Identifying Sensible Lexical Relations in Generated Stories

Melissa Roemmele
SDL Research
mroemmele@sdl.com

Abstract

As with many text generation tasks, the focus of recent progress on story generation has been in producing texts that are perceived to "make sense" as a whole. There are few automated metrics that address this dimension of story quality even on a shallow lexical level. To initiate investigation into such metrics, we apply a simple approach to identifying word relations that contribute to the 'narrative sense' of a story. We use this approach to comparatively analyze the output of a few notable story generation systems in terms of these relations. We characterize differences in the distributions of relations according to their strength within each story.

1 Introduction

Current text generation systems are frequently able to produce output that is linguistically well-formed with regard to sentence-level syntactic and lexical dependencies. Still, when people perceive the generated text as a whole, it often doesn't appear to "make sense". There are many dimensions to what qualifies a text as sensible. Recent work has focused on trying to model commonsense knowledge and reasoning via the domain of narrative. From the perspective of this work, stories encode the rich set of coherence relations between entities and events by which people interpret their experiences. This has led to frameworks that evaluate automated commonsense reasoning through story modeling tasks like predicting what happens next in a story (Mostafazadeh et al., 2016). Accordingly, the challenge of story generation systems is to express the same commonsense relations that establish the coherence of human-authored stories. One barrier to addressing this challenge is how to quantify the presence of these relations in a text. People can readily provide intuitive judgments about whether a story makes sense, but there has been little exploration of cues for these judgments that can be modeled by current NLP analyses, even relatively shallow ones. We address this in this work by examining a simple approach to detecting lexical relations that contribute to the coherence (or what we call 'narrative sense') of a story. We apply this approach to compare the output of a few different story generation systems according to these relations.

Evaluation in general is an ongoing challenge in text generation research, and particularly for open-ended content like stories. Some work has borrowed automated metrics used for evaluation in other generation tasks (e.g. BLEU for machine translation). However, such metrics expect that there is a fixed set of gold standard references to which output should be compared, which is not a fitting assumption for many story generation frameworks. If the task is to generate a story about a particular topic or to generate a story given an opening sentence, there is no finite set of "correct" stories that meet the objective. For this reason, most work relies on human judgment for evaluation (e.g. Fan et al., 2018; Holtzman et al., 2018; Roemmele and Gordon, 2018), often through a quantitative rating or ranking scheme for selected quality dimensions (e.g. asking "how coherent is this story?" or "which story is more coherent?" among a set of candidates). While these judgments are a reliable indicator of the relative impact of different generation models, they are costly in that they must be repeated for each new set of generated output. Moreover, relying on holistic ratings/rankings of quality does not provide insight into the text-level features that influence these judgments. Qualitative feedback is useful for this, but it can be difficult for people to precisely verbalize their intuition about what makes a generated text sound good or bad. Fully modeling this judgment may require sophisticated nat-

ural language understanding capabilities, but we can still investigate whether shallow indicators of this judgment are available.

As more text generation systems are being deployed, comparative evaluations between them are becoming increasingly important. Many researchers have released story generation models trained through their own particular experiments, including those described in Section 2. These already-trained models can be readily used by other NLP practitioners, but re-training them can often require significant time and resources due to their complexity. In some cases (e.g. the GPT-2 system described below), the procedure for training the model is not publicly available. Still, this does not mean any comparative evaluation between systems trained on different datasets is fruitless. Such evaluations may not be able to completely disentangle the contribution of a particular algorithmic approach versus that of the dataset itself, but they can still illuminate the relative impact of each model in the stories it produces. Moreover, they can also help scrutinize any qualitative claims made about the performance of a particular system, since sometimes such claims are based on a handful of carefully selected examples. Our vision is to move towards frameworks that can analyze the characteristics of story generation models even when they are presented as black boxes, simply by observing the stories they generate.

In this work, we analyze stories in terms of word relations in order to investigate whether such relations can be examined as an indicator of narrative sense. As outlined in Section 3, to capture word relations we use a generic NLP technique of calculating statistical word co-occurences in a corpus of stories, in particular by using the Pointwise Mutual Information (PMI) (Church and Hanks, 1990) statistic. The use of statistical word association measures in narrative modeling tasks is familiar. There is work on using these measures to evaluate coherence in news stories (Shahaf and Guestrin, 2010). Other work has used word co-occurrence statistics to predict commonsense cause-effect relations between sentences (Gordon et al., 2011; Luo et al., 2016; Riaz and Girju, 2013; Sasaki et al., 2017). A related line of research has focused on modeling pairs of verb-argument units in narrative text in order to induce story event sequences (Chambers and Jurafsky, 2008; McIntyre and Lapata, 2009; Rudinger et al., 2015). Other relevant tasks like emotional framing of narrative (Jurafsky et al., 2014), sentence completion based on reading comprehension (Woods, 2016), and creative language generation (Toivanen et al., 2014) have also been addressed using lexical association measures. Most relevant to generation evaluation, Sagarkar et al. (2018) demonstrated that quality ratings of generated stories correlated significantly with the average PMI score of their component word pairs found in a story corpus. Our work takes a look at the distribution of word pair PMI scores in stories generated by alternative approaches that have not yet been directly compared.

2 Generation Task

We examined four models that have specifically been applied to story generation and which generate stories based on a seed input text. We used 15,138 items from the test set of the English-language Reddit WritingPrompts dataset[1] (Fan et al., 2018) as these seed inputs. This dataset is derived from the /r/WritingPrompts subreddit, where users write and share fiction stories in response to a story premise (the prompt). Each item in this dataset consists of a prompt and a human-authored story. For each prompt, we generated a story with each of the models described below. We did not train our own version of these models but instead used the already-trained models provided by the respective authors on their linked GitHub repositories.

CREATIVEHELP[2] (Roemmele and Gordon, 2018): An RNN language model trained on a subset of 11,000 self-published fiction books in the Toronto BookCorpus[3].

GPT-2[4] (Radford et al., 2019): A transformer-based language model trained on several million webpages in the WebText corpus. The full model advertised in the paper is not publicly available, so we used the 'small' version of the model. While GPT-2 has been presented as a generic text generation system as opposed to one specifically for story generation, the examples presented by the authors include narrative text.

HIERARCHICAL[5] (Fan et al., 2018): A convolutional sequence-to-sequence model with self-

[1] See github.com/pytorch/fairseq for link to dataset
[2] github.com/roemmele/narrative-prediction
[3] yknzhu.wixsite.com/mbweb
[4] github.com/openai/gpt-2
[5] github.com/pytorch/fairseq

Prompt
On your way to Mars on the first voyage ever , you discover that you and the countless others on the ship are intended ritual sacrifices by Elon Musk . You now must avoid reaching Mars at all costs .

CREATIVEHELP	Your all - sister , state - and - seek , is exactly what you believed was William , WP thought , and agreed with the program . Elon Musk and Jett moved into their strong chamber . At the group of Ryan over from Tumblr , the third was fallen and a wise top dancer and a united East Africans . Here was a light , and I felt really fear or make out William . My heart leapt to my throat ; I 'd trapped Riley in no life between her But Anna was afraid . The real baby boy was doing something for me had led me out of his room . From the fire , it closed and connected with him . Elon Musk looked down at her hands and said , " not just you ! " She looked at me , then shrugged and
GPT-2	But if they can , do so . On Your Way To Mars on the first voyage ever , you discover that you and the countless others on the ship are intended ritual sacrifices by Elon Musk . You now must avoid reaching Mars at all costs . But if they can , do so . Borough of Babel : On your way to Pluto The planet currently inhabited by Jekyllands Islet and its looters bears benefit isty , albeit unable to argue that it is indeed inhabited . Or itself is ; consider the inhabitants of Copernicus themselves , as walking amongst them , does so represent the rarity of the copernican planet , which is understandably deplored . 343.FIRE Required Strange preview gained ID of Wurmby , and 25705 Heat up Any an important bounty hunter might pursue , the real tyrant looks like Javian 's gaze .
HIERARCHICAL	" I am an old man , old man. " " You are the last human to be here , you are the one who will never find us ! " " But I have been the one to bring you the one who will never come back ! " A long pause followed as a young man in a suit and helmet stepped in front of a large screen . " I am not a human man , but I am a man . I am the last human . I am the only human who can not be found . I have been the last human to be taken , you will become the last human . I am the only human . I have been the last human , and every human has been given to me . I am the last human
L2W	it is your duty to make sure that you are safe . " the two men exchanged a long look , then nodded and left . as the door closed behind them , the door swung shut behind them . the room was silent . the room was empty . the room was silent except for the sound of heavy breathing . the room was filled with the sounds of shuffling and shuffling . the room was pitch black and pitch black . the walls and floor of the room were strewn with debris and debris . the room seemed to be suspended above the floor . there were no lights in the room . no sign of anyone else . no one was in sight . there were only two chairs . one of them was a desk . a small desk . a desk sat on the floor
HUMAN	I hid under the tiny bed of my room , terrified . I covered my mouth , trying to muffle my heavy breathing . I heard its footsteps in the distance , dragging it 's feet as it walked . I heard it sniff the air , grunting , as it caught on to a scent . It started walking in a different direction , before suddenly breaking out into a sprint . I gasped instinctively , before realizing it was moving away from me . I lay there , in complete darkness , filled with fear . It had broken the main generators , and the backups only provided power to the main hallways . In the distance , I heard a faint scream , which was soon replaced by a deafening silence . I slowly moved out from under my bed . My hands were trembling , tears

Table 1: Example of a story generated by each model for a given prompt

attention on the decoder. The approach also uses a fusion mechanism that further encourages conditioning on the input while generating. This model was trained on the Reddit WritingPrompts dataset, which is the same dataset we use to seed generation in our work (we use the test set that was not observed by this model during training).

L2W[6] (Holtzman et al., 2018): An RNN language model enhanced with discriminator mechanisms that promote non-repetition, semantic entailment between sentences, relevance, and lexical diversity in the generated output. As with CREATIVEHELP, this model was trained on the BookCorpus stories.

One detail to note is that among these models, only the HIERARCHICAL model is specifically trained to observe the prompt as text that is independent from the generated story itself. The other models are designed for 'story continuation', i.e. generating the next segment of an initial story. Here, these models viewed the prompt as the initial sequence in the story which is continued by the generated text. However, we subsequently disregard the prompt in our analysis and instead focus only on the relations within the generated text itself. These intra-story relations can still be compared across models without consideration of their relevance to the input texts.

Our analysis requires that the generated stories be comparable in length, so we limited the length of each story to 150 tokens. In some cases, due to the design of each model (e.g. some models complete generation when an end-of-story token is generated), the resulting stories were shorter. There were also instances in which the human-

[6] github.com/ari-holtzman/l2w

authored story was shorter. Consequently, we filtered any set of stories associated with the same prompt where at least one of stories contained fewer than 150 tokens. This resulted in 13,453 stories being included in our analysis. Table 1 shows an example of a prompt and the generated stories for that prompt alongside the corresponding human-authored story (labeled HUMAN).

3 Narrative Sense Relations

In line with the discussion above, we refer to the lexical relations examined in this work as 'narrative sense' relations. By scoring word pairs according to how often they appear in the same story, higher scores will indicate pairs with a stronger relation across different stories, i.e. words for which it makes sense that they would appear in the same story.

Though the inputs we provide to the models come a particular genre of English-language stories (self-published internet fiction), we wanted to examine lexical relations that span across different types of stories. Accordingly, we derived the narrative sense relations from four highly-utilized story corpora described below that (to the best of our knowledge) were not observed by the models during training. Obviously, it is not possible to construct a dataset which has full coverage of all sensible pairs that could appear in a set of generated stories. We selected these four diverse corpora to aim for as broad of coverage as possible without overly biasing the dataset towards pairs contained in the training data for any one of the models.

ROCStories[7] (Mostafazadeh et al., 2016): 97,027 five-sentence narratives authored via crowdsourcing. Authors were specifically asked to write stories in simple English about common everyday scenarios.

Visual Information Storytelling (VIST)[8] (Huang et al., 2016): 50,200 five-sentence stories also authored through crowdsourcing. Authors were prompted to write a story from a sequence of photographs depicting a salient "storyable" event.

CMU Plots[9]: 58,862 book/movie plot summaries extracted from Wikipedia. We truncated each of these summaries to its first 150 tokens, consistent with the length of generated stories.

Children's Book Test[10] (Hill et al., 2016): 98 children's novels authored between 1850 and 1950 and freely available through Project Gutenberg (we used the training set only of the full dataset). We segmented each book into passages of 150 tokens, which resulted in 36,987 passages (we subsequently treated each passage as its own story).

We tokenized all 244,216 stories in these corpora and applied lemmatization to the word tokens[11]. Since our analysis targets content words, we removed punctuation/symbols, numbers, and all words included in an English stopword list. We also removed proper nouns in order to reduce story-specific relations such as entity names. We then established a vocabulary of words occurring in at least five stories. As mentioned above, we calculated the PMI co-occurrence of these words. PMI is calculated for each word pair $(w1, w2)$ based on how often the words appear together relative to their individual frequency:

$$PMI(w1, w2) = \frac{count(w1, w2)}{count(w1) * count(w2)} \quad (1)$$

Here, a co-occurrence between two words was counted any time they appeared in the same story, without regard to their order. There is one exception: when the words occur within the same trigram, they are not counted as a co-occurrence. Our aim in doing this was to minimize relations between words that are phrase-dependent in favor of capturing relations that span across the story. This, in addition to the filtering of stop words and ignoring word order, helps to separate narrative sense relations from words that are related by grammatical dependencies, which is not what we are targeting with this analysis.

Using this methodology, we extracted and computed PMI scores for 7,829,163 word pairs consisting of 23,592 lemmatized words in the given dataset. Scores are computed in log space, as shown in Table 3. The scores in this dataset range from -17.25 to -2.30, with a median of -11.66.

4 Analysis of Generated Stories

For each generated story, we applied the same processing done for the stories in the narrative sense relations dataset, i.e. lemmatizing and removing proper nouns, punctuation/symbols, and

[7] cs.rochester.edu/nlp/rocstories

[8] visionandlanguage.net/VIST

[9] cs.cmu.edu/~ark/personas; cs.cmu.edu/~dbamman/booksummaries.html

[10] fb.ai/babi

[11] Tokenization, POS tagging, lemmatization, and stopword removal was done with spaCy: spacy.io

	CREATIVEHELP	GPT-2	HIERARCHICAL	L2W	HUMAN
1. Total raw words	6943	58966	4942	3173	34401
2. Total recognized words	5008 (72.1%)	18165 (30.8%)	4617 (93.4%)	2937 (92.6%)	17068 (49.6%)
3. Mean stories per word	108.51	32.85	62.56	96.95	34.38
4. Mean words	40.39	44.35	21.46	21.17	43.62
5. Mean word pairs	832.26	1283.79	254.54	225.42	1159.00
6. Mean seen word pairs	790.59 (95.2%)	980.85 (77.1%)	242.33 (96%)	217.88 (97%)	937.72 (82.3%)

Table 2: Statistics for the number of unique words and word pairs across all 13,453 evaluated stories

stopwords. Each story is represented as a set of unique words (disregarding their frequencies), and all pairwise combinations between these words are considered in the analyses.

4.1 Word Statistics

Table 2 contains some descriptive statistics for the generated words/pairs according to each model. Note that the term 'word' in this table refers to a unique word type, since all token frequency information is disregarded. Not surprisingly, there are words in the generated stories that are not contained in the vocabulary for the narrative sense relations dataset. Line 1 reports the total number of unique words in each set of stories after filtering/lemmatization (raw words), while Line 2 shows the proportion of these words that also appear in the vocabulary for the narrative sense relations dataset (recognized words). There are many unrecognized words in the GPT-2 and HUMAN stories, but these stories also contain many more recognized words as well (and it should be considered that several of the unrecognized words occur very rarely in these stories, which is not conveyed in the table). With having smaller word sets, the majority of the words in the HIERARCHICAL and L2W stories are recognized. All subsequent lines in the table pertain to the recognized words. Line 3 reports the mean number of stories that each word generated by that model appears in. This is an indication of lexical diversity, where higher numbers indicate higher redundancy of words across stories generated by that model. For example, each of the 4,617 words among HIERARCHICAL stories occurs in 62.56 HIERARCHICAL stories on average. Consistent with the GPT-2 and HUMAN stories featuring a much broader set of words, these stories are much more diversified in their word selection. The CREATIVEHELP and L2W stories have more words that appear redundantly across stories, with less redundancy in the HIERARCHICAL stories. Line 4 reports the mean number of unique words per story. The CREATIVEHELP, GPT-2, and HUMAN stories have far more unique words than the HIERARCHICAL and L2W stories. This finding is qualitatively reflected in Table 1, where the examples for the latter models contain many repeated words. Lines 5 and 6 show the mean number of unique word pairs per story (where both words are recognized) and the proportion of these that also show up in the narrative sense relations dataset. Naturally, there are fewer word pairs for the HIERARCHICAL and L2W stories given that they contain fewer words overall. There is more coverage for these word pairs in the narrative sense relations dataset. Most of the CREATIVEHELP pairs are also recognized from this dataset. In contrast, the GPT-2 and HUMAN stories contain several word pairs that have not been observed in this dataset.

4.2 Distribution of Word Relations

We examined the word pairs for each model according to their PMI scores in the narrative sense relations dataset. All unseen word pairs were assigned the lowest score of the pairs in the dataset (-17.25). Table 3 illustrates the top 10 word pairs with the highest PMI in each of the stories from Table 1.

Figure 1 plots the binned distribution (binned using the Freedman-Diaconis rule (Freedman and Diaconis, 1981)) of PMI scores for all word pairs in the generated stories for each model. The y-axis represents the total number of word pairs with scores in the corresponding bin. The blue area of the graph includes all pairs, while the orange area represents the distribution when only the 100 highest-scoring pairs in each story are considered. The median of each distribution is indicated by the lines of the corresponding color. The plots convey some of the information in Table 2, particularly with regard to the HIERARCHICAL and L2W models generating fewer unique words and thus fewer pairs overall. These particular models also have a much more narrow score distribution, and a higher median score overall relative to the

CREATIVEHELP	GPT-2	HIERARCHICAL	L2W	HUMAN
((chamber, leapt), -9.26)	((ship, voyage), -9.0)	((pause, step), -10.92)	((chair, shuffling), -9.04)	((scent, sniff), -8.59)
((afraid, fear), -10.47)	((inhabitant, planet), -9.29)	((pause, suit), -11.3)	((pitch, swing), -9.66)	((generator, power), -8.97)
((’d, shrug), -10.57)	((inhabit, inhabitant), -9.51)	((follow, human), -11.52)	((breathing, sound), -10.22)	((instinctively, silence), -9.45)
((chamber, throat), -10.59)	((bounty, planet), -9.55)	((come, pause), -11.55)	((silent, strew), -10.24)	((gasp, grunt), -9.53)
((’d, say), -10.65)	((inhabit, ritual), -9.68)	((helmet, man), -11.67)	((floor, strew), -10.27)	((faint, instinctively), -9.54)
((leapt, seek), -10.75)	((planet, ship), -9.69)	((old, young), -11.68)	((chair, sit), -10.47)	((mouth, muffle), -9.62)
((believe, united), -10.78)	((tyrant, voyage), -9.82)	((follow, young), -11.71)	((nod, pitch), -10.61)	((faint, scent), -9.75)
((chamber, room), -10.79)	((inhabitant, tyrant), -10.04)	((pause, screen), -11.71)	((debris, floor), -10.63)	((breathing, darkness), -9.75)
((shrug, strong), -10.83)	((consider, preview), -10.1)	((follow, man), -11.85)	((nod, shut), -10.69)	((direction, scent), -9.89)
((chamber, heart), -10.84)	((sacrifice, tyrant), -10.21)	((human, large), -11.86)	((breathing, safe), -10.72)	((faint, tremble), -9.92)

Table 3: Highest-scoring word pairs for each story from the example in Table 1

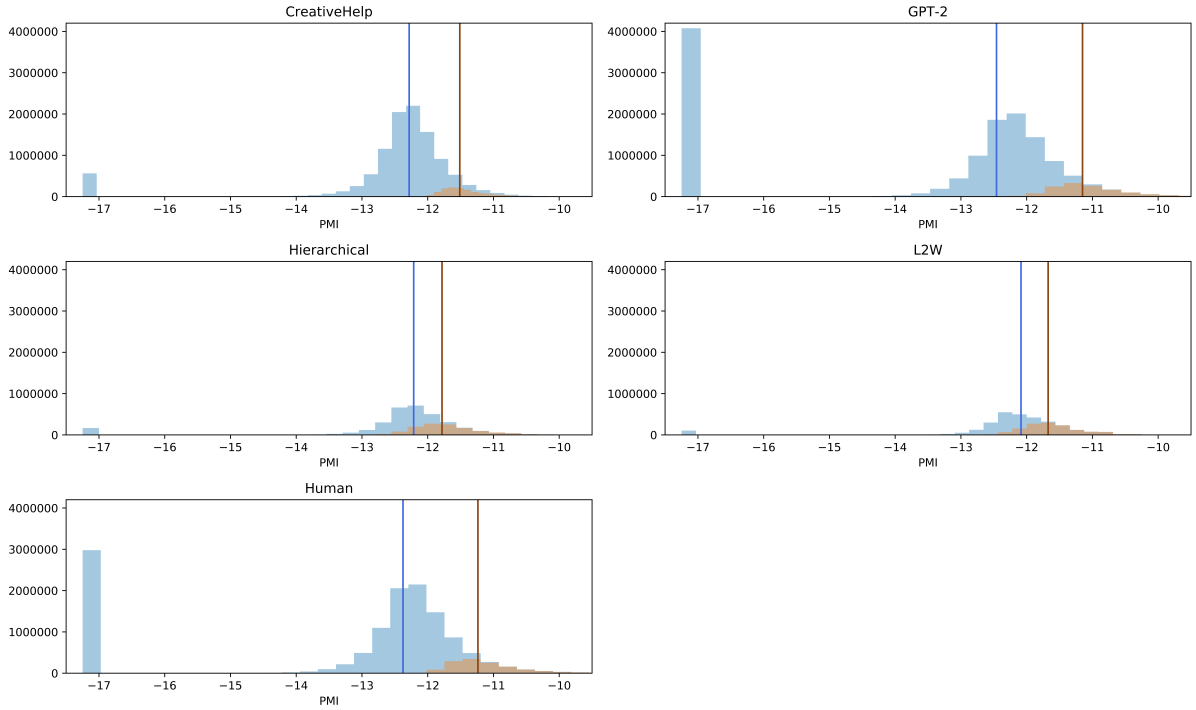

Figure 1: Distribution of word pair scores in generated stories for each model

other models. In contrast, the scores of the CREATIVEHELP, GPT-2, and HUMAN pairs are distributed across a wider range. The large number of pairs not observed in the narrative sense relations datasets for the GPT-2 and HUMAN stories is represented by the tall blue bar on the far left side of each of these plots (since the score of these pairs is set to the lowest PMI score in the narrative sense relations dataset). This causes the median of the full distribution for these models to be much lower. However, these stories also have many more pairs with higher PMI scores, signified by the large gap between the full distribution and the top-100 distribution, where the medians of the latter are much higher for these models. Thus, we can summarize that the HIERARCHICAL and L2W models tend to consistently generate moderately strong relations, but the GPT-2 and HUMAN stories are more likely to contain very strong lexical relations. Interestingly, the median score of the GPT-2 pairs among the top-100 distribution (-11.15) is actually slightly higher than the corresponding HUMAN median (-11.24).

4.3 Distinguishing Narrative Sense Relations

Figure 1 reveals differences in the distribution of lexical relation scores for each model, but the discrepancies in their individual word distributions make it difficult to draw conclusions about how much narrative sense is produced by each model. To try to further interpret the differences between the models, we designed a prediction task that tests whether the lexical relations in each story can be distinguished from spurious relations. In particular, for each generated story represented as a set of words, we artificially created a new story with the same number of words, where the words were randomly sampled from the set of all stories

CREATIVEHELP	GPT-2	HIERARCHICAL	L2W	HUMAN
5571 (41.4%)	6574 (48.9%)	6661 (49.5%)	6104 (45.4%)	7355 (54.7%)

Table 4: Total number of stories (among 13,453) exceeding narrative sense threshold for each model

generated by the corresponding model. Thus, the scores of any relations that emerge in these samples are accounted for by overall word frequency alone. Another way to think about this test is that it determines how easy it is to distinguish relations that occur within a given story to those that occur across different stories. We compared the distribution of scores for the original story to the distribution for the random story using the Wilcoxon rank-sum statistic (Wilcoxon, 1945), which evaluates the difference between two distributions. If this test indicated the original word pair scores were on average higher than the random scores (at a level of statistical significance $p < 0.10$), we assigned the original story a point indicating it exceeded the narrative sense threshold. Exceeding this threshold signifies confidence that the lexical relations between the words in the story 'make sense'. In this scheme, stories with high narrative sense should contain much higher scoring word pairs than would be expected to appear from random combinations of the same words. If there are never differences between these distributions, it suggests that the generated word relations occur largely by chance. Thus, stories with more distinct narrative sense relations should more often exceed the narrative sense threshold.

Table 4 shows the results of this analysis. Note that the narrative sense threshold is quite conservative due to the requirement that the difference between the original and random pairs be statistically significant. Thus, the absolute number of stories that exceed the threshold is low for all models, but we are only concerned with their relative difference. The CREATIVEHELP stories have the least distinct narrative sense relations, which is notable given that their median word pair score is higher than that of the HIERARCHICAL and L2W stories. This suggests many of the relations generated by CREATIVEHELP appear simply due to the number of combinations of words in these stories (since more combinations yields more opportunities to find high-scoring relations in the narrative sense relations dataset). As expected, the HUMAN stories exceed the narrative sense threshold the most often, meaning that their lexical relations are the least likely to be predicted by just the overall frequency of their words. This result also distinguishes the HUMAN stories from the GPT-2 stories, which otherwise show similar score distributions in Figure 1. While the GPT-2 model produces many strong narrative sense relations overall, from the result in Table 4 we can conclude that a single GPT-2 story tends to have less narrative sense than a HUMAN story when their respective overall word distributions are taken into account. Moreover, the HIERARCHICAL stories also demonstrate stronger narrative sense relations than the GPT-2 stories according to this analysis, even though the former produces fewer high-scoring pairs overall.

5 Conclusion

We demonstrated an analysis of lexical relations in generated stories with an emphasis on identifying 'narrative sense' relations that contribute to perceived story coherence. This work is intended to support the development of automated metrics that detect whether a generated text is sensible, in order to reduce the expense of exclusively relying on human judgment for this type of evaluation. We extracted word relations in the generated output of four published story generation systems that have not previously been compared on the same set of story inputs. We discovered interesting differences in the relations produced by each model, and presented a way to characterize these relations according to how well they can be discriminated from relations that appear by chance. These results indicate that the human-authored stories feature strong narrative sense relations that distinguish them from the generated stories. Differences among the generated models are also apparent. As future work, we can reproduce this analysis using a different narrative sense relations dataset to better determine the impact of this dataset on exposing these differences.

In this work, the narrative sense of a lexical relation is vouched for by its repeated appearance in other stories, so the focus is on rewarding models for producing these relations. An alternative analysis could instead look for relations that violate

some aspect of commonsense knowledge. This would shift the focus of the analysis to penalizing models for producing relations that detract from the coherence of the story. However, it is also important to point out that an ideal story generation system would model human creativity in producing content that has not been observed in any existing story. Presumably many of the previously unseen pairs appearing in the human-authored stories are reflective of this creativity while also not necessarily violating commonsense. Future work should examine how to evaluate the capacity of systems to induce novel lexical relations that support story coherence.

References

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *46th Annual Meeting of the Association of Computational Linguistics*, pages 789–797.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 889–898.

David Freedman and Persi Diaconis. 1981. On the histogram as a density estimator: L 2 theory. *Probability theory and related fields*, 57(4):453–476.

Andrew Gordon, Cosmin Adrian Bejan, and Kenji Sagae. 2011. Commonsense Causal Reasoning Using Millions of Personal Stories. *Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, pages 1180–1185.

F. Hill, A. Bordes, S. Chopra, and J. Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the Association for Computational Linguistics*.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Dan Jurafsky, Victor Chahuneau, Bryan Routledge, and Noah Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4).

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *15th International Conference on Principles of Knowledge Representation and Reasoning (KR-2016)*.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pages 839–849.

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

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France. Association for Computational Linguistics.

Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, page 21. ACM.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manasvi Sagarkar, John Wieting, Lifu Tu, and Kevin Gimpel. 2018. Quality signals in generated stories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 192–202. Association for Computational Linguistics.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 201712th International Conference on Computational SemanticsShort papers*.

Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, KDD '10, pages 623–632, New York, NY, USA. ACM.

Jukka Toivanen, Oskar Gross, and Hannu Toivonen. 2014. The officer is taller than you, who race yourself! using document specific word associations in poetry generation.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Aubrie Woods. 2016. Exploiting linguistic features for sentence completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 438–442, Berlin, Germany. Association for Computational Linguistics.