Entity Skeletons for Visual Storytelling

Khyathi Raghavi Chandu* Ruo-Ping Dong* Alan W Black

Language Technologies Institute, Carnegie Mellon University
{kchandu, awb}@cs.cmu.edu, ruopingd@alumni.cmu.edu

Abstract

We are enveloped by stories of visual interpretations in our everyday lives. Story narration often comprises of two stages, which are, forming a central mind map of entities and then weaving a story around them. In this paper, we address these two stages of introducing the right entities at seemingly reasonable junctures and also referring them coherently in the context of visual storytelling. The building blocks of the central mind map, also known as entity skeleton are entity chains including nominal and coreference expressions. We establish a strong baseline for skeleton informed generation and propose a glocal hierarchical attention model that attends to the skeleton both at the sentence (local) and the story (global) levels. We observe that our proposed models outperform the baseline in terms of automatic evaluation metric, METEOR. We also conduct human evaluation from which it is concluded that the visual stories generated by our model are preferred 82% of the times.

1. Introduction

> *"You're never going to kill storytelling because it's built in the human plan. We come with it." - Margaret Atwood*

Storytelling in the age of artificial intelligence is not supposed to be a built-in capability of humans alone. With the advancements in interacting with virtual agents, we are moving towards sharing this creative and coherent ability with machines as well. The evolution of storytelling spans from primordial ways of cave paintings and scriptures to contemporary ways of using multiple modalities, such as visual, audio and textual narratives. We address narrating a story from visual input, also known as *visual story telling* [12]. Generating textual stories from a sequence of images has gained traction very recently [8, 11, 13, 15, 21, 3]. Stories can be perceived as revolving around characters [17], events/actions [22, 19, 21], or theme [7]. Emulating a naturally generated story requires equipping machines to learn where to introduce entities, and more importantly, how to refer to them henceforth. The main task addressed in this paper is to introduce entities similar to how humans do and more importantly, referring them appropriately in subsequent usage for stories from images. We perform this in two phases: (1) Entity Skeleton Extraction, and (2) Skeleton Informed Generation. Here, a skeleton is defined as a simple template comprising of the entities and their referring expressions extracted using off-the-shelf NLP tools. This entity skeleton is also represented in different levels of abstractions to compose a generalized frame to weave the story. The entities can be reliably extracted from image captions which when used in conjunction with images result in a better coherent story.

2. Related Work

Visual Storytelling: [12] proposed visual storytelling dataset, comprising of sequences of story-like images with corresponding textual descriptions in isolation and stories in sequences. [13] proposed a seq2seq framework and [24] proposed late fusion techniques to address this task. We derive motivation from these techniques to introduce entities and references as skeletons. [20, 14] explored the task of generating a sequence of sentences for an image stream. [1] and [14] addressed syntactic and semantic coherence while our work is focused on content relevance.

Schema based generation: [9] was one of the initial works delving into how entities and their referring expressions are used in a discourse. Several efforts for narrative generation tasks have spawned from introducing a schema or a skeleton. While [17, 4, 2] explored using event representations, [5, 6, 26] proposed hierarchical frameworks for topically guided story. [25] used reinforcement learning to first generate skeleton (the most critical phrases) and then expand the skeleton to a complete sentence. Our work falls along the lines of generating a story from visual input based on entity skeletons.

3. Data Description

The visual storytelling is proposed as a multimodal grounded sequential generation dataset [12]. Formally, the dataset comprises of visual stories $\boldsymbol{S} = \{\boldsymbol{S_1}, \ldots, \boldsymbol{S_n}\}$. Each story in the dataset consists

Sentences from SIS	Surface	Nominalized	Abstract	Surface	Nominalized	Abstract
The cake was amazing for this event!	None	[0, 0]	None	event	[1, 0]	other
The bride and groom were so happy.	The bride and groom	[1, 0]	person	None	[0, 0]	None
They kissed with such passion and force.	They	[1, 1]	person	None	[0, 0]	None
When their son arrived, he was already sleeping.	their	[1, 1]	person	None	[0, 0]	None
After the event, I took pictures of the guests.	None	[0, 0]	None	event	[1, 0]	other

Table 1. Examples of three forms of Entity-Coreference Schema Representation

of a sequence of five story-like images, along with descriptions-in-isolation (DII) and stories-in-sequences (SIS). Each story can be formally represented as $\boldsymbol{S_i} = \{(\boldsymbol{I}_i^{(1)}, \boldsymbol{x}_i^{(1)}, \boldsymbol{y}_i^{(1)}), \ldots, (\boldsymbol{I}_i^{(5)}, \boldsymbol{x}_i^{(5)}, \boldsymbol{y}_i^{(5)})\}$, where $\boldsymbol{I}_i^{(j)}$, $\boldsymbol{x}_i^{(j)}$ and $\boldsymbol{y}_i^{(j)}$ are each image, single sentence in DII and single sentence in SIS respectively, and *i* refers to the *ith* example story. SIS and DII are supposed to be associated with each image, shown in Table 2. For the images for which the DII are absent, we use a pre-trained image captioning model [23] to make the dataset complete for our use case.

	Train	Val	Test
# Stories	40,155	4,990	5,055
# Images	200,775	24,950	25,275
# with no DII	40,876	4,973	5,195

Table 2. Details of the Dataset

4. Model Description

Our approach of using entity skeletons to generate a coherent visual story is divided into two phases: (1) Entity Skeleton Extraction, and (2) Skeleton Informed Generation. We will be releasing the codebase.

4.1. Entity Skeleton Extraction

The task is to introduce the characters in right times and refer to them appropriately henceforth. This means that we not only target the head mention of an entity but also cater to the corresponding appropriate coreference expressions. We define the skeleton as a linear chain of entities and their corresponding referring expressions. We first extract the coreference chains from the textual stories that are made up of SIS in the training data. This is done by using version 3.7.0 of Stanford CoreNLP toolkit [16]. These three ways of representing skeletons are described in detail next. An example of the three forms are depicted in Table 1

1. Surface form Coreference Chains: The skeleton for each story is represented as $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_5\}$, where $\boldsymbol{c}_j$ is the coreference word in *jth* sentence. The skeleton word is *None* when there is no word corresponding to that coreference chain in that sentence.

2. Nominalized Coreference Chains: This form disintegrates the properties of presence and absence of the entity words and whether the word is present in the noun or the pronoun form. The skeleton for each story is represented as $\{[h, p]_1, \ldots, [h, p]_5\}$. Here, $h \in \{0,1\}$, is a binary variable indicating if there is a coreference mention, i.e 1 if there is a mention in the skeleton chain and 0 if it is None. Similarly, $p \in \{0,1\}$ is a binary variable indicating that the word is head mention i.e, the word is in the noun form if it is 0 and pronoun form if it is 1.

Models	Entity Forms	Meteor	Dist.	Avg # entities
Baseline	None	27.93	1.02	0.4971
+Entities	Surface	27.66	1.02	0.5014
MTG (α(0.5))	Surface	27.44	1.02	0.9554
MTG (α(0.4))	Surface	27.59	1.02	1.1013
MTG (α(0.2))	Surface	27.54	1.01	0.9989
MTG (α(0.5))	Nominal	**30.52**	1.12	0.5545
MTG (α(0.5))	Abstract	27.67	1.01	0.5115
Glocal Attention	Surface	**28.93**	1.01	**0.8963**

Table 3. Automatic Evaluation of Story Generation Models

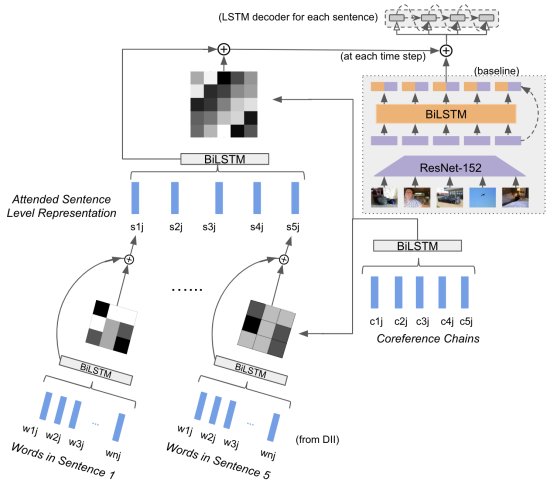

Figure 1. Architecture of Glocal Hierarchical Attention on Entity skeleton coreference chains to perform Visual Storytelling

3. Abstract Coreference Chains: Thi s form represents entities in abstract categories such as *person, object*, *location* etc., We use Wordnet [18] to derive these properties.

4.2. Schema Informed Generation

In this section, we describe a baseline and a second baseline that accesses the skeleton information for fair comparison. We then move onto discussing two models that incorporate the three forms of entity skeletons.

For simplicity in formal representation, we use the following notations. t and τ indicates the t^{th} step or sentence in a story and τ^{th} word within the sentence respectively. I_t, x_t, y_t, represent image, DII, SIS for a particular time step. k_t is the skeleton coreference element for that par-

ticular sentence. Here k can take any of the three forms of coreference chains discussed previously, which is word itself (surface form) or a pair of binary digits (nominalization) or noun properties (abstract). Note that k is not used in this baseline model.

1. Baseline Model: Our baseline model has an encoder-decoder framework that is based on the best performing model in the Visual Story Telling challenge in 2018 [13] that attained better scores on human evaluation metrics. Image features are extracted from the penultimate layer of ResNet-152 [10]. The encoder part of the model is represented as the following which comprises of two steps of deriving the local context features l^t and the hidden state of the t^{th} timestep of the BiLSTM that gives the global context.

$\boldsymbol{l}_t = \textit{ResNet}(\boldsymbol{I}_t)$

$\boldsymbol{g}_t = \textit{Bi-LSTM}([l_1, l_2 ... l_5]_t)$

The latent representation obtained from this encoder is the *glocal* representation $[\boldsymbol{l}_t, \boldsymbol{g}_t]$, where [..] represents augmentation of the features. This *glocal* vector is used to decode the sentence word by word. The generated words in a sentence from the decoder $\hat{\boldsymbol{w}}_{\boldsymbol{t}}$ is obtained from each of the words $\hat{\boldsymbol{w}}^{\boldsymbol{\tau}}$ that are the outputs that are also conditioned on the generated words so far $\hat{\boldsymbol{w}}_{\boldsymbol{t}}^{<\boldsymbol{\tau}}$ with τ^{th} word in the sentence being generated at the current step. The baseline model is depicted in the right portion of the Figure 1.

$\hat{\boldsymbol{w}}_t \sim \prod_\tau Pr(\hat{\boldsymbol{w}}_t^\tau | \hat{\boldsymbol{w}}_t^{<\tau}, [\boldsymbol{l}_t, \boldsymbol{g}_t])$

2. Skeleton Informed Baseline Model: For a fair comparison with our proposed approaches, we condition the decoder on not only the glocal features and the words generated so far, but also the surface form of the words.

$\hat{\boldsymbol{w}}_t \sim \prod_\tau Pr(\hat{\boldsymbol{w}}_t^\tau | \hat{\boldsymbol{w}}_t^{<\tau}, [\boldsymbol{l}_t, \boldsymbol{g}_t, \boldsymbol{k}_t])$

3. Multitask Story Generation Model (MTG): Incorporating the entity skeleton information directly in the decoder might affect the language model of the decoder. Instead of augmenting the model with skeleton information, we enable the model to predict the skeleton and penalize it accordingly. The main task here is the generation of the story itself and the auxiliary task is the prediction of the entity skeleton word per time step. Each of these tasks are optimized using cross entropy loss. The loss for generation of the story is $\mathbf{L}_1$ and the loss to predict the skeleton of the model is $\mathbf{L}_2$. We experimented with different weighting factors for α which are presented in Table 3.

$$\sum_{I_t, x_t, y_t \in \mathbb{S}} \alpha \mathbf{L}_1(I_t, y_t) + (1-\alpha)\mathbf{L}_2(I_t, y_t, k_t)$$

Note that we do not use k as a part of the encoder even in this model but only use them to penalize the model when the decoded sentence does not contain skeleton similar to k.

4. Glocal Hierarchical Attention: This multitasking model does not explicitly capture the relationship or focus on the words within a sentence or across the five sentences with respect to the skeleton in consideration. Hence, we went one step further to identify the correlation between the coreference skeleton with different levels including within a sentence (i.e, at word level) and across sentences (i.e, at sentence level). We use attention mechanism to represent these correlations. Figure 1 depicts the entire glocal hierarchical attention model with the encoder decoder framework on the right and the two stages of attention on the left.

Local Attention: This is to identify the correlation between words in each sentence to the coreference skeleton words. Since we use the skeleton words as they appear to attend to the words in DII, we use the surface form notation in this model. As we have seen, the surface form skeleton is represented as $C = \{c_1, c_2.., c_5\}$. The vocabulary of these surface form skeleton words is limited to 50 words in the implementation. The surface skeleton form C is passed through a Bi-LSTM resulting in hidden state H_k which is of 1024 dimensions. This hidden state is used to perform attention on the input words of DII for each image. Note here that the skeleton words for coreference chains are extracted from SIS (i.e, from $\{y_1, y_2.., y_5\}$), from which the hidden state is extracted, which is used to perform attention on the individual captions (DII i.e, $\{x_1, x_2.., x_5\}$). The skeleton remains the same for all the sentences. The skeleton form is passed through a Bi-LSTM resulting in $H_k \in \mathbb{R}^{k \times 2h}$, where hidden dimension of the Bi-LSTM is h. Each x in the story (with n words in a batch) is passed through a Bi-LSTM with a hidden dimension of h, resulting in $H_w \in \mathbb{R}^{5 \times n \times 2h}$. This then undergoes a non-linear transformation. Attention map for the word level is obtained by performing a batch matrix multiplication (represented by $\otimes$) between the hidden states of the words in a sentence and the hidden states of the entity skeleton. In order to scale the numbers in probability terms, we apply a softmax across the words of the sentence. Essentially, this indicates the contribution of each word in the sentence towards the entity skeleton that is present as a query in attention. This is the *local attention* $A_w \in \mathbb{R}^{5 \times n \times k}$ pertaining to a sentence in the story.

$$A_w = softmax(H_w \otimes H_k)$$

Glocal Attention: We then perform *global attention*, which is at the entire story level. For this, the locally attended representation of each sentence is then augmented with the output of the Bi-LSTM that takes in DII. The attended representation for each of the k words are concatenated and projected through a linear layer into 256 dimensions (P_w). This goes in as sentence representation for each of the s_{ij} (where i is the index of the sentence in the story and j corresponds to the story example) as shown in Figure 1. The word representations at each time step are obtained by augmenting the corresponding vectors from H_w and P_w. These form our new sentence embeddings. These sentence embed-

Models						Phenomena
SIS	we went to the stadium early to eat and sight see before the game .	the view was incredible . you could see the entire city .	we got to our seats , and could n't believe how close to the field they were .	we could see all the action .	once the national anthem was sung , and the first pitch was thrown , the excitement began . it was a great game !	
Baseline Model	the city was a great place to visit .	i had a great time .	there were many people there .	we got to see a lot of cool things .	it was a lot of fun .	- Characters in the story are mentioned as "many people" instead of "we" (sentence 3).
Glocal Hierarchical Attention Model	we saw the **building** was packed .	i was excited to see my favorite team .	we were all excited to see the **game** .	we all got together to watch .	it was a great **game** .	+ characters ('we' and 'it') were introduced at the right time + Important entities were mentioned (building, game)

Figure 2. Qualitative Analysis

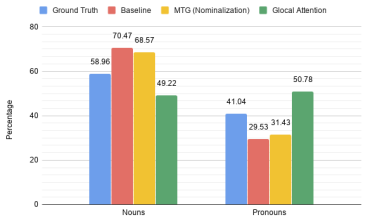

Figure 3. Percentage of Entities in the form of Nouns and Pronouns in the generated stories

dings are again passed through a Bi-LSTM to get a sentence level representation. This process is done for each sentence in the story (which are the replications as shown in the left portion of Figure 1). This results in a latent representation of the story $H_s \in \mathbb{R}^{5\times 2h}$. Along the same lines of local attention, we now compute story level hierarchical global attention to result in $A_s \in \mathbb{R}^{5\times k}$.

$$A_s = softmax([H_w, P_w] \otimes H_k)$$

The attended vectors from A_w and A_s of size *nk* and *k* respectively are concatenated in each sentence step in the decoder from the baseline model. This is shown in the top right corner of Figure 1 (although the Figure depicts concatenation for single time step).

5. Quantitative and Qualitative Analysis

We perform automatic evaluation with METEOR score for generation. The results are shown in Table 3. However, our main target is to verify whether the story adheres to the entity skeleton form that is provided. Hence we also compute the distance between the binary vectors of length 5 constructed by extracting entities in ground truth and the generated stories (*Dist.* measure). As we can see, the Euclidean Distance is not very different in each of the cases. However, we observe that the multitasking approach (MTG) is performing better with nominalization form of entity skeletons as compared to the baselines and other forms of entity skeleton representations as well. The *glocal* model described performs attention on the surface words only and hence the experiment includes only this configuration. We observe that glocal attention model outperforms the baseline model. However, there is a scope for improvement when the attention mechanism is performed on nominalized skeleton representation, which we leave for the future work.

To analyze the number of entities generated, we calculated percentages of nouns and pronouns in the ground truth and generated stories, presented in Figure 3. In the nouns section, baseline model seemed to have over-generated nouns in comparison to both of our proposed models. While MTG model also has over-generated the nouns, our glocal attention model has generated fewer nouns compared to ground truth. While the MTG model generated higher number of pronouns in comparison to the baseline, the glocal attention model seemed to have generated even higher percentage of pronouns. Despite this over-generation, glocal attention model is the closest to the number of pronouns in the ground truth stories. Coming to the diversity of entities generated in stories, we calculate the average number of distinct entities present per story for each of the models. These numbers are shown in the last column of Table 3. This number is closer to that of the ground truth for the glocal attention model assuring that there is sufficient diversity in the entity chains that are generated by this model.

Qualitative Analysis: Figure 2 presents an image sequence for a story along with the corresponding ground truth (SIS) and the generated stories. The positive and the negative phenomena observed are presented in the last column. The Glocal Hierarchical Attention Model is able to capture the skeleton words right in comparison to the baseline model.

Human Evaluation: We conduct preference testing for 20 randomly sampled stories by asking 5 subjects the following question *'preferred story from images'*. Our *glocal hierarchical attention model* is preferred 82% and 64% of the times compared to baseline model and MTG model with nominalized representation respectively.

6. Conclusion and Future Work

Automatic storytelling has been a dream since the emergence of AI.Our work is inspired from the intuition that humans form a central mindmap of a story before narrating it. In this work, this mindmap is associated with entities (such as persons, locations etc.,) to incorporate content relevance. We present our work on introducing entity and reference skeletons in the generation of a grounded story from visual input. We observe that our *MTG* and *glocal hierarchical attention* models are able to adhere to the skeleton thereby producing schema based stories with seemingly on-par and sometimes better results. These stories depict better naturalness in human evaluation. We plan on applying our methods to other forms of conditions to generate storytelling such as semantic representations, graphs and prompts.

References

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. *arXiv preprint arXiv:1606.07493*, 2016. 1

[2] Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara Martin, and Mark Riedl. Guided neural language generation for automated storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 46–55, 2019. 1

[3] Khyathi Chandu, Eric Nyberg, and Alan W Black. Storyboarding of recipes: Grounded contextual generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6040–6046, 2019. 1

[4] Elizabeth Clark, Yangfeng Ji, and Noah A Smith. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2250–2260, 2018. 1

[5] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 1

[6] Angela Fan, Mike Lewis, and Yann Dauphin. Strategies for structuring story generation. *arXiv preprint arXiv:1902.01109*, 2019. 1

[7] Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás. Story plot generation based on cbr. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 33–46. Springer, 2004. 1

[8] Diana Gonzalez-Rico and Gibran Fuentes-Pineda. Contextualize, show and tell: a neural visual storyteller. *arXiv preprint arXiv:1806.00738*, 2018. 1

[9] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Chao-Chun Hsu, Szu-Min Chen, Ming-Hsun Hsieh, and Lun-Wei Ku. Using inter-sentence diverse beam search to reduce redundancy in visual storytelling. *arXiv preprint arXiv:1805.11867*, 2018. 1

[12] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016. 1

[13] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. Glac net: Glocal attention cascading networks for multi-image cued story generation. *arXiv preprint arXiv:1805.10973*, 2018. 1, 3

[14] Yu Liu, Jianlong Fu, Tao Mei, and Chang Wen Chen. Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1

[15] Stephanie Lukin, Reginald Hobbs, and Clare Voss. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32, 2018. 1

[16] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. 2

[17] Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[18] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[19] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, 2016. 1

[20] Cesc C Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. In *Advances in neural information processing systems*, pages 73–81, 2015. 1

[21] Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, 2018. 1

[22] Elena Rishes, Stephanie M Lukin, David K Elson, and Marilyn A Walker. Generating different story tellings from semantic representations of narrative. In *International Conference on Interactive Digital Storytelling*, pages 192–204. Springer, 2013. 1

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[24] Marko Smilevski, Ilija Lalkovski, and Gjorgi Madzarov. Stories for images-in-sequence by using visual and narrative components. *arXiv preprint arXiv:1805.05622*, 2018. 1

[25] Jingjing Xu, Yi Zhang, Qi Zeng, Xuancheng Ren, Xiaoyan Cai, and Xu Sun. A skeleton-based model for promoting coherence among sentences in narrative story generation. *arXiv preprint arXiv:1808.06945*, 2018. 1

[26] Lili Yao, Nanyun Peng, Weischedel Ralph, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. *arXiv preprint arXiv:1811.05701*, 2018. 1