Identifying Narrative Clause Types in Personal Stories

Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran and Marilyn A. Walker
Natural Language and Dialog Systems Lab
University of California Santa Cruz
Santa Cruz, CA 95064, USA
{reid,elahe,maw}@soe.ucsc.edu, tcorcora@ucsc.edu

Abstract

This paper describes work on automatically identifying categories of narrative clauses in personal stories written by ordinary people about their daily lives and experiences. We base our approach on Labov & Waletzky's theory of oral narrative which categorizes narrative clauses into subtypes, such as ORIENTATION, ACTION and EVALUATION. We describe an experiment where we annotate 50 personal narratives from weblogs and experiment with methods for achieving higher annotation reliability. We use the resulting annotated corpus to train a classifier to automatically identify narrative categories, achieving a best average F-score of .658, which rises to an F-score of .767 on the cases with the highest annotator agreement. We believe the identified narrative structure will enable new types of computational analysis of narrative discourse.

1 Introduction

Sharing personal experiences by storytelling is a fundamental aspect of human social behavior (Fivush et al., 2005; Fivush and Nelson, 2004; Habermas and Bluck, 2000; Bamberg, 2006; Thorne, 2004; Bohanek et al., 2008; Thorne and Nam, 2009; McLean and Thorne, 2003; Pratt and Fiese, 2004). Humans appear to be wired to engage with information that is narratively structured (Gerrig, 1993; Bamberg, 2006; Bruner, 1991), and telling stories provides a critical developmental and societal function, by serving as a means to reinforce community value systems and to define individual identity (Thorne and Shapiro, 2011; Thorne et al., 2007). This has led some theorists to claim that "the stories they tell" is the defining aspect of both individuals and cultures.

Unlike any prior time in human history, personal narratives about many life experiences are being told online, and are widely available in social media sources such as weblogs. A personal narrative about an arrest is shown in Fig. 1, and one about a protest is in Fig. 4. Narratives such as these provide a valuable resource for learning a wealth of commonsense knowledge about people, the types of activities they engage in, and the attitudes they hold. They are also well suited to learning about causal and temporal relationships between events because narrative interpretation explicitly depends on the coherence of these relationships (Graesser et al., 1994; Elson, 2012; Gordon et al., 2011; Hu et al., 2013).

This paper applies and tests a narrative clause labeling scheme to personal narratives. Our scheme is derived from Labov & Waletzky's (henceforth **L&W**) theory of oral narrative (Labov, 1997; Labov and Waletzky, 1967). L&W's theory distinguishes (1) clauses that indicate causal relationships (ACTION), from (2) clauses that provide traits or properties of the setting or characters (ORIENTATION), from (3) clauses describing the story characters' emotional reactions to the events (EVALUATION).

We adopt L&W's theory for three reasons. First, we believe that the narrative structure of personal narratives posted on weblogs will be more similar to oral narrative than they are to classical stories. Second, we believe that any narrative discourse typology must at least distinguish ACTION, from ORIENTATION, and EVALUATION. Third, personal stories found on the web are often noisy and difficult to interpret; they do not always clearly follow well defined narrative conventions. A deep analysis

Proceedings of the SIGDIAL 2014 Conference, pages 171–180,
Philadelphia, U.S.A., 18-20 June 2014.

#	Category	Story Clause
1	Orientation	Now, on with this week's story...
2	Orientation	The last month has been hectic.
3	Orientation	Turbo charged.
4	Orientation	Lot's of work because I was learning from Tim, my partner in crime.
5	Orientation	This hasn't been helped by the intense pressure in town due to the political transition coming to an end.
6	Orientation	This week things started alright and on schedule.
7	Action	But I managed to get myself arrested by the traffic police (rouleage) early last Wednesday.
8	Action	After yelling excessively at their outright corrupted methods
9	Action	and asking incessently for what law I actually broke,
10	Action	they managed to bring me in at the police HQ.
11	Action	I was drawing too much of a curious crowd for the authorities.
12	Action	In about half an hour at police HQ I had charmed every one around.
13	Action	I had prepared my "gift" as they wished.
14	Evaluation	Decision witheld, they decided that I neednt to bother,
15	Evaluation	they liked me too much.
16	Evaluation	I should go free.
17	Action	I even managed to meet famous Raus, the big chief.
18	Evaluation	He was too happy to let me go when he realized I was no one.
19	Action	But then, a Major at his side noticed my Visa was expired.
20	Evaluation	Damn!
21	Orientation	My current Visa is being renewed in my other passport at Immigration's.
22	Evaluation	Fuck.
23	Evaluation	In custody, for real.

Figure 1: An excerpt from an example story from our corpus annotated with the L&W categories.

and annotation scheme, such as the one employed by DramaBank (Elson and McKeown, 2010; Elson, 2012) that extends theories of narrative structure and plot units (Stein et al., 2000; Lehnert, 1981), offers many advantages. However, acquiring this level of analysis on user generated content, such as blog stories, is resource intensive.

Research on computational models of narrative structure typically focus on inferring the causal and temporal relationships between events (Goyal et al., 2010; Chambers and Jurafsky, 2009; Riaz and Girju, 2010; Beamer and Girju, 2009; Do et al., 2011; Manshadi et al., 2008; Gordon et al., 2011; Hu et al., 2013). Yet L&W point out that stories are not just about the events that occur. In fact, L&W say that stories that are only about events are boring. Current methods for inferring narrative structure, including our own (Hu et al., 2013), do not distinguish event clauses from other narrative clause types. But note that actions only constitute about one third of the clauses in the narratives in Fig. 1 and Fig. 4.

Sec. 2 provides more detail about L&W's theory. Sec. 3 describes the annotation experiments and efforts to improve annotation reliability. Sec. 4 presents experiments on learning to automatically classify L&W categories, where we examine the the most predictive features, and the effect of annotator agreement on classification accuracy. We achieve a best average F-score of .658, which rises to an F-score of .767 on the cases with the highest annotator agreement. We analyze the types of errors the classifier makes in Sec. 5.1 and conclude in Sec. 6.

2 Labov & Waletzky's Theory of Narrative Discourse

L&W's theory of oral narrative defines a story as a series of ACTION clauses (events), of which at least two must be temporally joined (e.g., clauses 7-13 in Fig. 1 and clauses 7-11 in Fig. 4) (Labov and Waletzky, 1967; Labov, 1997) Stories also contain ORIENTATIONS (setting the scene, describing the characters), e.g. utterances 1-6 in Fig. 1. An orientation clause introduces the time and place of the events of the story, and identifies the participants of the story and their initial behavior. To properly understand narrative structure, orientations need to be identified as a separate type of utterance distinct from events. L&W define two other structural types called ABSTRACT and CODA. The ABSTRACT is an initial narrative clause summarizing the entire sequence of events. A CODA is final clause which returns the narrative to the time of speaking, indicating the end of the narrative. The CODA often provides the "moral" of the story.

The final element of a story according to L&W is EVALUATION, which L&W identify as essential to every story. According to L&W, evaluation gives the reason for telling the story, or the point of the story: without EVALUATION there is no story, merely a boring recitation of events. L&W state that the EVALUATION clauses may also provide information on the consequences of the events as they

relate to the goals and desires of the participants, and can be used to describe the events that did not occur, may have occurred, or could occur in the future in the story. Clauses 14-16 and 18 in Fig. 1 provide the narrator's evaluation of the transpiring events as well as introducing possible but unrealized alternative timelines. In theories of narrative identity (McAdams, 2003; Thorne, 2004), evaluation performs two additional functions: (1) it expresses the tellers opinion and in doing so reflects the value system of that person and their community; (2) it constructs and maintains relations between the teller and the listener. Clauses 20 and 22 illustrate these functions where the narrator directly reveals his affective response to the prior events.

3 Dataset

Corpus of Personal Stories. Previous work (Gordon and Swanson, 2009) showed that about 5% of all weblog entries are personal stories describing an event in the author's daily life. They developed an automatic classifier for identifying personal narratives from a random sample of 5,000 posts from a corpus of 44 million entries available as part of the ICWSM 2010 dataset challenge (Burton et al., 2009). 229 of these posts were manually labeled as personal stories. Our experiments are based on 50 of these 229 stories.

Annotation Process. L&W's theory applies to subsentence discourse units in a narrative. It is an open question what level of phrasal granularity is appropriate to apply to written narratives. Here, we treat each independent clause as the basic unit of discourse and manually segment each story in our dataset using this definition. This results in a collection of 1,602 independent clauses. We then divided the 50 stories into 4 groups and annotated them in batches among 3 annotators in order to refine our annotation guidelines and process. This dataset is freely available at https://nlds.soe.ucsc.edu/lw.

Previous work has applied L&W's theory to Aesop's fables and achieved high levels of interannotator agreement and extremely high machine learning accuracies (Rahimtoroghi et al., 2013). However personal narratives clearly provide a more challenging context for annotation. There was a high level of disagreement after the initial round of annotation. We found at least 6 primary sources of disagreements:

- Clauses of more than one category are common with rising action and evaluation, e.g. a clause containing elements of orientation, action, and evaluation: *After leaving the apartment at 6:45 AM, flying 2 hours, taking a cab to Seattle, and then driving seven hours up to Whistler including a border crossing, it's safe to say that I felt pretty much like a dick with legs.*
- Actions that are implied but not explicitly stated in the text.
- Stative descriptions of the world as a result of an action that are not intuitively orientation.
- Stative descriptions of the world that are localized to a specific place in the narrative, which is problematic to L&W's definition of orientation.
- Subjective clauses in scene setting are usually orientation, but are lexically similar to evaluation.
- Disambiguating the functional purpose of clauses that describe actions, but may be intended to set the scene as opposed to the rising action.
- Disambiguating the functional purpose of subjective language in the description of an event or state, e.g., *The gigantic tree outside my window*, *The radiant blue of the sky*.

After several rounds of annotation we stabilized on a labeling scheme that hierarchically extends the original L&W categories, along with annotation guidelines that annotators could use to disambiguate recurring problematic cases. The final set of extended category labels along with two reduced hierarchical mappings are shown in Table 1.

STATIVE-LOCAL CONTEXT is a category for distinguishing stative descriptions of the world, that are not intuitively orientation. For example:

- I saw the sign I expected to turn south on Hwy 138. *The traffic sign pointed to Sutherlin and Roseburg,*

The clause in italics is a stative that describes the sign seen in the previous action. It is clearly not an action or evaluation, but is not intuitively an orientation, because it is so locally dependent.

STATIVE-IMPLIED ACTIONS are clauses, which do not explicitly mention an action or event, but imply one that is necessary to maintain the causal or temporal coherence of the remaining story. For example: *After that, we decided to walk some more.* In the context of the story it is necessary to know that they *actually* did walk some more in order to interpret the other actions described in the narrative. Implied actions are often passive constructions that describe a state of the world that could only be true

Label Set	κ	Labels						
Extended	0.582	¬ Story	Orient	Action	Eval	Local Context	Implied Action	Consequence
Stative	0.597	¬ Story	Orient	Action	Eval	Stative	Stative	Stative
L&W	0.630	¬ Story	Orient	Action	Eval	Orient	Action	Eval

Table 1: The extended L&W label categories and two reduced hierarchical mappings.

if an action had taken place. For example: *We were at the convention center in about 10 minutes.*

STATIVE-CONSEQUENCE is a category that describes the state of the world that has resulted as a consequence of an action, but does not directly evaluate the goals, intentions or desires of the participants. For example, clause 23 in Fig. 1.

Using this extended label set we were able to achieve an inter-annotator agreement between the 3 annotators of 0.582 using Fleiss' κ on assigning categories to clauses. We also mapped the full set of labels to a smaller subsets to see if the finer grained distinctions helped improve reliability on more coarse grained labeling schemes. The extended labels we included were generally different types of stative descriptions of the world, which were all mapped to a single category for the *Stative* label set. Finally, we mapped each extended label to an original L&W category that we thought best fit the original definitions. When mapping back to these reduced label sets we were able to increase the κ to 0.597 for the stative set and 0.630 for the original L&W categories. This result indicates that we can achieve higher reliability by ensuring that the annotators think carefully about particular kinds of distinctions between different stative clauses.

Gold standard labels were selected based on a simple majority of the annotator assignments. When no annotators agreed on a label, one of the selected labels was chosen at random. Once completed there were 424 action clauses, 702 evaluations, 26 not stories, 306 orientation, 17 stative consequences, 12 implied actions and 115 local contexts. Note that EVALUATION and ORIENTATION clauses that would not be distinguished from ACTION by previous work constitute two thirds of the clauses.

4 Experiments

The triply annotated dataset described above was used as training and test data for experiments on learning to automatically label narrative clauses. 40 narratives were randomly selected to be used as training and development data and the remaining 10 narratives for test data. The average story in the training data had 29.3 clauses with the shortest story consisting of 4 and the maximum consisting of 100. The average story in the test data had 43 clauses with the shortest story consisting of 4 and the maximum consisting of 114.

Feature Set	Description
Linguistic	Parts of Speech, Number of Characters in post, Average Word Length, Unigrams, Bigrams
Lexical and Sentiment Categories	LIWC counts and frequencies, negation
Story Position	First Clause, Last Clause, Position in the story binned into ten story regions

Table 2: Feature Sets for L&W Classification.

To derive feature representations of each type of narrative clause we started with the features presented in (Rahimtoroghi et al., 2013). We refined these by examining L&W's descriptions of distinguishing features of each category. Table 2 summarizes the features we automatically extracted from all narrative clauses in the weblogs.

First, we used the Stanford Parser to distinguish independent and dependent clauses and kept track separately of features that occurred in both types of clause. This is because L&W state that the unit of analysis should be an independent clause with its subordinate clauses, but we felt that these were exactly the cases that often caused difficulties during annotation. However distinguishing between the features occurring in the two clause types would allow us to determine if and when the features of the subordinate clause were relevant or more informative for automatic classification. Then, within both dependent and independent clauses, we distinguished the part-of-speech of the main verb (POS), whether the clause contained a negation (Negate), lexical semantic categories from LIWC (Pennebaker et al., 2001), dependency relations (DEP), lexical unigrams (STEM), and whether the verb was one of a class of verbs that are likely to be stative.

We also developed a set of features describing the relative position of the clause in the story (BinPosition, FirstClause, LastClause), because different story regions are often associated with different

Feature	Gain	Act	Ori	Eval	¬
POS:IND-VBD	0.128	**0.084**	0.002	0.031	0.011
BinPosition0	0.076	0.017	**0.042**	0.014	0.003
FirstClause	0.044	0.010	**0.019**	0.011	0.003
POS:IND-VBZ	0.042	**0.029**	0.008	0.002	0.003
IND-Negate	0.040	0.025	0.000	**0.013**	0.002
IND-Copula	0.039	**0.030**	0.004	0.005	0.001
POS:IND-:	0.036	0.001	0.000	0.002	**0.033**
IND-FirstPerson	0.035	**0.017**	0.004	0.002	0.013
IND-LIWC_Motion	0.034	**0.021**	0.003	0.006	0.004
POS:IND-VBP	0.033	**0.023**	0.001	0.007	0.002

Table 3: The 10 most highly correlated features with each label and cumulatively over all the labels using mutual information and information gain.

clause types. For example, in Fig. 1 and Fig. 4, the beginning of the story contains more ORIENTATION clauses, while ACTION clauses are concentrated in the middle of the story. The EVALUATION clauses typically occur part-way through the story where they provide the narrator's reaction to story events. See Table 2.

In total there were 3,510 unique binary valued features extracted from our training dataset. We used mutual information to find the features that had the highest correlation with each category and the information gain over all the labels. The 10 highest valued features are in Table 3, e.g. the top feature is when the part-of-speech (POS) of the main verb in the independent clause (IND) is past tense (VBD).

This feature encoding was used for machine learning experiments with classification algorithms from Mallet (McCallum, 2002): Naive Bayes (NB) (Witten and Frank, 2005), Confidence Weighted Linear Classifier (CWLC) (Dredze et al., 2008), Maximum Entropy (ME)(Witten and Frank, 2005) and a sequential classifier (CRF) (Lafferty et al., 2001).

5 Evaluation and Results

We evaluate the performance of our classifiers with experiments using the 50 annotated stories. Using the 40 stories in the training set we calculated the information gain for each feature (see Table 3). For each subset of the highest valued features (in the range of 2^2-2^{12}), we performed a 10-fold cross-validation on the training data and assessed the performance of each classifier to find the right number of features. Within each fold of the cross-validation we also perform a simple grid search for the optimal hyper-parameters of the model (e.g., the prior in the ME and CRF models) using only the data within the training fold.

The feature selection experimental results are shown in Table 4. We report the optimal number of features and the corresponding macro F-score, weighted by the relative frequency of each category, for each algorithm and label set. For all algorithms, performance increases for label sets with higher levels of abstraction. The Naive Bayes and CRF models perform better with a small subset of the features, while the ME and CWLC algorithms use a much larger subset. Surprisingly the sequential classifier has the lowest F-score and Naive Bayes performs the best. A * indicates a significant improvement over CRF at $p < 0.05$ using a two-sided t-test. No other differences were significant.

	Extended		Stative		L&W	
Classifier	#	F1	#	F1	#	F1
CRF	2^7	0.61	2^9	0.61	2^7	0.65
CWLC	2^{11}	0.67	2^{11}	0.68*	2^{11}	0.73*
ME	2^{11}	0.67	2^{10}	0.68*	2^{10}	0.73*
NB	2^9	0.68*	2^9	0.70*	2^{10}	0.76*

Table 4: The optimal number of features found for each model and the average F-score obtained using a 10-fold cross-validation on the training data.

Using the optimal number of features obtained from this search we trained a model for each algorithm using the entire training dataset and selecting the hyper-parameters as before. We applied these models to the unseen test data and evaluated the performance of each classifier as applied to the entire set of clauses and to individual narratives.

We first computed the precision, recall and F-score aggregated over all the clauses in the test set. Table 5 summarizes the results for each classifier and label set. The left hand side of the table shows the macro precision, recall and F-score weighted by the relative frequency of each label. The right hand side of the table shows the F-score of each individual label separately. On this evaluation, Naive Bayes outperforms all other methods on all label sets. Overall, precision and recall are relatively balanced achieving a maximum F-score of 0.689 when the labels are mapped back to the original L&W categories. The CRF does surprisingly well considering its poor performance during the feature selection search. The classifiers perform the poorest on orientation clauses and the best on evaluation clauses.

As mentioned above, the annotation task is highly subjective, requiring interpreting the narrative and the author's intention, which prevents us from obtaining high levels of inter-rater agreement. Because of the noise in the annotations, the standard evalua-

		Overall				Per Label						
						L&W				Stative		
Label Set	Model	Prec	Rec	F1	κ	Ori	Eva	Act	¬	Imp	Loc	Con
Extended	CRF	0.568	0.626	0.593	0.419	**0.532**	0.727	0.651	0.000	0.000	0.041	0.000
	CWLC	0.567	0.616	0.582	0.398	0.377	0.763	0.612	0.000	0.000	0.087	0.000
	ME	0.597	0.649	0.614	0.450	0.496	0.767	**0.667**	0.000	0.000	0.085	0.000
	NB	**0.625**	**0.656**	**0.623**	**0.459**	0.478	**0.781**	0.650	0.000	0.000	**0.174**	0.000
Stative	CRF	0.563	0.591	0.574	0.370	0.412	0.695	0.628	0.000		**0.235**	
	CWLC	0.572	0.621	0.587	0.403	0.417	0.760	0.614	0.000		0.077	
	ME	0.610	0.644	0.611	0.441	0.464	0.759	0.673	0.000		0.118	
	NB	**0.650**	**0.667**	**0.638**	**0.477**	**0.496**	**0.779**	**0.676**	0.000		0.226	
L&W	CRF	0.650	0.665	0.656	0.468	0.556	0.742	0.640	0.000			
	CWLC	0.624	0.647	0.632	0.424	0.480	0.747	0.609	0.000			
	ME	0.681	0.700	0.688	**0.517**	**0.580**	**0.780**	0.670	0.000			
	NB	**0.687**	**0.705**	**0.689**	0.514	0.565	**0.780**	**0.687**	0.000			

Table 5: The performance of each of classifier on the test set when all clauses are aggregated together.

Agreement	Total #	Prec	Rec	F1	κ	Ori	Eva	Act	¬
1 of 3	15	0.333	0.400	0.339	0.069	0.000	0.625	0.333	0.000
2 of 3	146	0.597	0.610	0.580	0.405	0.472	0.700	0.622	0.000
3 of 3	269	0.770	0.773	**0.767**	0.607	0.667	0.824	0.746	0.000
All	430	0.687	0.705	0.689	0.514	0.565	0.780	0.687	0.000
Adjusted	430	0.646	0.660	0.643	0.447	0.516	0.745	0.623	0.000

Table 6: Performance measures for different levels of agreement among the annotators.

tion metrics may hide information about the ability of the classifiers to learn from our feature set. For example, the best performing classifier (NB) incorrectly labeled 127 clauses out of 430 possible in the test set. However, 44 of these errors agreed with at least one annotator, but were counted as entirely incorrect in the previous evaluations.

To address these concerns we also evaluated the performance of the the best performing classifier based on the level of agreement of each instance using two different approaches. See Table 6. The first approach was inspired by the approach in (Louis and Nenkova, 2011) where the clauses in the test set are binned based on the number of annotators who agreed with the gold standard label. The performance is then calculated for each bin. The first three rows of Table 6 show the performance for the different levels of agreement in the dataset. There were only 15 clauses in the test set where there was no agreement at all. It is unsurprising that when the annotators could not agree on a label the system performed near chance levels. However, **when all three annotators agreed on the gold standard label the F-score improved to 0.767**. As a comparison, the F-score of the entire test set was 0.689 as shown in the row labeled *All*.

Our second approach is based on the proposal of Tetreault et al. (Tetreault et al., 2013). They introduce a modification to the standard precision, recall and F-scores that takes into account the level of agreement of each instance, where the values of true-positives and false-negatives are assigned fractional counts based on the proportion of annotators who assigned that label. The final row of Table 6 provides the results using these adjusted values.

Label Set	Model	Min	Max	Mean $\pm$ CI
Extended	CRF	**0.333**	**0.763**	0.540 $\pm$ 0.080
	CWLC	0.276	**0.763**	**0.582** $\pm$ 0.099
	ME	**0.333**	0.753	0.572 $\pm$ 0.088
	NB	**0.333**	0.741	0.573 $\pm$ 0.093
Stative	CRF	0.298	**0.762**	0.521 $\pm$ 0.099
	CWLC	**0.345**	0.758	**0.591** $\pm$ 0.090
	ME	0.333	0.753	0.562 $\pm$ 0.098
	NB	0.333	0.758	0.582 $\pm$ 0.088
L&W	CRF	0.333	0.837	0.609 $\pm$ 0.097
	CWLC	**0.458**	**0.877**	**0.658** $\pm$ 0.081
	ME	0.333	0.830	0.649 $\pm$ 0.095
	NB	0.333	0.851	0.647 $\pm$ 0.096

Table 7: Summary statistics of the F-score, with 95% confidence intervals, when evaluating stories.

We also investigated the performance of the classifiers when evaluating each story separately. Table 7 summarizes these results. Each classifier was applied to the clauses of the 10 narratives in the test set and the F-score was computed for each narrative individually. The table shows the minimum, maximum and average F-score with 95% confidence in-

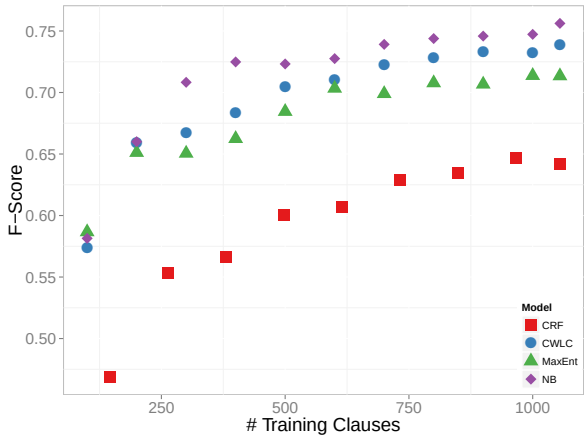

Figure 2: Learning curves of the Naive Bayes classifier using the optimal number of features.

tervals over the 10 narratives.

The CWLC performed the best on this test and the performance of all the algorithms generally improved using the higher-level label sets. The results also show that there is a high variance in performance between stories, with a minimum F-score of 0.458 and a maximum of 0.877 for the CWLC on the L&W label set. This indicates that some clauses are ambiguous and difficult to label, but also that some *stories* are more difficult to classify.

To assess whether more annotated data could improve performance, we ran a series of learning curves in Fig. 2. Only the training data was used for these experiments. The curves were created by randomly sampling 90% of the data for training and 10% for testing. A model was trained, using the same process as above, on successively larger subsets of the data and applied to the 10% held out clauses. This process was repeated 10 times and the mean F-Score is reported. In nearly all cases, the performance of classifiers is still increasing when all of the data is used indicating that we have not exhausted the expressive power of our features and more annotated data would be useful. However, we also see we can reach about 93% of our maximum performance with only a few hundred examples. We plan to increase the size of our annotated data set in future work.

5.1 Error Analysis

Our results to date indicate that we achieve an overall F-score of 0.689, and that our classifiers are most accurate for the *evaluation* and *action* categories. See Table 6. Fig. 3 presents a confusion matrix

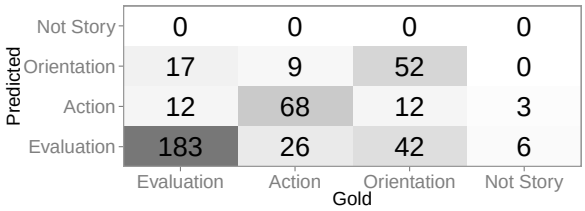

Figure 3: Confusion matrix for the best classifier.

showing the frequency of predicted labels against the gold standard labels for the Naive Bayes classifier on the L&W label set. With the exception of *not story* there are cases of confusion between all categories. However, in the vast majority of cases both action and orientation are confused with evaluation and the classifier overpredicts evaluation.

We also categorized errors for the Naive Bayes classifier into the the 4 sources of errors in the predictions shown in Table 8. The most common errors involved clauses with lexical INDICATORS that are highly correlated with one category, but in the context and interpretation of the story actually function as a different type. For example, **unfortunately**, **could** and **n't** are all strong indicators of evaluation, but in this case the primary function of the clause is to set the scene for the rest of the story, i.e., orientation. The interpretation of these clauses is clear to a human, despite the lexical items misleading the classifier.

Another source of error is when the function of the clause in the narrative is ambiguous (PURPOSE in Table 8). While there may be some misleading lexical indicators in these clauses, there were often no strongly correlated words, such as the adjectives and modal verbs in EVALUATIONS. The distinction in these cases is that the primary function of the clause within the story is unclear, even to a human reader. Unsurprisingly, most of the examples in this category had low inter-rater agreement.

Some of the clauses contain figurative language or complex constructions that require a significant amount of world knowledge and INFERENCE to interpret. For example, understanding the INFERENCE clause in Table 8 requires recognizing the metaphor about rabbit food in order to identify the subjective evaluation the narrator is making.

There are also cases of clauses that contain MULTIPLE categories, at least partially because of the granularity of our segmentation. In the example in Table 8 a new character, Alejandrio, is introduced and a rising action is described, trekking to the waterfall.

Error Type	Freq	Gold	Pred	Example
Indicators	57	Ori	Eva	So, **unfortunately** I **couldn't** make the Gamesindustry.biz party tonight.
Purpose	20	Ori	Eva	I know it is a remarkable haircut because on the way home a handsome young Moroccan man nearly died to tell me how beautiful I was.
Inference	6	Eva	Ori	That's that rabbit food that all of those Northeastern Kerry voters...
Multiple	4	Act	Ori	We trekked to a waterfall in the park with the help of Alejandrio a 65 year old Honduran guy who not only walked quicker than us but also carried all the water.
Unclear	39	Ori	Eva	We have diners out east,
Not Story	7	Not	Act	scroll down to the Hobbit post,

Table 8: Several common sources of errors with a prototypical example.

Our annotation guidelines instructed us to prefer actions in these types of clauses, however, both ORIENTATION and ACTION are present in this situation.

There were also 39 clauses that were labeled incorrectly that had no clear reason (UNCLEAR) for being mislabelled. We also explicitly excluded the 7 clauses marked not part of the story.

The types of errors described above are not mutually exclusive and in some cases are causally related. For example, the purpose of a clause may be ambiguous because it contains conflicting lexical indicators. Similarly, a clause containing multiple categories will likely have strong lexical indicators from each of these categories so that the classification algorithms cannot disambiguate among possible labels. This might be improved by more data, more sophisticated semantic features, or possibly an analysis focused on discourse relations, such as those in the PDTB (Louis et al., 2010; Prasad et al., 2008), or Elson's STORY INTENTION GRAPH (Rishes et al., 2013; Elson and McKeown, 2010; Elson, 2012).

6 Discussion

This paper describes work on categorization of narrative clauses based on Labov & Waletzky's theory of oral narrative, applied to personal narratives written by ordinary people. We show that we can automatically classify narrative clauses with these categories achieving an overall F-score of 0.689, which is substantially higher than a random (0.250) or majority class (0.437) baseline, which increases to an F-score of .767 on the cases where all three annotators agreed. The learning curves plotted in Fig. 2 clearly suggest that more training data would be beneficial before we investigate more complex features and learning algorithms.

We believe the ability to automatically perform this type of simple narrative analysis will enable and improve many other types of deeper narrative understanding (Rahimtoroghi et al., 2014; Hu et al., 2013). For example, causal and temporal relationship extraction methods that focus only on clauses in the same *action* sequence be more accurate, because they exclude disconnected events from the orientation or evaluation sections. This type of analysis will also enable new methods for learning attitudes and values of societal groups based on the different affective evaluations that are provided in response to action clauses.

Our results also highlight several properties of the data. Performance is different for results by story rather than over all clauses. This indicates that some stories are more difficult to classify than others and that ambiguous clauses are not uniformly distributed but are likely to be correlated with particular authors or writing styles. In other work, we have started to investigate whether we can automatically rate the temporal coherence of personal narratives (Ryan et al., 2014). We can use this to identify stories with utterances that are likely to be difficult to classify because of the poor quality of the narrative input. These cases are unlikely to have usable narrative structure.

Acknowledgments

This research was supported by NSF Grants IIS #1002921 and IIS #123855. The content of this publication does not necessarily reflect the position or policy of the government, and no official endorsement should be inferred.

Appendix A

See Fig. 4 for an additional example labelled with L&W Categories.

#	Category	Story Clause
1	Abstract	Today was a very eventful work day.
2	Orientation	Today was the start of the G20 summit.
3	Orientation	It happens every year
4	Orientation	and it is where 20 of the leaders of the world come together to talk about how to run their governments effectively and what not.
5	Orientation	Since there are so many leaders coming together their are going to be a lot of people who have different views on how to run the government they follow so they protest.
6	Orientation	This week things started alright and on schedule.
7	Action	There was a protest that happened along the street where I work
8	Action	and at first it looked peaceful until a bunch of people started rebelling
9	Action	and creating a riot.
10	Action	Police cars were burned
11	Action	and things were thrown at cops.
12	Orientation	Police were in full riot gear to alleviate the violence.
13	Action	As things got worse tear gas and bean bag bullets were fired at the rioters
14	Action	while they smash windows of stores.
15	Evaluation	And this all happened right in front of my store
16	Evaluation	which was kind of scary
17	Evaluation	but it was kind of interesting
18	Coda	since I've never seen a riot before.

Figure 4: A personal narrative about a protest, with narrative categories of Labov & Waletzky, 1967.

References

Michael Bamberg. 2006. Stories: Big or small: Why do we care? *Narrative inquiry*, 16(1):139–147.

Brandon Beamer and Roxana Girju. 2009. Using a bi-gram event model to predict causal potential. In *Computational Linguistics and Intelligent Text Processing*, p. 430–441. Springer.

Jennifer G Bohanek, Kelly A Marin, and Robyn Fivush. 2008. Family narratives, self, and gender in early adolescence. *The Journal of Early Adolescence*, 28(1):153–176.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18:1–21.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 spinn3r dataset. In *Proc. of the Third Annual Conf. on Weblogs and Social Media*.

N. Chambers and D. Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proc. of the 47th Annual Meeting of the ACL*, p. 602–610.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, p. 294–303.

Mark Dredze, Koby Crammer, and Fernando Pereira. 2008. Confidence-weighted linear classification. In *Proc. of the 25th international conference on Machine learning*, p. 264–271. ACM.

D.K. Elson and K.R. McKeown. 2010. Building a bank of semantically encoded narratives. In *Proc. of the Seventh International Conf. on Language Resources and Evaluation (LREC 2010)*.

David K Elson. 2012. Detecting story analogies from annotations of time, action and agency. In *Proc. of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*.

Robyn Fivush and Katherine Nelson. 2004. Culture and language in the emergence of autobiographical memory. *Psychological Science*, 15(9):573–577.

Robyn Fivush, Jennifer G Bohanek, and Marshall Duke. 2005. The intergenerational self: Subjective perspective and family history. *Individual and collective self-continuity. Mahwah, NJ: Erlbaum*.

R.J. Gerrig. 1993. *Experiencing narrative worlds: On the psychological activities of reading*.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conf. on Weblogs and Social Media, Data Challenge Workshop*.

Andrew Gordon, Cosmin Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth Conf. on Artificial Intelligence (AAAI-11)*.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proc. of the 2010 Conf. on Empirical Methods in Natural Language Processing*, p. 77–86.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

T. Habermas and S. Bluck. 2000. Getting a life: the emergence of the life story in adolescence. *Psychol Bull*, 126(5):748–69.

Zhichao Hu, Elahe Rahimtoroghi, Larissa Munishkina, Reid Swanson, and Marilyn A Walker. 2013. Unsupervised induction of contingent event pairs from film scenes. In *Proc. of Conf. on Empirical Methods in Natural Language Processing*, p. 370–379.

W. Labov and J. Waletzky. 1967. Narrative analysis: Oral versions of personal experience. In J. Helm, ed., *Essays on the Verbal and Visual Arts*, p. 12–44.

W. Labov. 1997. Some further steps in narrative analysis. *Journal of narrative and life history*, 7:395–415.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

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

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *International Joint Conf. on Natural Language Processing*, p. 605–613.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit relations. In *Proc. of the 11th Annual SIGdial Meeting on Discourse and Dialogue*.

Mehdi Manshadi, Reid Swanson, and Andrew S Gordon. 2008. Learning a probabilistic model of event sequences from internet weblog stories. In *Proc. of the 21st FLAIRS Conf.* .

Dan P McAdams. 2003. Identity and the life story. *Autobiographical memory and the construction of a narrative self: Developmental and cultural perspectives*, 9:187–207.

Andrew Kachites McCallum. 2002. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu.

K.C. McLean and A. Thorne. 2003. Adolescents' self-defining memories about relationships. *Developmental Psychology*, (39):635–645.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Inquiry and Word Count: LIWC 2001*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proc. of 6th International Conf. on Language Resources and Evaluation (LREC 2008)*.

Michael W Pratt and Barbara H Fiese. 2004. *Families, Stories, and the Life Course: An Ecological Context.*

Elahe Rahimtoroghi, Reid Swanson, and Marilyn A. Walker. 2013. Evaluation, orientation, and action in interactive storytelling. In *Proc. of Intelligent Narrative Technologies 6*.

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, Marilyn A. Walker, Kenji Sagae, and Andrew S. Gordon. 2014. Minimal narrative annotation schemes and their applications. In *Proc. of Intelligent Narrative Technologies 7*.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC)*, p. 361–368. IEEE.

Elena Rishes, Stephanie Lukin, David K. Elson, and Marilyn A. Walker. 2013. Generating dierent story tellings from semantic representations of narrative. In *Int. Conf. on Interactive Digital Storytelling, ICIDS'13*.

James Owen Ryan, Marilyn A. Walker, and Noah Wardrip-Fruin. 2014. Toward recombinant dialogue in interactive narrative. In *7th Workshop on Intelligent Narrative Technologies*.

Nancy L. Stein, Tom Trabasso, and Maria D. Liwag. 2000. A goal appraisal theory of emotional understanding: Implications for development and learning. In M. Lewis and J. M. Haviland-Jones, ed, *Handbook of emotions (2nd ed.)*, p. 436–457.

Joel Tetreault, Martin Chodorow, and Nitin Madnani. 2013. Bucking the trend: improved evaluation and annotation practices for esl error detection systems. *Language Resources and Evaluation*, p. 1–27.

A. Thorne and V. Nam. 2009. The storied construction of personality. In Kitayama S. and Cohen D., ed, *Handbook of Cultural Psychology*, p. 491–505.

A. Thorne and L. A. Shapiro. 2011. Testing, testing: Everyday storytelling and the construction of adolescent identity. *Adolescent Vulnerabilities and Opportunities: Developmental and Constructivist Perspectives*, 38:117.

A. Thorne, N. Korobov, and E. M. Morgan. 2007. Channeling identity: A study of storytelling in conversations between introverted and extraverted friends. *Journal of research in personality*, 41(5):1008–1031.

Avril Thorne. 2004. Putting the person into social identity. *Human Development*, 47(6):361–365.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA.