} = 0.2 * h_{\text{comment}} + 0.5 * h_{\text{tldr}} + 0.3 * h_{\text{prompt}}$$

This outperformed each of the three alone, achieving an RMSE of 5.1 sentences. The weights for each heuristic were tuned on the development set. For stories without a tl;dr, that heuristic was set to 0. The sentence in the story with the highest heuristic score was selected as the MRE.

In 52 of the 99 stories in the development set, we found that multiple, consecutive sentences were labeled by our annotator as MREs. The average number of consecutive MREs was 2.5 sentences. To reflect this, we labeled our training set in three-sentence blocks. The sentence selected by our labeling heuristic, along with the immediately preceding and following sentences, were all labeled as MREs. The result was the weakly-labeled training set in Table 2.

		Number of Sentences	
Data Set	Stories	MRE	Total
dev*	99	169	1528
seed*	82	184	958
tuning*	95	212	1301
testing*	193	444	2771
training	4178	11205	67954

Table 2: Distribution of labels (*manual).

4 Modeling Narrative

Our approach to modeling narrative is based on both Labov (2013) and Prince (1973). We claim that Labov's MRE is Prince's change of state with the added requirement of reportability or interestingness – in fact, all three components of Prince's minimal story have equivalences in Labov.

Labov and Waletzky (1967) proposed three components of narrative: the Orientation, which we equate with Prince's starting state; the Action, the chain of events culminating in the MRE; and the Evaluation, the author's perspective on the story. Labov (2013) adds three more components: the Resolution, equivalent to Prince's ending state, and the Abstract and Coda, where the author introduces and concludes the story.

We focus on Prince's claim that stories are about change. Polanyi (1985) observes that the turning point of a story is marked by a change in style, formality of language, or emphasis in the telling of the story. Labov (2013) likewise observes that a change in verb tense often accompanies MREs. We hypothesize that the MRE should be found at a point of change in the story.

We score each sentence according to three views of narrative: syntax, semantics, and affect.

Syntax. We model Polanyi's claim that a change in formality marks the changing point by including metrics of sentential syntax; we use the syntactic complexity of a sentence as an approximation for formality. The complexity of a sentence also reflects emphasis – short, staccato sentences bear more emphasis than long, complicated ones. We use the length of the sen-

tence, the length of its verb phrase, and the ratio of these two lengths; the depth of the sentence's parse tree (Klein and Manning, 2003), the depth of its verb phrase's subtree, and the ratio of these two depths. We also use the average word length for the sentence and the syntactic complexity formula proposed by Botel and Granowsky (1972), which scores sentences on specific structures, such as passives, appositives, and clausal subjects. Finally, we use the formality and complexity dictionaries described in Pavlick and Nenkova (2015), which provide human formality judgments for 7,794 words and short phrases and complexity judgments for 5,699 words and phrases. We score each sentence by averaging across all words and phrases in the sentence.

Semantics. As the MRE is surprising and shocking, we expect it to be dissimilar from the surrounding sentences; we use semantic similarity to surrounding sentences as a measure of shock. Our semantic scores are the bag-of-words cosine and the latent semantic similarity scores for adjacent sentences (Guo and Diab, 2012).

Affect. A change in affect reflects a change in style, and we expect the MRE to occur at an emotional peak. We use the Dictionary of Affect in Language (DAL) (Whissell, 1989), augmented with WordNet for coverage (Miller, 1995). The DAL represents lexical affect with three scores: evaluation (*ee*, hereafter 'pleasantness' to avoid confusion with Labov's Evaluation), activation (*aa*, activeness), and imagery (*ii*, concreteness). We also use a fourth score, the activation-evaluation (AE) norm, a measure of subjectivity defined by Agarwal et al. (2009):

$$norm = \frac{\sqrt{ee^2 + aa^2}}{ii}$$

For each of these four word-level scores, we calculate a sentence-level score by averaging across the words in the sentence using the finite state machine described by Agarwal et al. We expect the sentences surrounding an MRE to be more subjective and emotional as the impact of the MRE becomes clear. We also expect a build-up in activeness and intensity, peaking at the MRE.

To model change over the course of a narrative, we look for changes in the syntactic, semantic, and affectual scores. To illustrate this, Figure 1 shows the activeness and pleasantness DAL scores for the example narrative given in Section 1. We can see how the MRE is the most exciting sentence in the story – global maximum in activation – as well as the most horrifying – global minimum in pleasantness. The overall shape of the activeness scores reflects Prince's three components of a minimal story: low initial activation (starting state) and low final activation (ending state) with a build up to a peak at the MRE (change in state) between them.

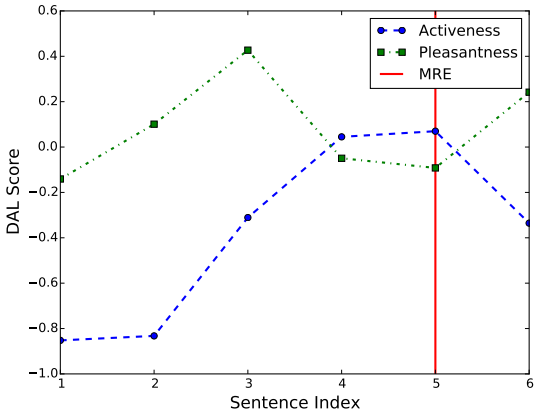

Figure 1: Activeness and pleasantness scores.

5 Experiments

Using our Reddit dataset and change-based model of narrative, we conducted two experiments on automatically identifying MREs. We compare our results with three baselines: random, our labeling heuristic, and the last sentence of the story (best-performing heuristic not used in labeling).

As described in Section 3.2, we labeled our training set in blocks of three consecutive MREs, centered on the sentence from each narrative that was selected by our heuristics. To account for this, in our experiments and baselines, we predicted the presence of an MRE in a three-sentence block. In testing, we considered a predicted block to be correct if it contained at least one gold-label MRE.

5.1 Features

Change-based Features. For each of the fifteen metrics in Section 4, shown in Table 3, we first smooth the scores by applying a Gaussian filter. We also tried weighted and exponential moving averages, as well as a Hamming window, but the Gaussian performed best in experiments on our tuning set. We then generate 11 features for each sentence: the metric score at the sentence; whether or not the sentence is a local maximum or minimum; the sentence's distance from the global

Type	Metric Names
Syntactic	sentlength, vplength, lengthratio, sentdepth, vpdepth, depthratio wordlength, structcomplexity, wordformality, wordcomplexity
Semantic	cossimilarity, lssimilarity
Affectual	pleasantness, activation, imagery, subjectivity

Table 3: The fifteen metrics for change.

maximum and minimum; the difference in score between the sentence and the preceding sentence, the difference between the sentence and the following sentence, and the average of these differences (approximating the incoming, outgoing, and self- slopes for the metric); and the incoming, outgoing, and self- differences of differences (approximating the second derivative).

Other Features.

- The tense of the main verb and whether or not there is a shift from the previous sentence. Labov (2013) suggests a shift between the past and the historical present near the MRE.
- The position of the sentence in the narrative.
- The bag-of-words cosine similarity and latent semantic similarity between the sentence and the first and second sentences in the narrative. The MRE usually appears near the end of a story, but Labov (2013) notes that the Abstract, a short introduction that occurs in some narratives, often refers to the MRE.

5.2 Distant Supervision

Our first experiment used a distant supervision approach with our automatically-labeled training set. Distant supervision has previously been applied to NLP problems such as sentiment analysis (Go et al., 2009; Purver and Battersby, 2012; Suttles and Ide, 2013) and relation extraction (Mintz et al., 2009; Yao et al., 2010; Hoffmann et al., 2011; Nguyen and Moschitti, 2011; Krause et al., 2012; Min et al., 2013; Xu et al., 2013).

We classify blocks of three sentences as containing the MRE or not. The two classes, *MRE* and *not-MRE*, were weighted inversely to their frequencies in the weakly-labeled set, and all features were normalized to the range $[0, 1]$. We trained an SVM with margin $C = 1$ and an RBF kernel with $\gamma = 0.001$, chosen using grid search on our tuning set (Pedregosa et al., 2011).

Trial	Precision	Recall	F-Score
Last sent. baseline	0.208	0.112	0.146
Heuristic baseline	0.107	0.333	0.162
No change*	0.146	0.378	0.211
Random baseline	0.185	0.586	0.281
Change only*	0.351	0.685	0.466
All features*	**0.398**	**0.745**	**0.519**

Table 4: Distant supervision results (*$p < 0.01$).

The results of the distant supervision experiment are shown in Table 4. Our best results use all features, but, notably, using the change-based features alone achieves significant improvement over the three baselines ($p < 0.00005$). The 'no change' trial used the metric scores themselves and the 'other' features but none of the change-based features, such as slopes and proximity to global extremes. This feature set was outperformed by the random baseline ($p < 0.0024$), supporting our hypothesis that it is change in a metric, rather than the score itself, that predicts MREs.

Because we used an non-linear kernel, we were not able to examine feature weights directly. Instead, Table 5 shows the results of a logistic regression model trained on our features. The 10 best features are shown, along with their weights and 95% confidence intervals. From feature 8, we see that the MRE is found in sentences near the narrative's global minimum in imagery (the Evaluation), but feature 1 indicates that sentences containing the MRE show a sharp increase in imagery compared to the previous sentences. The MRE is described in a burst of vivid language, followed by more abstract author opinions .

Features 2 and 9 indicate that the MRE tends to be described using informal language – a textual echo to Labov's observation that the subjects of his sociolinguistic interviews spoke less formally and more colloquially as they relived the climaxes of their stories (Labov, 2013). Feature 3 suggests that sentences containing the MRE are similar to the surrounding sentences. While we expected MRE sentences to be different from their neighbors due to the unusual and shocking nature of the MRE, this feature seems instead to reinforce the idea that MREs tend to described over the course of multiple, consecutive sentences, rather than in a single

	Feature Name	Weight	Confidence Interval
1.	incomingd2_imagery	4.174	(4.062, 4.287)
2.	distancefrommin_wordformality_neg	4.109	(3.952, 4.265)
3.	cossimilarity_adjacent	3.618	(3.425, 3.812)
4.	distancefrommin_activeness	3.377	(2.855, 3.298)
5.	sentdepth	3.364	(3.138, 3.590)
6.	distancefrommin_wordlength_neg	3.321	(3.018, 3.624)
7.	distancefrommin_vpdepth	3.034	(2.823, 3.247)
8.	distancefrommin_imagery_neg	2.790	(2.524, 3.056)
9.	wordformality_neg	2.329	(2.226, 2.432)
10.	incomingd2_vplen	2.128	(1.938, 2.318)

Table 5: Top 10 features.

sentence. From feature 4, we see, as expected, that the MRE is far from the narrative's global minimum in activeness, as it is the end of a chain of events, far away from the stative Orientation.

Finally, features 5 and 10 suggest that MRE sentences are not only long, but much longer than the preceding sentences, and feature 6 indicates that MRE sentences are close to the global minimum in average word length. Shorter average word length is expected, as an indicator of both informal word choice and emphasis. Long sentences, however, suggest a domain difference between our work on text and Labov's work on transcribed speech. Looking over our development set, we find that many authors combine the description of the MRE with evaluative material in a single sentence, resulting in a longer and more syntactically complex MRE sentence than is found in Labov's data.

5.3 Self-Training

Our second experiment used a self-training approach, where a classifier uses a small, labeled seed set to label a larger training set. Self-training has been applied to parsing (McClosky et al., 2006; Reichart and Rappoport, 2007; McClosky and Charniak, 2008; Huang and Harper, 2009; Sagae, 2010) and word sense disambiguation (Mihalcea, 2004). With the same parameters as in the distant supervision experiment, we trained an SVM on our hand-labeled seed set of 958 sentences. We used this initial model to relabel the training set. All sentences where this labeling agreed with our automatically-generated heuristic labels were added to the seed set and used to train a new model, which was in turn used to label the remaining sentences, and so on until none of the current model's labels agreed with any of the remaining heuristic labels. Figure 2 shows the learning curve for the self-training experiment, along with the growth of the self-training set.

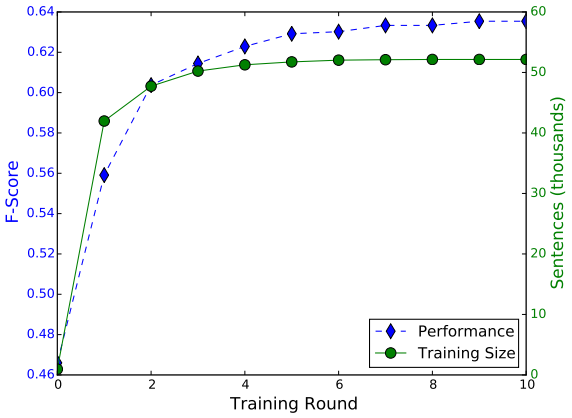

Figure 2: Learning and training set size curves.

The results of the self-training experiment are shown in Table 6. We achieve the best performance, f1 = 0.635, after 9 rounds of self-training. Self-training terminated after 10 rounds, but the 10th round had no effect on performance.

Trial	Precision	Recall	F-Score
Random baseline	0.185	0.586	0.281
Seed only*	0.374	0.617	0.466
Dist. supervision*	0.398	0.745	0.519
Self-training*	**0.478**	**0.946**	**0.635**

Table 6: Best self-training results (*$p < 0.01$).

The initial model, trained only on the seed set, performed nearly as well as our distant supervi-

sion experiment. This illustrates that quantity of data does not overcome the use of accurate manual labels on a small dataset. As described in Section 3.2, the distant supervision labels were based on a linear combination of three heuristics that achieved at best an RMSE of 5.1 sentences. However, with self-training, we can exploit the noisy heuristic labels by using only those labels that agree with the seed-trained model, thus reducing the amount of noise. 52,147 of the 67,954 weakly-labeled sentences were used in self-training.

6 Discussion and Future Work

Identifying MREs is a hard problem. A human annotator can rely on world knowledge to find the most shocking and impactful event in a story, but we do not have access to that knowledge. Additionally, MREs are rare, comprising 15% of the sentences in our hand-annotated datasets. MREs comprise just over 16% of our weakly-labeled training set, but as we discuss below, there is too much noise in the automatically-generated labels.

Despite the difficulty of the task, our experiments show that our change-based model of narrative is effective for identifying MREs, and this model provides evidence supporting the change-in-state view of narrative suggested by Prince (1973), Polanyi (1985), and Labov (1997). We achieve high recall with self-training (95%), but precision is low across the board. This suggests that, while MREs do occur at extremes in syntactic complexity, semantic similarity, and emotional activation, there may be many non-MRE local extremes throughout a narrative.

Examining our results, we find a few common sources of error. False positive sentences tend to have high imagery and activeness. In Table 5, we saw that imagery and activeness alone do not indicate the presence of the MRE. An MRE sentence is not just active; it is separated from the stative introduction by the other events of the story. Nor is it enough for a sentence to have high imagery; the MRE is more vividly described than the preceding events – we see again the importance of change in our model of narrative. False negatives tend to have high scores in syntactic complexity and formality. As low formality was one of our stronger predictors of MRE sentences, we may need to adjust these features in future work.

We also hope to refine our automatically-generated labels in future work. Our self-training experiment showed that 27% of our automatically-generated labels were too noisy to use. We also hope to improve our filters for automatically discarding non-story text. We currently reject texts shorter than three sentences, based on Prince's three-part definition of a story. In spite of this filtering, 7% of our 500 manually-labeled texts were identified as non-stories by our annotator. Extrapolating to our training set, we suspect that over 300 of our training 'narratives' are not narratives at all.

Finally, we hope to explore other theories of narrative analysis that could suggest new ways to quantify change in narrative. Prince, Polanyi, and Labov propose a high-level view of personal narrative: stories are centered around reportable events that cause a change in state for the author. This work tested fifteen surface-level features that reflect this change in state. Are there others? Or is a deeper semantic understanding of the starting and ending states of stories required?

7 Conclusion

We have described a new model of narrative based on Prince (1973), Polanyi (1985), and Labov (1997). Our model tracks story characterstics over the course of a narrative, capturing change in complexity, meaning, and emotion.

We have created a corpus of 4,896 personal narratives, taking advantage of AskReddit, a community where members often prompt each other for stories. Our experiments on this corpus show that our change-based model is able to identify MREs. They also demonstrate that large quantities of hand-labeled data are not required for this task. Our distant supervision and self-training approaches successfully use data weakly labeled using heuristic rules that leverage the comment thread structure of Reddit content. We believe these Reddit stories are representative of the short, personal narratives found online in blogs or discussion forums, and so this work should be useful for finding MREs in a variety of online personal narratives. The one difference between this data and stories from other online sources is the prompt. A personal narrative posted to someone's personal blog is unlikely to have a prompt. We use the prompt for our heuristic labeling, so our automatic labels on non-Reddit data may be noisier, but many blog posts also have titles or tags that may be just as useful.

Identifying MREs is a hard problem that has not

previously been addressed in work on computational narrative. We have shown that the high-level view proposed by linguistic theories of narrative – that stories are about change – holds true. Measuring change over the course of a narrative yields better results than other features and baselines.

Why do we care about MREs? Polanyi (1976) asserts that "one does not produce a narrative text for no reason at all." The Most Reportable Event is that reason. It is the point of the story; the shortest possible summary; the answer to the question, "So what?". It could be used to generate titles or summaries to be used in organizing stories for readers to browse, or it could be used in recommendation systems to help readers find related stories. In future work we hope to be able to generate a text description the full MRE, which would be better suited to summarization or generating headlines, rather than identifying sentences that refer to it. We hope this work will encourage others to further investigate the Most Reportable Event.

Acknowledgments

This work was partially supported by NSF Contract No. IIS-1422863.

References

Apoorv Agarwal, Fadi Biadsy, and Kathleen McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Langauge Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

Morton Botel and Alvin Granowsky. 1972. A formula for measuring syntactic complexity: A directional effort. *Elementary English*.

Eric Gilbert. 2013. Widespread underprovision on reddit. *Proceedings of the 2013 conference on Computer supported cooperative work*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*.

Andrew S. Gordon, Qun Cao, and Reid Swanson. 2007. Automated Story Capture From Internet Weblogs. *Proceedings of the Fourth International Conference on Knowledge Capture*.

Andrew S. Gordon and Reid Swanson 2009. Identifying Personal Stories in Millions of Weblog Entries. *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1*

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.

Zhongqiang Huang and Mary Harper. 2009. Self-Training PCFG grammars with latent annotations across languages. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*.

Eric Jones, Travis Oliphant, Pearu Peterson, and others. 2001. SciPy: Open source scientific tools for Python. http://www.scipy.org/ Online; accessed 26 Jan. 2015.

Dan Klein and Christopher Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*.

Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. 2012. Large-scale learning of relation-extraction rules with distant supervision from the web. *The Semantic WebISWC 2012*, 263-278. Springer Berlin Heidelberg, Berlin.

William Labov. 1997. Some further steps in narrative analysis. *Journal of Narrative and Life History*, 7:395-415.

William Labov. 2013. *The Language of Life and Death*. Cambridge University Press, Cambridge, UK.

William Labov and Joshua Waletzky. 1967. Narrative Analysis: Oral Versions of Personal Experience. *Essays on the Verbal and Visual Arts*, 12-44. June Helm (Ed.). University of Washington Press, Seattle, WA.

Annie Louis and Ani Nenkova. 2013. What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of ACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*.

David McClosky and Eugene Charniak 2008. Self-training for biomedical parsing. *Proceedings of the*

46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers.

Neil McIntyre and Mirella Lapata. 2009. Learning to Tell Tales: A Data-driven Approach to Story Generation. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Rada Mihalcea. 2004. Co-training and self-training for word sense disambiguation. *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*.

George Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. *Proceedings of NAACL-HLT 2013*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*.

Jessica Ouyang and Kathleen McKeown. 2014. Towards automatic detection of narrative structure. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Ellie Pavlick and Ani Nenkova. 2015. Inducing Lexical Style Properties for Paraphrase and Genre Differentiation. *Proceedings of NAACL-HLT 2015*.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12: 2825-2830.

Livia Polanyi. 1976. Why the Whats are When: Mutually Contextualizing Realms of Narratives. *Proceedings of the second Annual Meeting of the Berkeley Linguistic Society*.

Livia Polanyi. 1985. *Telling the American story : a structural and cultural analysis of conversational storytelling* Ablex Publishing, Norwood, NJ.

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

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Elahe Rahimtoroghi, Reid Swanson, Marilyn A. Walker, and Thomas Corcoran. 2013. Evaluation, Orientation, and Action in Interactive StoryTelling. *Proceedings of Intelligent Narrative Technologies 6*.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Kenji Sagae. 2010. Self-training without reranking for parser domain adaptation and its impact on semantic role labeling. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Philipp Singer, Fabian Flck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? *Proceedings of the companion publication of the 23rd international conference on World wide web companion*.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. *Computational Linguistics and Intelligent Text Processing*, 121-136. Springer Berlin Heidelberg, Berlin.

Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran and Marilyn A. Walker. 2014. Identifying Narrative Clause Types in Personal Stories. *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Cynthia Whissell. 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4:113-131. Academic Press, London.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction without labelled data. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.