Recognizing Coherent Narrative Blog Content

James Ryan[1] and Reid Swanson[2]

[1] Expressive Intelligence Studio, University of California, Santa Cruz
[2] Institute for Creative Technologies, University of Southern California
jor@soe.ucsc.edu, rswanson@ict.usc.edu

Abstract. Interactive storytelling applications have at their disposal massive numbers of human-authored stories, in the form of narrative weblog posts, from which story content could be harvested and repurposed. Such repurposing is currently inhibited, however, in that many blog narratives are not sufficiently *coherent* for use in these applications. In a narrative that is not coherent, the order of the events in the narrative is not clear given the text of the story. We present the results of a study exploring automatic methods for estimating the coherence of narrative blog posts. In the end, our simplest model—one that only considers the degree to which story text is capitalized and punctuated—vastly outperformed a baseline model and, curiously, a series of more sophisticated models. Future work may use this simple model as a baseline, or may use it along with the classifier that it extends to automatically extract large numbers of narrative blog posts from the web for purposes such as interactive storytelling.

Keywords: coherence, blogs, content harvesting, machine learning

1 Introduction

Millions of weblog posts are published online each day, making up a text corpus of immense proportions for use in natural language processing (NLP) applications. A particular focus of some of these applications has been blog posts that are *personal narratives*, in which the author recounts a series of related events that happened in the past to her or a close associate. These narrative blog posts, examples of which are shown in Figure 1 and Figure 2, present massive potential in the form of story content that could be harvested and appropriated for use in interactive storytelling applications.

One promising avenue for such work would be to harvest content in support of a narrative-based virtual agent, a format that recent work on intelligent virtual agents suggests can be effective in a wide variety of applications. Narrative-based interaction discussing past emotional experiences, for example, is known to be highly beneficial for well-being: meta-analyses of 200 studies in which people actively retell personal stories demonstrate a striking impact on physical health in the form of reduced doctor visits and improved immune antibody responses [10, 38]. Building a narrative-based agent typically requires specification of the narrative world, with its causal and temporal relationships, as well as all the

Keep thinking of those accomplishments, guys. It's really neat to experience the payoff. My small victory today was in going to get my hair cut and dealing with my nerves. As always, I seem to forget how it goes because it's been a couple of months so I anxiety comes back despite having done this many times before, and it had me on edge about calling to see if my hairdresser was working today. I was worried she'd pick up like last time and I'd have no idea what to say and sound foolish like last time, but I made myself call anyway, despite feeling unprepared. Then when I got there, she wasn't ready for me yet so I sat down and felt my heart pounding and I wasn't sure what to do with my hands and worried about my posture. But I repeated some rational statements to myself about how I'd done this before and I could do it again etc, and within a few minutes I felt much better. Actually, I was more sociable than my hairdresser this time (she's bubbly so that's unusual) and I think I did just fine.

Fig. 1: An anonymously authored blog personal narrative about social anxiety.

possible interactions between the virtual agent and the user. This handcrafting approach has been used to produce compelling interactive agents for health and wellness [5], training [47], and entertainment [1]. But because considerable investment is required to author for this approach, recent work has investigated the prospect of harvesting narrative content from weblogs and developing methods to repurpose such material according to application-specific aims [35, 46, 15]. As an example, let us consider the personal narrative about social anxiety shown in Figure 1. Perhaps stories like this could provide narrative content that could be repurposed for a virtual storytelling agent specifically targeted at treating social anxiety. Similarly, stories about other medical conditions, such as strokes [15], could be used to support storytelling agents built to help users who are living in the wake of those circumstances.

More broadly, narrative blog content could be repurposed for use in a wide array of storytelling applications, and in fact a number of existing projects have already done this. An early example, *Buzz*, is an installation piece that renders harvested blog stories as monologues delivered by virtual agents [35]. More recently, in [20] blog stories are performed as dialogues between two embodied agents with adaptable personalities that affect the retellings. This is made possible by first encoding the blog stories into a semantic representation that the agent dialogue is generated from. Other recent applications have used this same method to generate story retellings that vary according to adjustable parameters, such as the teller's personality, using content originating in fables [41], videogames [31, 2], and blogs [32, 30]; a corpus of semantic encodings of blog stories has even been released [29]. In [26], character dialogue is generated by repurposing natural language extracted from books. Other projects have demonstrated the plausibility of learning narrative models from story corpora [24, 25, 19, 17]. Another line of work has repurposed narrative blog content for mixed-initiative story cocreation. An example of this is *Say Anything* [46], a system that collaborates with human users to build stories composed of user-submitted free text and related content that is extracted from blogs on the fly. More recently, Roemmele and Gordon [43] introduced *Creative Help*, which is a collaborative authoring tool that extends this method to give the user more authorial control [43], with plans for further extensions utilizing deep learning [42]. Clearly, the repurposing

of narrative blog content affords a rich space of expressive possibility for works of interactive storytelling.

While in a single week the blogosphere may produce a million or more personal narratives, these must be sifted out from the non-story blog posts that outnumber them considerably—previous work estimates that only around 15% of blog posts are narratives [14]. Gordon and Swanson [13] built a classifier to automatically identify stories from non-stories in weblog posts with a 66% precision, which they used to construct a corpus of over one million personal narratives. Many of these stories, however, are not sufficiently *coherent* for use in secondary applications. In a narrative that is not coherent, the order of the events in the narrative is not clear given the text of the story. For example, there may be a confusing disconnect between fabula and discourse.[3] We specifically refer to this concept as *temporal coherence*. (Figure 2 gives examples of both coherent and incoherent narrative blog posts taken from Gordon and Swanson's corpus.) This lack of coherence can trouble the kinds of promising applications of interactive storytelling that are made possible by harvesting coherent narrative blog content, an area we outlined above. This is because the appeal of harvesting narrative material for free is offset when additional inference is needed to determine the actual ordering of the events in the extracted stories. Non-narrative applications are undermined too, such as the emerging area that uses blog stories to learn causal relations between everyday events in support of automated knowledgebase construction [11, 12, 33, 15].

Our goal in this work is to identify the stories in Gordon and Swanson's corpus that have the *strongest temporal coherence*, specifically through the use of machine learning techniques that may do such sifting automatically. This is related to work in automated discourse coherence models that aims primarily to improve multi-document summarization tasks. A wide range of techniques have been developed in this area [4, 3, 23, 45, 8, 9, 39, 27, 28, 16], but these tend to require high-quality NLP tools, such as syntactic parsers and entity recognizers. In our work, we focus on shallow features because the very noisy nature of blog text is not handled well by current NLP tools.

To our knowledge, this represents the first investigation into developing automatic methods for estimating the temporal coherence of blog narratives. If successful, this project would yield a number of interesting benefits. First, it would produce a new corpus containing blog stories that are coherent enough to be used in the kinds of secondary applications that we have referenced above. Second, since the larger corpus of blog stories that we are sifting is varied and domain-independent, our project could provide a generalized method for the recognition of coherent narrative blog content—one that could be used by anyone to extract such content from other sources. Lastly, the development of such a generalized method could tell us about what makes blog stories coherent, thereby shedding light on the very essence of temporal coherence in narrative, more broadly.

[3] Here, we acknowledge that such decoupling is used productively in certain storytelling forms and that audience confusion may be socioculturally subjective [22, 34].

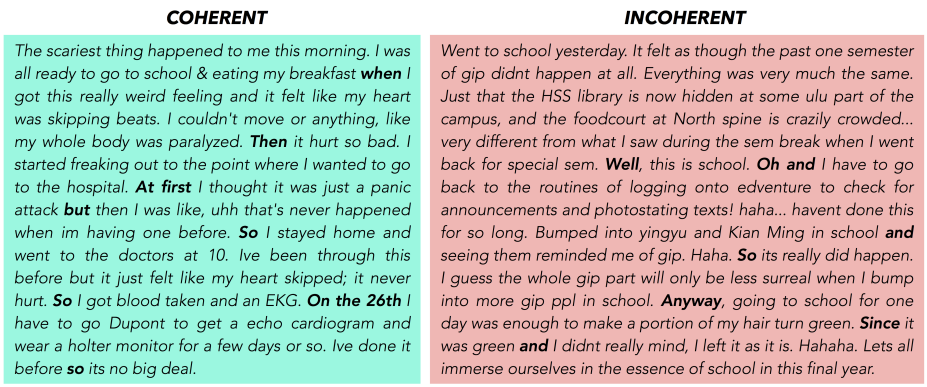

Fig. 2: Coherent and incoherent personal-narrative blog posts, authored anonymously. Notice that the order of events is much easier to infer in the coherent narrative. Discourse connectives are emphasized in **boldface**.

In this paper, we outline the development of several regression models that automatically predict the temporal coherence of personal narratives found in weblogs. These models were trained on 200 blog posts that were sampled from Gordon and Swanson's corpus and annotated for their coherence by Amazon Mechanical Turkers. Several types of features were extracted from the text of these posts and used to train regression models. As a baseline, we used twelve standard readability metrics similar to and including the Flesch-Kincaid grade-level measure of text readability [21] to train a comparable model, which underperformed each of ours in an evaluation procedure. In the end, our simplest model—one that only considers text punctuation and capitalization—performed best, showing high enough correlation with temporal coherence to be of immediate practical use in corpus sifting. While this does not teach us much about temporal coherence in narrative, or even in blog stories specifically, the failures of our more sophisticated models may. Moreover, the simplicity of our best-performing model (and the classifier that it extends) indicates that our technique could easily be duplicated by others to obtain huge numbers of blog posts from the wild that are temporally coherent narratives. We are excited at the prospect of future applications of interactive storytelling that could be enabled by such abundances of raw narrative material.

2 Background

This study extends work by Gordon and Swanson, who built a corpus of 1.7M personal narratives [13], as described above. This corpus was extracted from the ICWSM 2009 Spinn3r Dataset, a collection of 44M English blog posts published between August and October of 2008 [6], using a linear classifier. However, as noted above, much of these automatically recognized stories are not sufficiently coherent for use in secondary applications. As an extension to this earlier work, we see our current project as concerning the development of a *second-order* clas-

sifier that may be used in conjunction with Gordon and Swanson's. Specifically, we envision an extraction pipeline that proceeds as follows: assemble a large collection of blog posts scraped from the web; use Gordon and Swanson's classifier to refine this collection to a subset that are identified as stories; finally, further refine this set to only the stories that our classifier recognizes as being temporally coherent.

3 Method

In this section, we describe our data collection and annotation processes, feature extraction methods, and model training and evaluation procedures in detail.

3.1 Data Collection and Annotation

We randomly sampled 200 blog posts from Gordon and Swanson's corpus, described above, for annotation in a task posted online through Amazon Mechanical Turk [36]. Only posts of between 150–200 words were selected during sampling, and abridged posts in the corpus, identifiable by a final ellipsis, were also excluded. Figure 2 shows two typical examples taken from this sample.

After selection, the 200 posts were subjected to a crowdsourced annotation procedure administered online as an Amazon Mechanical Turk Human Intelligence Task (HIT) [36]. Each task in the HIT included two triplets of blog posts that were preceded by a set of instructions defining narrative temporal coherence. For each triplet, the annotator was asked to rank its three stories in order of temporal coherence. Additionally, as a way to ensure that the tasks were completed in good faith, annotators were required to give a one-sentence summary characterizing the main point of the narrative. Every blog post appeared in two triplets, and each triplet was annotated by three separate individuals. In total, 70 unique annotators attempted the HIT, of which the work of three was rejected for being incomplete or indicative of poor-faith effort. Of the 67 annotators whose work was accepted, each completed an average of three tasks. Using the annotator's choices for most-coherent narrative for each triplet, Cohen's kappa coefficient was 0.41 across all annotators, indicating moderate agreement.

As each blog post appeared in two triplets that in turn were each annotated three times, each post received six coherence rankings in total. These rankings were treated as real numbers—`1.0` for *most coherent*, `2.0` for *second-most coherent*, and `3.0` for *least coherent*—and were used to compute mean coherence rankings for each post, which we then used as the posts' gold-standard labels.

3.2 Feature Extraction

Over 40,000 total features across several features sets were extracted from each blog post. These sets, shown in Table 1, include both low-level and high-level linguistic features, and also superficial text features pertaining to capitalization and punctuation. Feature values were generally frequencies, as we explain below.

Unigrams are the lowest-level linguistic features that were extracted. Intuitively, these may not seem to be good predictors of narrative structure, but these were the best performing feature used in Gordon and Swanson's classifier

Features	N	Description
Unigrams	5,374	Relative frequency of the lexical unigrams.
Bigrams	24,202	Relative frequency of the lexical bigrams.
POS unigrams	39	Relative frequency of the POS-tag unigrams.
POS bigrams	861	Relative frequency of the POS-tag bigrams.
LIWC	81	Relative frequency of Linguistic Inquiry and Word Count categories.
Discourse	83	Combination of relative frequency of discourse-connective categories and the relative frequency of the individual connectives from the Penn Discourse Treebank 2.0 annotation manual [40].
Capitalization/ punctuation	3	Percentage of words capitalized, percentage of sentence-initial words capitalized, and percentage of unigrams that are punctuation.
Readability	12	Scores from standard readability metrics.
Syntax	2,533	Combination of average per-word probability of parse trees (Syntax:TreeScoreMean) and relative frequency of each context-free production rule.

Table 1: Feature sets extracted for each blog post.

[13], and so we deemed them a good starting point. For similar reasons, we also extracted lexical bigram features. Across all 200 blog posts, there were 5,374 unique unigrams and 24,202 unique bigrams. Values for these features represent the frequency—normalized to the maximum frequency count for that feature, given all 200 blog posts—with which a particular n-gram occurred in the post at hand. At a higher linguistic level, we calculated relative frequencies for POS-tag unigrams and bigrams using the Penn Treebank tagset [44]. These are easy features to extract and could conceivably be predictive of narrative coherence—for instance, it could be that certain verb tenses are used more often in coherent narratives. At a higher level yet, we also parsed the stories to extract two features that characterize syntactic well-formedness: the average probability of their parse trees and the relative frequency of each context-free production rule.

We also extracted features corresponding to categories from the Linguistic Inquiry and Word Count (LIWC) dictionary [37], which maps over 2500 English words to one or more syntactic or semantic categories. These include categories pertaining to positive emotions, negative emotions, causal words, time markers, as well as function words, past-tense words, present-tense words, and several

more classes. Among these, we were especially interested in the causal-word and time-marker categories, which we expect to appear in coherent narratives. The features that we most expected to predict narrative coherence were those corresponding to the 43 categories of *discourse connectives* given in the Penn Discourse TreeBank 2.0 Annotation Guidelines [40]. As many discourse connectives specifically describe causal or temporal relations, it is quite intuitive that these would appear often in coherent narratives. 254 unique discourse connectives are included across the 43 Penn Discourse TreeBank categories, some examples of which are shown in Table 2. While several connectives appear in multiple categories, due to resource constraints we did not enact any sense-disambiguation procedure. Values for the features in these two sets were relative frequencies, as with the above feature sets.

Our final feature set is not linguistic, but rather comprises three superficial text features pertaining to capitalization and punctuation. These were included not because they cause narrative coherence, but because they could conceivably be *associated* with narrative coherence. That is, it is reasonable that writers of well-formed blog narratives would use capitalization and punctuation more often than writers of poorly formed ones. The three features in this set were the percentage of words that are capitalized, percentage of sentence-initial words that are capitalized, and percentage of unigrams that are punctuation. These percentages were not normalized in the way that the above feature values were.

3.3 Model Derivation and Evaluation

Each of the feature sets described in Section 3.2, as well as a superset of all features and additional sets derived by feature selection, were used to train regression models with WEKA [48]. Additionally, we built a simple baseline model using twelve readability metrics similar to and including the Flesch-Kincaid grade-level measure of text readability [21]—standard measures that could conceivably predict narrative coherence. After initial exploration using several algorithms trained on the various feature subsets, we found the M5P model tree with default parameters to give the best results, so we used this exclusively. We employed 10-fold cross-validation to evaluate the models.

4 Results and Discussion

Table 2 shows the performance of the models that we trained. For each model, correlation coefficients and root relative squared errors were averaged across ten trials of 10-fold cross-validation, *i.e.*, over testing on 100 10% folds. A 100% root relative squared error is equivalent to always estimating the *mean* label value (*i.e.*, the mean coherence ranking, which is naturally `2.0`).

Clearly, the baseline model trained using text readability metrics was outperformed by the other models. Unexpectedly, the best model was the simplest—one that considers only the percentage of words that are capitalized, percentage of sentence-initial words that are capitalized, and percentage of unigrams that are punctuation. We believe that these superficial text features do not *cause* narrative coherence, but rather are merely associated with it. That is, writers of

Feature subset	10-fold cross-validation	
	Correlation coefficient	**Root relative squared error**
Capitalization/punctuation	0.474	88.2%
Feature selection (top 30)	0.454	89.7%
POS unigrams	0.385	93.94%
LIWC	0.379	93.96%
POS bigrams	0.376	98.52%
All features	0.331	98.75%
Unigrams	0.309	103.53%
Syntax	0.301	102.21%
Discourse	0.294	96.84%
Bigrams	0.278	102.71%
Readability (baseline)	0.141	103.75%

Table 2: Performance of the models trained using each feature set. Coefficients and errors represent averages across ten iterations of 10-fold cross-validation.

temporally coherent narrative blog posts are likely to use more punctuation and capitalization than writers of incoherent posts are. In any case, this is another example of a cheap and simple feature set performing surprisingly well [18].

The model that we did expect to perform best, trained on the Penn Discourse Treebank discourse-connective categories, did not produce an exceptionally high correlation coefficient. A possible explanation comes from the fact that we performed no sense disambiguation on the connectives appearing in the blog posts. Because of this, the counts for *every* category that was associated with *any* sense of a connective token were incremented for each appearance of that token, and this may have brought consequential noise into these feature values.

Interestingly, models trained with feature selection perform better on this task than models that use all the features. We examined the top selected features in order to further investigate why some features were not performing as well as expected. As shown in Table 3, none of the discourse-connective features appear among these, although we do observe other features appearing here that belong to sets whose models performed poorly. Another potential issue is the small amount of training data that we used—only 200 stories. While our data collection was fairly costly, it appears that collecting annotations for several hundred more blog narratives would be helpful. Figure 3 plots correlation of the best-performing model with the test set as a function of training set size, suggesting that more data would help performance.

Features	**Weight**
LIWC:Unique Words	0.154
SyntaxRule:ROOT → S	0.150
Syntax:TreeScoreMean	0.149
LIWC:Total Function Words	0.121
POS Bigrams:[DT, NN]	0.112
LIWC:Dictionary Words	0.111
Unigrams:'-'	0.104
SyntaxRule:NP → NP NP	0.103
Readability:Gunning-Fog	0.103
Readability:Flesch-Kincaid	0.103
POSUnigrams:':'	0.103
POSBigrams:[NNP, :]	0.103
POSBigrams:[:, NNP]	0.103
Unigrams:';'	0.103
Unigrams:'.'	0.103
LIWC:Dash	0.103
POSUnigrams:CC	0.103
POSBigrams:[:, CD]	0.102
POSBigrams:[:, NN]	0.102
CaptPunct:SentenceCap	0.098

Table 3: The top twenty features (determined by feature selection), along with their weights. A model trained using the top thirty performed second-best.

5 Conclusion and Future Work

Interactive storytelling applications have at their disposal massive numbers of human-authored stories, in the form of narrative blog posts, from which story content could be harvested and repurposed. Such repurposing is currently inhibited, however, in that many blog narratives are not sufficiently *coherent* for use in these applications. In a narrative that is not coherent, the order of the events in the narrative is not clear given the text of the story. In this paper, we have presented the results of a study exploring automatic methods for estimating the coherence of narrative blog posts. In the end, we found that our simplest and best-performing model, one that considers only the degree of capitalization and punctuation in a blog narrative, shows correlation coefficients sufficient for immediate use in extracting many thousands of temporally coherent stories from Gordon and Swanson's corpus of over one million blog posts [13].

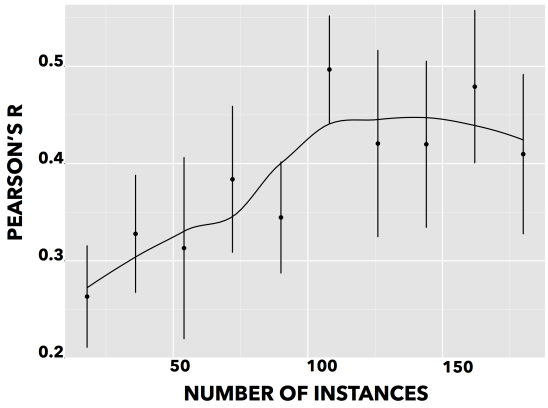

Fig. 3: A learning curve showing model performance as a function of training set size, which suggests that more data would help.

While the model trained using text readability metrics served as a suitable baseline here, it is not surprising that it was outperformed. This measure, though in wide use, merely considers the number of words per sentence and the number of syllables per word in a text. Of course, as our best-performing model was hardly more sophisticated, this baseline model could have also proven to correlate well with narrative temporal coherence—but it did not. Future work exploring automatic methods for determining the temporal coherence of narrative blog content should use our punctuation-and-capitalization model as a baseline.

We still maintain that discourse connectives, particularly ones cuing temporal or causal relations, may be a good predictor of narrative temporal coherence. Unfortunately, we believe that our method for extracting features of this type—particularly our lack of disambiguation of connective sense—introduced enough noise to significantly hamper performance of the corresponding model. In future work, connective-sense disambiguation should be used before extracting such features, and ideally a much larger training set should be used, as our findings suggest that this may produce better or more discriminative models. We also believe, following findings by others [39, 7], that we will have to incorporate more complex features and leverage models that consider the sequential nature of text.

While additional work is needed to home in on the recognizable expressions of temporal coherence in surface-level story discourse—one of the goals of this study—both our regression model and the classifier that it extends are simple enough for others to use for their own purposes. Given how feasible this makes the extraction of massive numbers of coherent blog stories from the web, we are excited at the prospect of future applications of interactive storytelling that would be enabled by such abundances of raw narrative material.

6 Acknowledgments

This work would not have been possible without Marilyn Walker, who provided mentorship and funded the annotation procedure presented in this paper.

References

1. Adam, C., Cavedon, L.: A companion robot that can tell stories. In: Proc. Intelligent Virtual Agents (2013)
2. Antoun, C., Antoun, M., Ryan, J.O., Samuel, B., Swanson, R., Walker, M.A.: Generating natural language retellings from Prom Week play traces. Proc. Procedural Content Generation (2015)
3. Barzilay, R., Lapata, M.: Modeling local coherence: an entity-based approach. In: Proc. ACL (2005)
4. Barzilay, R., Lee, L.: Catching the drift: Probabilistic content models, with applications to generation and summarization. In: Proc. NAACL HLT (2004)
5. Bickmore, T., Schulman, D., Yin, L.: Engagement vs. deceit: Virtual humans with human autobiographies. In: Proc. IVA (2009)
6. Burton, K., Java, A., Soboroff, I.: The ICWSM 2009 Spinn3r dataset. In: Proc. Weblogs and Social Media (2009)
7. Eisenberg, J.D., Yarlott, W.V.H., Finlayson, M.A.: Comparing extant story classifiers: Results & new directions. In: Proc. CMN (2016)
8. Elsner, M., Charniak, E.: A unified local and global model for discourse coherence. In: Proc. NAACL (2007)
9. Elsner, M., Charniak, E.: Coreference-inspired coherence modeling. In: Proc. ACL (2008)
10. Frisina, P.G., Borod, J.C., Lepore, S.J.: A meta-analysis of the effects of written emotional disclosure on the health outcomes of clinical populations. Nervous and Mental Disease 192(9) (2004)
11. Gerber, M., Gordon, A.S., Sagae, K.: Open-domain commonsense reasoning using discourse relations from a corpus of weblog stories. In: Proc. Formalisms and Methodology for Learning by Reading (2010)
12. Gordon, A., Bejan, C., Sagae, K.: Commonsense causal reasoning using millions of personal stories. In: Proc. AAAI (2011)
13. Gordon, A., Swanson, R.: Identifying personal stories in millions of weblog entries. In: Proc. Weblogs and Social Media (2009)
14. Gordon, A.S., Cao, Q., Swanson, R.: Automated story capture from internet weblogs. In: Proc. Knowledge Capture (2007)
15. Gordon, A.S., Wienberg, C., Sood, S.O.: Different strokes of different folks: Searching for health narratives in weblogs. In: Proc. Social Computing (2012)
16. Guinaudeau, C., Strube, M.: Graph-based local coherence modeling. In: Proc. ACL (2013)
17. Guzdial, M., Harrison, B., Li, B., Riedl, M.O.: Crowdsourcing open interactive narrative. In: Proc. Foundations of Digital Games (2015)
18. Hand, D.J., et al.: Classifier technology and the illusion of progress. Statistical Science (2006)
19. Harrison, B., Riedl, M.O.: Towards learning from stories: An approach to interactive machine learning. In: Proc. AAAI (2015)
20. Hu, C., Walker, M.A., Neff, M., Tree, J.E.F.: Storytelling agents with personality and adaptivity. In: Proc. IVA (2015)
21. Kincaid, J.P., et al.: Derivation of new readability formulas for Navy enlisted personnel. Tech. rep., DTIC Document (1975)
22. Labov, W.: Uncovering the event structure of narrative. Round Table on Language and Linguistics (2003)

23. Lapata, M., Barzilay, R.: Automatic evaluation of text coherence: Models and representations. In: Proc. IJCAI (2005)
24. Li, B., Lee-Urban, S., Appling, D.S., Riedl, M.O.: Crowdsourcing narrative intelligence. Advances in Cognitive Systems 2(1) (2012)
25. Li, B., Lee-Urban, S., Johnston, G., Riedl, M.: Story generation with crowdsourced plot graphs. In: Proc. AAAI (2013)
26. Li, B., Thakkar, M., Wang, Y., Riedl, M.O.: Data-driven alibi story telling for social believability. In: Proc. Social Believability in Games (2014)
27. Lin, Z., Ng, H.T., Kan, M.Y.: Automatically evaluating text coherence using discourse relations. In: Proc. ACL: HLT (2011)
28. Louis, A., Nenkova, A.: A coherence model based on syntactic patterns. In: Proc. EMNLP-CoNLL (2012)
29. Lukin, S.M., Bowden, K., Barackman, C., Walker, M.A.: Personabank: A corpus of personal narratives and their story intention graphs. In: Proc. LREC (2016)
30. Lukin, S.M., Reed, L.I., Walker, M.A.: Generating sentence planning variations for story telling. In: Proc. SIGDIAL (2015)
31. Lukin, S.M., Ryan, J.O., Walker, M.A.: Automating direct speech variations in stories and games. In: Proc. GAMNLP (2014)
32. Lukin, S.M., Walker, M.A.: Narrative variations in a virtual storyteller. In: Proc. IVA (2015)
33. Manshadi, M., Swanson, R., Gordon, A.S.: Learning a probabilistic model of event sequences from internet weblog stories. In: Proc. FLAIRS (2008)
34. Michaels, S.: "Sharing time": Children's narrative styles and differential access to literacy. Language in Society 10(03) (1981)
35. Owsley, S.H., Hammond, K.J., Shamma, D.A., Sood, S.: Buzz: Telling compelling stories. In: Proc. Multimedia (2006)
36. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. Judgment and Decision Making 5(5) (2010)
37. Pennebaker, J.W., Francis, L.E., Booth, R.J.: LIWC: Linguistic Inquiry and Word Count (2001)
38. Pennebaker, J.W., Seagal, J.D.: Forming a story: The health benefits of narrative. Clinical Psychology 55(10) (1999)
39. Pitler, E., Nenkova, A.: Revisiting readability: a unified framework for predicting text quality. In: Proc. EMNLP (2008)
40. Prasad, R., et al.: The Penn Discourse Treebank 2.0 annotation manual (2007)
41. Rishes, E., Lukin, S.M., Elson, D.K., Walker, M.A.: Generating different story tellings from semantic representations of narrative. In: Proc. ICIDS (2013)
42. Roemmele, M.: Writing stories with help from recurrent neural networks. In: Proc. AAAI (2015)
43. Roemmele, M., Gordon, A.S.: Creative help: A story writing assistant. In: Proc. ICIDS (2015)
44. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision) (1990)
45. Soricut, R., Marcu, D.: Discourse generation using utility-trained coherence models. In: Proc. COLING/ACL (2006)
46. Swanson, R., Gordon, A.S.: Say anything: A massively collaborative open domain story writing companion. In: Proc. ICIDS (2008)
47. Traum, D., et al.: Hassan: A virtual human for tactical questioning. In: Proc. SIGDIAL (2007)
48. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, CA (2005)