Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types

Belen Saldias
MIT Media Lab
Cambridge MA 02139, USA
belen@mit.edu ✉

Deb Roy
MIT Media Lab
Cambridge MA 02139, USA
dkroy@media.mit.edu

Abstract

Sharing personal narratives is a fundamental aspect of human social behavior as it helps share our life experiences. We can tell stories and rely on our background to understand their context, similarities, and differences. A substantial effort has been made towards developing storytelling machines or inferring characters' features. However, we don't usually find models that compare narratives. This task is remarkably challenging for machines since they, as sometimes we do, lack an understanding of what similarity means. To address this challenge, we first introduce a corpus of real-world spoken personal narratives comprising 10,296 narrative clauses from 594 video transcripts. Second, we ask non-narrative experts to annotate those clauses under Labov's sociolinguistic model of personal narratives (i.e., action, orientation, and evaluation clause types) and train a classifier that reaches 84.7% F-score for the highest-agreed clauses. Finally, we match stories and explore whether people implicitly rely on Labov's framework to compare narratives. We show that `actions` followed by the narrator's `evaluation` of these are the aspects non-experts consider the most. Our approach is intended to help inform machine learning methods aimed at studying or representing personal narratives.

1 Introduction

We can develop the ability to retrieve a story that we have experienced or heard when someone else is telling a story. We find ourselves thinking about our story, and so we think that we know what is coming next in our friend's story. However, in order for computers to match stories automatically, we need to understand what "matching" implies and what aspect of a story should be attended to.

There have been some attempts to match stories (Nguyen et al., 2014; Chaturvedi et al., 2018)

and to understand human judgment about matched stories (Nguyen et al., 2014; Reagan et al., 2016). Nevertheless, these efforts have been mostly developed in supervised scenarios that already have a set of matched stories in hand, and they are mostly focused on non-personal narratives (e.g., fictional). From these insightful works, however, we want to explore the understanding that when we consider stories to be similar, we attend to some aspects more than others, stressing the need for comparison of different aspects rather than at a global level.

As a first effort towards our purpose, we collect the largest annotated corpus of spoken personal narratives to our knowledge, comprising 10,296 narrative clauses from 594 stories. We use transcripts of Roadtrip Nation (RTN) videos[1], where professionals share stories about their lives and career pathways. As for the annotation task, we ask Mechanical Turkers to annotate each clause under Labov's sociolinguistic model of personal narratives (Labov et al., 1967), where a narrative is defined by a structural component, which includes a temporal organization (action clauses) and contextual orientation (orientation clauses), and an evaluation component (evaluation clauses), which represents storytellers'/characters' needs and desires (explained in more depth in section 3).

Next, aiming to automatically tag stories, we develop a model to classify these clauses that reaches 84.7% F-score for the highest-agreed clauses. Once we can automatically differentiate among clause types, we would like to use them to compare stories, but, do ordinary people rely on these clause types to compare narratives? To approach that question, we pair stories and run experiments to understand to what extent ordinary people (as opposed to literary experts) rely on Labov's model to think about similarities among these stories.

---

[1] https://roadtripnation.com/

Our approach is intended to help inform machine learning methods aimed at studying personal narratives and at modeling abstract information extraction. To the best of our knowledge, this work is the first to propose and develop an approach to understand whether ordinary people rely on Labov's framework to compare personal narratives and what they perceive as similarities among those narratives. We show that actions followed by the narrator's evaluation of these are the aspects non-experts consider the most when they compare stories. Our main contributions can be summarized as follows:

- We acquire annotations to comprehensively label real-world spoken personal narratives, amounting to 10,296 clauses under Labov's clause types, and develop a straightforward strategy to classify those clauses.

- We explore to what extent people rely on Labov's framework to compare stories and show that people tend to recognize better similarities in action and evaluation clauses.

The rest of the paper is organized as follows. In section 2, we present some main related work. In section 3, we specify the story aspects to be used in our experiments. In section 4, we describe the uniqueness of our introduced narrative corpus. In section 5 and 6 we describe results, and we end with conclusion and future directions in section 7.

2 Related Work

Our work is preceded by substantial efforts toward document (Blei et al., 2003; Dai et al., 2015; Yang et al., 2016) and story (Mostafazadeh et al., 2016; Chaturvedi et al., 2018; Iyyer et al., 2016; Antoniak et al., 2019; Fu et al., 2019) representation. We find that most approaches to text similarity focus on non-narrative corpora (vor der Brück and Pouly, 2019; Lin et al., 2013; Cer et al., 2017). We also observe that most works in stories have been developed for non-personal narratives.

An specific approach to story matching was proposed by Chaturvedi et al. (2018), who used movie remakes from Wikipedia as paired stories and showed that even in that scenario it was challenging to match the remakes. Additionally, their method does not generalize well to other story types (or even movie plots) since they include specific movie parameters, like characters' name and gender, as the basis of their solution, which does not apply to our case since we do not attempt to match stories based on these surface-level indicators. The closest work to ours was done by Nguyen et al. (2014), who proposed a set of crowdsourcing tasks to analyze perception of similarity in folk narratives. They tried various approaches to retrieve these narratives. Nevertheless, they had in hand a set of metadata labels that allowed them to match narratives prior to any experiment.

How we narrate our stories was initially studied by Labov et al. (1967). More recently, Swanson et al. (2014) proposed the first mechanism to automatically classify Labov's clauses (action, orientation, and evaluation-type clauses) in personal narratives based on clauses' syntactical structure, namely part-of-speech (POS). By using 50 short stories from online mini-blogs, of diverse topics and structures, they developed a well-defined set of instructions to properly annotate Labov's clause types (referred to as *baseline* method and dataset onward). However, personal narratives from spoken stories set a more challenging context for both annotation and collection (see section 4). We get inspiration from these works to approach clause types classification using newer techniques like word embeddings (Pennington et al., 2014) and neural networks (Kim, 2014).

Furthermore, as we learn to disentangle narrative dimensions or aspects (namely, action, orientation, and evaluation-type clauses), we can use them for other story representation tasks. For instance, identifying the clauses within a story that tell people's intents/desires, reactions, and evaluation of the events (e.g., emotions) can help train and evaluate models aimed at detecting, or planning plots conditioned on, those underlying intentions and reactions (Rashkin et al., 2018; Guan et al., 2019).

3 Story aspects of comparison

Stories can be thought to be similar in a variety of dimensions; unlike most non-narrative texts, stories have "meta" dimensions that go beyond what is said (context of a story, actions that happen, emotional content, speaker's backgrounds, among others). In this work, we explore to what extent Labov's model for personal narratives underlies how non-expert people perceive story similarities. We focus on the following three aspects:

Temporal organization (action clauses): These clauses express a series of events. The narrator might play with the story's chronology, causing

differences between narrative structures of one narrator and another.

Contextual world (orientation clauses): These clauses describe information about the context in which actions occur; they serve to orient the audience about people, places, time, and behavioral situations.

Human needs and desires (evaluation clauses): These clauses give significance and tell about the purpose of telling that story; they express the narrator's needs and desires.

See figure 1 for an example of a narrative annotated under Labov's model for personal narratives.

4 Narrative corpus

We introduce the largest dataset of annotated spoken personal narratives to our knowledge, from now on referenced as Roadtrip Nation or RTN corpus. These narratives were obtained from transcripts of stories video-recorded by Roadtrip Nation (RTN). In those videos, people from many backgrounds share stories about their lives and career pathways. The corpus comprises 10,296 narrative clauses from 594 stories (each one told by a different person), which account for more than 10 hours of people telling stories, each one averaging 17.1 clauses or 62 seconds long, where each clause has on average 11 tokens.

To split narratives into clauses, we proceed as follows. For every sentence in the story, we take every independent clause along with its dependent clause, which account for one narrative clause. To determine clauses, we rely on top-level S* (S, SINV, SBAR, SBARQ, SQ) tags from Penn Treebank II (Bies et al., 1995). For each top-level S* tag, we take its subtree along with hanging prepositions, conjunctions, and adverbs.

While we propose to automatically split our data, Swanson et al. (2014)'s data (our *baseline* dataset) was split by trained humans. We compared our strategy implemented using NLTK with their strategy by spliting their stories as well; we found that our method differs at most in one clause from their manually split stories.

4.1 Uniqueness of this narrative corpus

This corpus is particularly well-suited to study oral personal narratives for a few reasons. First of all, these stories were all video-recorded and manually transcribed (by Roadtrip Nation (RTN)[2]). These

Story clause	Clause type
I had no photography experience	Orientation
, so I started from scratch.	Action
I was an older student and didn't have any bad habits.	Orientation
So I became obsessed with that and became a double major with a specialization in photojournalism, minor in art.	Action
And, uh, worked very hard at that then.	Action
If I could wish upon you one thing, it would be that whatever you find to earn a living, you find something that you love to do.	Evaluation

Figure 1: A fragment of a personal narrative in the RTN corpus annotated by Turkers using Labov's model.

stories are raised from spontaneous questions during real-world interviews to adults conducted by high school or college students, which produces a fluid and constantly changing dynamic.

Additionally, we recognize the storytellers' awareness of the listeners due to the presence of oral discourse markers that are prominent in oral narratives, such as "you know," "right," "anyway," "like," "ah," "uh," among others. Particularly, "you know" is the most frequent bigram in our dataset (0.5% of all bigrams, 437 appearances) compared to the baseline dataset to study Labov's model, which has "you know" mentioned only 3 times throughout all stories (Swanson et al., 2014). Furthermore, we find that the word "you" appears in the RTN stories an average of 5.3 times per story vs. 1.4 times in the baseline stories.

Besides giving a background (orientation clauses) and telling events (action clauses), RTN stories are specifically produced to display meaningful life experiences or pathway decisions to make the listener reflect or engage with the stories. These purposes emphasize Labov's evaluative function (evaluation clauses) of describing the storyteller's motivation in telling their story.

Here are two randomly sampled full transcripts (i.e., RTN stories), where we can see some of the spectrum of the stories in this corpus:

1. *"In college, I was figuring my life out. I didn't have an exact plan in terms of what I wanted to do. Everybody that acted in the capacity of a guidance counselor to me helped mold me into where I am today. For instance, when I was in high school, my guidance counselor told me , Chris, based on what I know about you, I know you love to be in big cities. I know you love to study human behavior and psychology. We discussed where I might end up in college , so I chose to go to NYU based on that feedback. And when I got my first job in marketing analytics, it's when I realized that hey, this is really cool, I actually really like this. Don't feel like you have to know all the answers right now. The more*

strict you are in terms of what you think you want to do, the less options you'll have. So think outside the box and keep an open mind."

2. *"And slowly and slowly, I started doing small jobs, you know, like, you know, I think one of my first jobs was doing, like, you know, ironing Peter Gabriel's suit and giving him powder for Good Morning America. Like, you know, kind of little things like that. But already, working with musicians, I was like, 'This is is where I belong.' So, a magical thing happened at this time. I got introduced to Lenny Kravitz, and Lenny Kravitz, at the time, ah, was, uh, a poor, starving musician. Eventually, after working with Lenny for a long time, my work started to grow, and I was working with more and more people and doing other things. So, I realized that ... the next step for me... would be to work on a movie."*

Note that in written stories (such us the ones in the baseline dataset), all the oral discourse markers present in this last story can be proofread and extracted. However, these are inherent to spontaneous oral narratives.

Finally, even though we ran our experiments prompting Turkers to "focus on the content and not the speakers' characteristics such as accent or gender" (first note in full instructions), the released dataset[3] includes speakers' gender to encourage further analysis across people with different backgrounds but similar stories. From results in section 6.2, we estimate that Turkers were rarely biased in their assessment of similarity towards gender because when they were asked to explain why two stories were similar, not one reason related to gender (out of 180 explained reasons).

4.2 Annotation process

We followed the annotation guidelines, for Labov's model extended label set, constructed by trained researchers in Swanson et al. (2014) to explain to Mechanical Turkers how to annotate our clauses. Since both domains of stories (RTN vs. baseline data from Swanson et al. (2014)'s) are different, we ran earlier small quality-control experiments to understand whether workers could reach an agreement and, if so, generate labeled stories to add as examples to the task description. Turkers were also invited to provide feedback during these early experiments; after two iterations, we converged to a

detailed task description. Finally, each story was assigned to three different workers, and an average of 2.23 workers agreed on every clause.

Aiming for clean annotations, along with injecting gold examples to reduce randomness, workers were rewarded $1.35 per story, were restricted to living in a English-speaking country, had a HIT Approval Rate $\geq$ 99 and Number of HITs Approved $\geq$ 500, and had been granted Masters status on the platform. We made annotation tasks full description, some audible stories, and collected data for this task available at https://github.com/social-machines/acl-nuse-personal-narratives. Gold labels were assigned by simple majority, and for those clauses without agreement, we randomly selected one of the assigned labels by annotators. Find the label distribution in table 1. Overall, we have 9,234 clauses with at least 2 Turkers agreed on them, and 3,495 clauses with 3 Turkers agreed on them.

5 Narrative clauses classification

Learning to classify narrative clauses can help us disentangle personal narratives' dimensions. Our specific intention is to understand how this decomposition helps compare stories in different aspects (clause types are assumed to be aspects or dimensions within stories for this work). Additionally, each of these clause types can be used independently for various objectives. For instance, action-type clauses could guide events extraction where, even though the narrator might play with the story's chronology, having these clauses apart can help find causal or temporal orders. Also, identifying orientation-type clause can help create a grounded understanding of the story, where actions and emotions depend on the story's environment described in these clauses. Finally, evaluation-type clauses could bring to surface narrators' mental states,

Clause type	RTN	*baseline*
Action	26.7% (2.15)	24.2%
Evaluation	40.0% (2.29)	50.0%
Orientation	29.7% (2.24)	24.2%
Not story	3.6% (2.13)	1.6%
Total clauses	10,296	1,602

Table 1: Label distribution. Find between parentheses the average agreement for each clause type. Note that the evaluation clause type is the most common clause type in both datasets.

which could push forward research on language models conditioned on mental states (Rashkin et al., 2018).

We propose to use a convolutional neural network (CNN) with max-over-time pooling to classify clauses (Zhang and Wallace, 2015). More specifically, our model consists of a non-static CNN as in Kim (2014), where we initialize embeddings using $d = 300$-dimensional GloVe pretrained vectors (Pennington et al., 2014) and concatenate to each vector a one-hot vector (45-dimensional) that encodes POS tags associated with every token. We perform 1-max pooling with ReLU activations over each map generated by filters of sizes 2, 3, and 4; we use 30 filters per size. Then, we use two linear ((90, 45), (45, 3)) layers with dropout of 0.3 before the final softmax layer. We also explored fine-tuning BERT (Devlin et al., 2018) and found that, in most tried scenarios, this simple word-based CNN-based model outperformed BERT in accuracy, maybe due to the small fine-tuning dataset.

We randomly split the RTN dataset into 86% for training, 7% for validation, and 7% for testing, removing the "not story"-clause type. This gives 7,698 training, 619 test, and 634 validation clauses with agreement ≥ 2. Our vocabulary has around 6,000 tokens, including an unknown word token we use for uncommon words (≤ 2 appearances).

For training, we used 60 epochs and early stopping based on the validation error. We trained with different number of filter, linear layer sizes, batch sizes and learning rates set through experimentation based on performance. We find our best results using Adam with a learning rate of 5e-5 and use batch sizes of 64.

5.1 Baseline

We compare our best architecture to the baseline approach proposed by Swanson et al. (2014). To reproduce this baseline, we follow the authors' feature engineering approach and use their data split. By running experiments, we observed correspondence with the top 5 feature-relevance ranking that the baseline model found (POS:IND-VBD being the top 1). This informed our decision of using POS in our proposed approach as well. Note that, originally, the baseline model also included relative clause position within a story (which we are not including here since we mostly care about the clause purpose given its language), lexical seman-

tic categories from LIWC (Pennebaker et al., 2001), dependency relations (DEP), and lexical unigrams (STEM). Using all these engineered features, Swanson et al. (2014) reached an F-score of 76.7% on the cases with the highest annotator agreement. We refrained from using all but part-of-speech (POS) engineered features and still achieved 72.7% F-score by replicating their approach (see table 2).

5.2 Results

We report results for models trained and tested with (disjoint) sets composed only of clauses where at least two annotators agreed on their corresponding clause types, and as described in section 4.2, gold truth labels were assigned by simple majority.

Results are shown in table 2. Our results demonstrate that a simple CNN with pre-trained embedding and no feature engineering reaches high performance in our RTN dataset. Furthermore, we can see that our proposed model (trained on RTN data) still achieved high performance while evaluated on the baseline test set, even though these datasets are from different domains. On the other hand, the baseline support vector machine (SVM, linear and l1-penalized) (Cortes and Vapnik, 1995) model performs poorly when evaluated in RTN data, likely because it only uses POS (syntactic) features to represent clauses, and both written (baseline) and spoken (RTN) clauses pose different challenges in syntactical structure. We address these challenges by taking advantage of word embeddings' representational power. *From this, we see that our approach (model and dataset) can be generalized to the baseline dataset better than the other way around.*

Model	RTN test	*baseline* test
CNN - RTN	84.7%	*62.9%
SVM - *baseline*	*37.1%	72.7%
RF - RTN	48.3%	*52.5%
Random (see table 1)	40%	50%

Table 2: F-scores of our model vs. the *baseline* for clauses of highest agreement (= 3) in test sets. "- RTN" (236 clauses) and "- *baseline*" (238 clauses) refer to what dataset was used for training and validation. "*" implies testing in a domain that was not part of the training set (RTN vs. baseline dataset), where we trained in one dataset and predicted on the other. Among the different feature-based models that we tried, a linear l1-penalized support vector machine (SVM) and a random forest (RF) reached highest performance. For clauses with agreement ≥ 2, we obtained 68.31% F-score (619 clauses).

Additionally, note that if a model always predicts the most common label (or randomly assigns them), the micro-F1-score (i.e., accuracy) for RTN would be 40% and for the baseline 50%. We found that when we used the feature-engineering approach proposed by Swanson et al. (2014) in the RTN corpus, the best trained and tested standard model, a random forest with 100 estimators (RF) (Breiman, 2001), does not perform well in this new corpus. Though, from table 2, we also see that it still does better than random (third vs. fourth row). This result suggests that sentence structure and part-of-speech (POS) do not generalize well to classify narrative clause types, as one would expect from POS being predominant in the top 10 most relevant features in this feature-engineering (original and baseline) approach. While the baseline model found POS features to be highly relevant, since our model uses word embeddings, POS information only contributed $2 - 3\%$ to the F-score. Furthermore, these results stress the difference between both story domains: video-recorded spoken narratives (RTN) vs. mini-blog written stories (baseline from Swanson et al. (2014)).

To sum up, the fact that a simple CNN performs well on this classification task, as illustrated in table 2, reflects the high disentangling power that Labov's model proposes for analyzing spoken personal narratives. Finally, since we can automatically annotate and thus disentangle narrative clauses under this framework, our approach shows to be plausible, so we now proceed to explore aspects of similarity.

6 People's perception of similarity

Aiming to understand the aspects (i.e., clause types) that ordinary people attend to the most when they think about similarities among stories, we proceeded as follows. We represent each story as a set of narrative clauses, where each clause is initially encoded into a high-dimensional vector by using the Universal Sentence Encoder (USE) introduced by Cer et al. (2018). Next, given stories *s* and *s'*, for each clause in *s* we find the closest clause in cosine similarity in *s'* ($s \rightarrow s'$), and vice versa ($s' \rightarrow s$), and obtain an average similarity score. Using this mechanism, we match stories only at clause-type subsets (action, evaluation, or orientation-type only). Finally, we sample 60 story pairs with average cosine similarity ≥ 0.5 for one of the clause types matches. See appendix A for some sample

matched stories.

For our experiments, we use these 60 stories, which are presented to Turkers in audio form only (as opposed to transcript text). While reading and listening might require different attention spans, since Labov's sociolinguistic model focuses on stories that are produced orally (just like these) and these are short stories – 62 seconds long on average – we rely on Turkers' auditory cognitive processing.

6.1 Annotation task: matching stories

We prompted: *"Which one of the following stories, A or B, was the most similar to the main story (and why)?"*. Each main story was annotated twice, switching order for A and B; one of these stories is matched at only one clause type level and the other is randomly selected. Table 3 shows these results.

Match only at	% of times detected
Action	67.8%
Evaluation	60.9%
Orientation	48.0%

Table 3: What aspects are paid attention. For those stories matched at the action-clause level, 67.8% of times Turkers recognized the matched story accurately, and selected the random story the remaining 32.2% of the times (these action-level matched stories were more than two times recognized correctly than incorrectly). Stories matched in evaluation-type clauses were also recognized accurately 60.9% of the times, which is 50% more than those stories that were wrongly recognized (60.9% vs. 39.1%). As for orientation-level matches, these were recognized somewhat randomly (48% of the times Turkers selected the matched stories and 52% of the times they selected a random story). Some reasons behind mismatches could be (1) that Turkers might be paying attention to other not covered aspects (further explored in section 6.2), (2) some randomness on annotations, and (3) the matching strategy.

From this experiment, we conclude that action and evaluation-type clauses were relevant for non-experts when they compared stories for similarity. Hence, our hypothesis on whether ordinary people rely on these Labov's aspects to compare narratives proved to be true for both action and evaluation aspects of a story but not for the orientation aspect.

6.2 Map of crowdsourced aspects to Labov's

Trying to understand how Turkers perceived the different aspects, and where mismatches could possibly come from, for the same 60 stories, we selected the story C that has score ≥ 0.5 at a given match

and has the smallest matching score for the other clause types. We asked *"Explain in what aspects (at least three) are the following personal narratives similar"*, hoping that Turkers would give reasons related to the matched dimensions. Note that with this open-ended question, Turkers were invited to think about *any aspects* that came to mind, thus we did not impose aspects on them beforehand.

Next, we map their responses to Labov's aspects; for example, the explanation *"They both have pessimistic thoughts..."* refers to how a narrator feels or perceives the situation → `evaluation` clause type. Some results of this mapping strategy are illustrated in table 4, and results for this mapping process are summarized in table 5.

Explanation	Mapped aspects
"Both started out in one direction and switched to a different field."	`action`
"Both people spoke about intense passion for something."	`evaluation`
"Both were from small towns."	`orientation`
"Both speakers suggest [ac] *pursuing their career goals* [ev] *makes them a better person* in real life too."	`action, evaluation`

Table 4: Examples of mapped explanations. We analyzed every open-ended explanation given by Turkers and mapped them to Labov's model according to what aspects these explanations were mostly referring to. Note that not all explanations were granular, hence, for some of them we highlighted more than one aspect (see fourth row in this table).

We show that for action- and evaluation-type clauses, Turkers mentioned aspect of similarity related to these clauses at least twice as often (as relevant) as the less relevant aspect in the matched stories, which (again) proves that these Labov's clause types can work as aspects of similarity.

As for orientation-type clauses, while they are still identified as reasons for similarity as illustrated in table 5, these are not the main reason to match two stories. We argue that this is due to the nature of our prompts to Turkers, which specifically asked for "stories" (section 6.1) or "narratives" (section 6.2); in ordinary people's mind (i.e., non-narrative experts), both of these concepts might not relate to the physical space or context where events and emotions/intentions happen, causing Turkers to not pay as much attention to them. It might also be that since all RTN stories are within the pathways/inspiration/career domain, people get

engaged with that part as opposed to if our domain were more diverse in topics, which would then have led people to recur to the orientation aspect (background/set-up/place) to match them in the absence of common feelings or similar actions/decisions among stories.

Match at	Action	Evaluation	Orientation
Action	**100**%	88%	44%
Evaluation	95%	**90**%	45%
Orientation	92%	96%	**58**%

Table 5: Aspects referenced in 180 explanations of similarity (3 for each of 60 stories). As expected from results in table 3, explanations related to action and evaluation aspects are highly present in detected reasons for similarity. We see that, for most story pairs, Turkers gave explanations regarding actions that happen within stories. In particular, for pairs matched at action-clause level, every pair was said to be similar due to similar actions. For evaluation-clause level matches, we find explanations mapped to that aspect twice as often as for the least present aspect (90% vs. 45%). Finally, while orientation-type clauses were not perceived as a main similarity aspect (see table 3), we find that for stories matched at orientation clauses, Turkers recognized this aspect to be a reason for similarity more often than for any other matches (58% / 45% = 1.28 times).

7 Conclusion

We introduce the largest corpus of annotated spoken personal narratives, to our knowledge, and develop a straightforward method to classify these narratives' clauses using Labov's sociolinguistic model. Our model's high performance in classification reflects the disentangling power that Labov's model offers for analyzing oral personal narratives. Only by being trained in our introduced corpus, our model performs well in an earlier proposed dataset of written stories. Furthermore, we propose the first attempt to understand whether ordinary people (i.e., non-narrative experts), such as Mechanical Turkers, rely on Labov's model to compare personal stories, and show that these people do rely on two out of three Labov's aspects of narrative. Namely, action-type and evaluation-type clauses are perceived as central aspects of comparison, but the same does not apply to, and remains unresolved for, orientation-type clauses. One natural next step would entail shedding light on how different questions' wording and emphasis, aimed at matching stories, affect what people think of as similarity aspects. We hope that these precursory findings about

the aspects that proved to underlie story-matching could also be used in a broader set of tasks, such as finding causal or temporal relationships between events, inferring mental states, or grounding actions and emotions in a story's set-up.

Finally, we acknowledge that we have only scratched the surface of this wonderfully rich space of personal narrative representations and of what people focus on when they compare stories. Our overarching goal, of modeling human judgment of narrative similarity and building a machine capable of replicating that behavior, leaves untouched several questions that future research should explore. For example, what other aspects should be examined to represent personal narratives, how to decide the relative relevance of these aspects, and how to model similarity judgments within aspects.

Acknowledgments

References

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. *University of Pennsylvania*, 97:100.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Tim vor der Brück and Marc Pouly. 2019. Text similarity estimation based on word embeddings and matrix norms for targeted marketing. 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 1827–1836.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Snigdha Chaturvedi, Shashank Srivastava, and Dan Roth. 2018. Where have i heard this story before? identifying narrative similarity in movie remakes. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 673–678.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Andrew M Dai, Christopher Olah, and Quoc V Le. 2015. Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Liye Fu, Jonathan P Chang, and Cristian Danescu-Niculescu-Mizil. 2019. Asking the right question: Inferring advice-seeking intentions from personal narratives. *arXiv preprint arXiv:1904.01587*.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

William Labov et al. 1967. Waletzky (1967). narrative analysis: Oral versions of personal experience. *Essays on the verbal and visual arts*, pages 12–44.

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee. 2013. A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, 26(7):1575–1590.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 321–330. ACM.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

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.

Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn Walker. 2014. Identifying narrative clause types in personal stories. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 171–180.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

A   Sample matched stories

These stories were matched in the action-clause level (stories A and B, with a similarity score of 0.58), and in the orientation-clause level (stories B and C, score of 0.50). Note that some clauses are not displayed due to space limitations.

Narrative clause	Clause type
"Did I have the pathway figured out, by no means, no at that time, right?	evaluation
So I also got involved in a atmospheric chemistry lab	**action**
, so nothing to do with animals	orientation
, but a lot to do with the environment.	orientation
I loved that, but I was like, well	evaluation
, I really wanna still apply this to animals.	evaluation
So I went on to graduate school	**action**
, and I enjoyed teaching, cuz I also worked as a teaching assistant at CSU Long Beach."	**action**

Story A

Narrative clause	Clause type
"When I was in school, I wanted to be a doctor.	**orientation**
I went to college	**action**
and I realized I actually didn't wanna be a doctor.	evaluation
I wanted to do something more in public health.	**orientation**
And so I went to graduate school	**action**
and I ultimately got a PhD in international relations and global health cuz I'm interested in this question on sort of a global level.	**action**
So although I started off wanting to be a doctor and although I never became a doctor, except that I guess I do get to be called Dr. Clinton because I have a doctorate degree.	**orientation**
I've figured out what my passion is and how to do that in a way that feels right for me."	evaluation

Story B

Narrative clause	Clause type
"Up until the time I got to college	**orientation**
, I still had the aspiration to maybe go to medical school.	**orientation**
Until I started to really reality hit in that my family wasn't very financially well off	evaluation
, and the reality of the fact that medical school costs a lot of money , takes a long time.	evaluation
And then it kinda broad my horizons a little bit in that I could explore some other options."	action

Story C