

TARN-VIST: Topic Aware Reinforcement Network for Visual Storytelling

Weiran Chen, Xin Li, Jiaqi Su, Guiqian Zhu, Ying Li, Yi Ji, Chunping Liu

{wrchen2023, 20224227052}@stu.suda.edu.cn, czu_lixin@163.com

School of Computer Science and Technology, Soochow University, Suzhou, China

ABSTRACT

As a cross-modal task, visual storytelling aims to generate a story for an ordered image sequence automatically. Different from the image captioning task, visual storytelling requires not only modeling the relationships between objects in the image but also mining the connections between adjacent images. Recent approaches primarily utilize either end-to-end frameworks or multi-stage frameworks to generate relevant stories, but they usually overlook latent topic information. In this paper, in order to generate a more coherent and relevant story, we propose a novel method, Topic Aware Reinforcement Network for Visual StoryTelling (TARN-VIST). In particular, we pre-extracted the topic information of stories from both visual and linguistic perspectives. Then, we apply two topic-consistent reinforcement learning rewards to identify the discrepancy between the generated story and the human-labeled story so as to refine the whole generation process. Extensive experimental results on the VIST dataset and human evaluation demonstrate that our proposed model outperforms most of the competitive models across multiple evaluation metrics.

CONTRIBUTIONS

- We first take advantage of CLIP and RAKE together to extract topic information of stories from both visual and linguistic perspectives.
- To make full use of the topic information, we design reinforcement learning rewards for topic consistency based on the extracted topic information and cosine similarity.
- Experimental results on the VIST dataset and human evaluation demonstrate that our proposed model outperforms most of the leading models on multiple evaluation metrics.

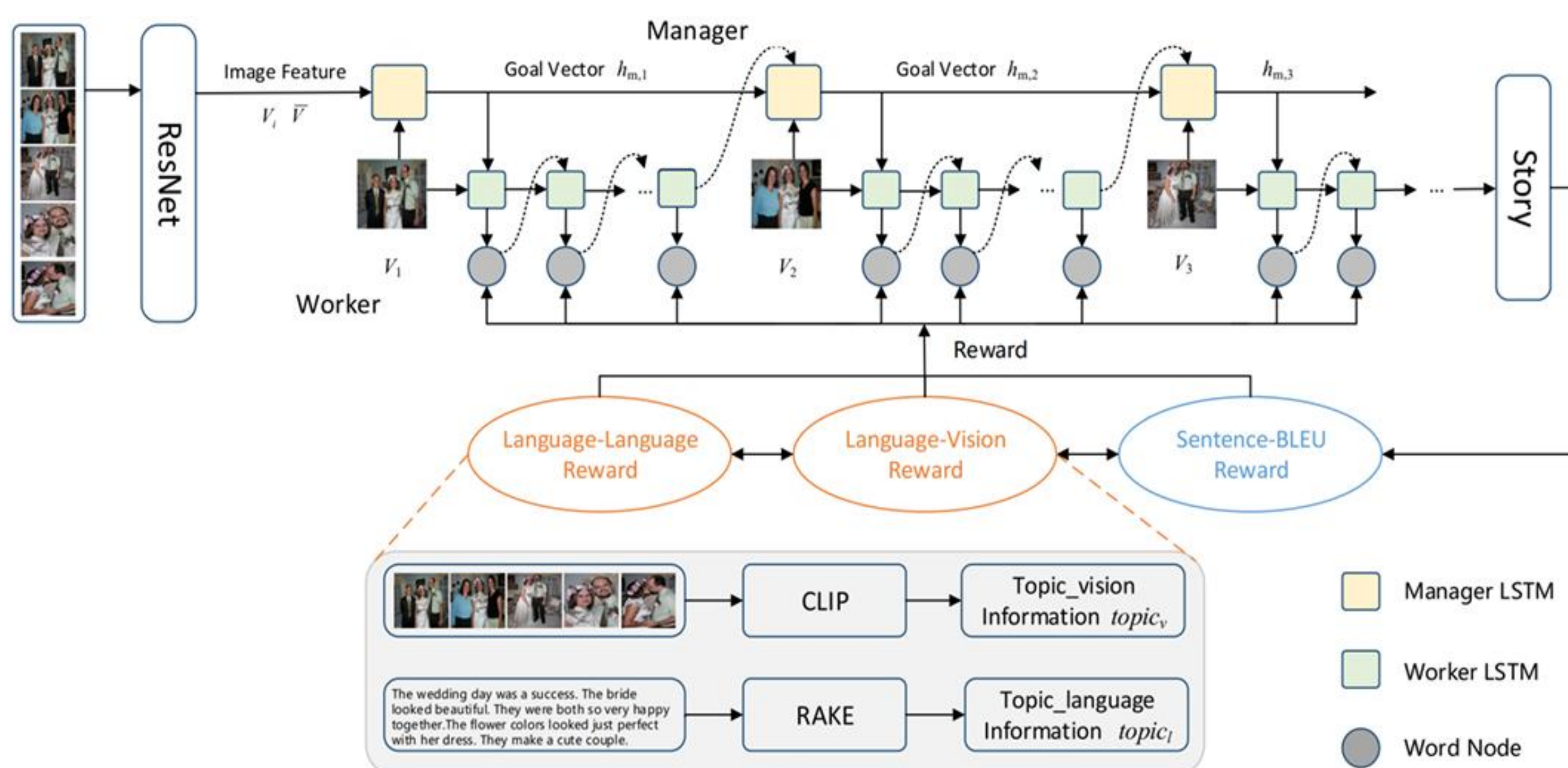


Figure 1: Overview of TARN-VIST. In our model, image features are obtained by the pre-trained ResNet and then fed into the hierarchical decoder which consists of a manager LSTM and a worker LSTM to generate a sample story. Once the candidate story is generated, the two topic consistency rewards are combined to refine the generation process. Furthermore, we also set a classical sentence-level BLEU reward to control the fluency of the generated story.

Table 1: Quantitative results on the VIST dataset for surface-level-based automatic metrics. For all these metrics, higher score means better performance.

Model	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Seq2seq	-	-	3.5	31.4	-	6.8	-
H-Attn-Rank	-	-	-	34.1	29.5	7.5	-
BARNN	-	-	-	33.3	-	-	-
SRT	43.4	21.4	5.2	12.3	-	11.4	-
XE-ss	62.3	38.2	13.7	34.8	29.7	8.7	-
AREL	63.7	39.0	14.0	35.0	29.6	9.5	8.9
HPSR	61.9	37.8	12.2	34.4	31.2	8.0	-
HSRL	-	-	12.3	35.2	30.8	10.7	7.5
SGVST	65.1	40.1	14.7	35.8	29.9	9.8	-
ReCo-RL	-	-	12.4	33.9	29.9	8.6	8.3
INet	64.4	40.1	14.7	35.6	29.6	11.0	-
IRW	66.7	41.6	15.4	35.6	29.6	11.0	-
CKAKS	-	-	12.0	35.4	30.0	10.5	-
LGMT	67.5	41.6	15.1	35.6	29.7	10.0	-
Sentistory	64.8	39.8	14.2	35.3	29.8	9.7	-
TARN-VIST	69.0	43.5	13.4	35.8	29.5	12.1	11.3

Algorithm: Visual Perspective Topic Information Extraction Process

Input: Image Sequence I with 5 images and *Candidate-concept*

Output: Topic Information

1. Initialise *Candidate-concept* $\leftarrow []$
2. /* Candidate-Concept Extraction */
3. **for** $i = 1$ to 5 **do**
4. Extract top-3 concept c_i^j from Image i with *Clarifai's* Image Dection API
5. Filter out some useless concepts and assemble concepts into sentence s_i^j
6. *Candidate-concept.append(s_i^j)*
7. **end for**
8. /* Image Encoding */
9. **for** $i = 1$ to 5 **do**
10. Image-feature $i = \text{CLIP-Image-Encoder}(\text{Image } i)$
11. **end for**
12. Image-mean-feature = mean(Image-feature i)
13. /* Text Encoding */
14. Text-features = CLIP-Text-Encoder(*Candidate-concept*)
15. /* Similarity Calculation */
16. Similarity = Image-mean-feature @ Text-features
17. Topic Information = Similarity[0].topk(1)
18. return Topic Information

Table 2: Quantitative results on the VIST dataset for semantic understanding evaluation metric. For all these metrics, higher score means better performance.

Method	BERTScore	TARN-VIST	Tie
KE-VIST (No KG)	28.25	17.21	43.63
KE-VIST (With Open IE)	29.12	17.93	46.85
KE-VIST (With VG)	29.16	18.03	47.54
PR-VIST	27.64	18.09	48.92
TARN-VIST	30.47	18.51	49.43

Table 3: Human Pairwise Evaluation between TARN-VIST and other models. For each pairwise comparison, each of three columns stands for the percentage of volunteers that prefer this story to the other one, and consider both stories are of equal quality.

Aspect	KE-VIST	TARN-VIST	Tie	PR-VIST	TARN-VIST	Tie
Relevance	19%	67%	14%	30%	60%	10%
Coherence	22%	50%	28%	18%	62%	20%
Information Richness	28%	64%	8%	33%	50%	17%

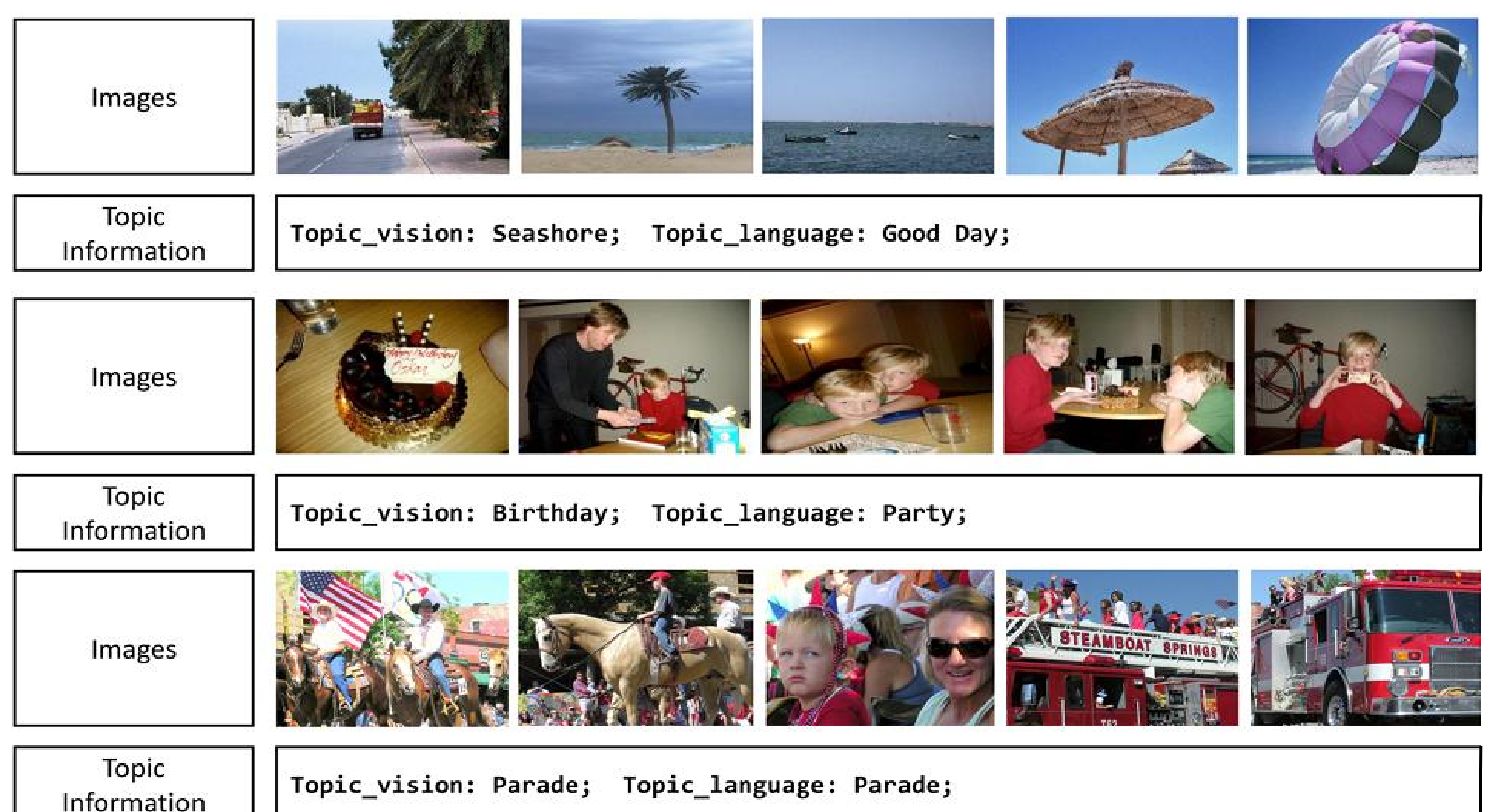


Figure 2: Examples of extracted topic information.



Figure 3: Example story generated from TARN-VIST and competitive baselines.