# Toward a Better Story End: Collecting Human Evaluation with Reasons

**Yusuke Mori**[1] **Hiroaki Yamane**[2,1] **Yusuke Mukuta**[1,2] **Tatsuya Harada**[1,2]
[1]The University of Tokyo, [2]RIKEN
{mori, mukuta, harada}@mi.t.u-tokyo.ac.jp
hiroaki.yamane@riken.jp

## Abstract

Creativity is an essential element of human nature used for many activities, such as telling a story. Based on human creativity, researchers have attempted to teach a computer to generate stories automatically or support this creative process. In this study, we undertake the task of story ending generation. This is a relatively new task, in which the last sentence of a given incomplete story is automatically generated. This is challenging because, in order to predict an appropriate ending, the generation method should comprehend the context of events. Despite the importance of this task, no clear evaluation metric has been established thus far; hence, it has remained an open problem. Therefore, we study the various elements involved in evaluating an automatic method for generating story endings. First, we introduce a baseline hierarchical sequence-to-sequence method for story ending generation. Then, we conduct a pairwise comparison against human-written endings, in which annotators choose the preferable ending. In addition to a quantitative evaluation, we conduct a qualitative evaluation by asking annotators to specify the reason for their choice. From the collected reasons, we discuss what elements the evaluation should focus on, to thereby propose effective metrics for the task.

## 1 Introduction

Creativity is vital to human nature, and storytelling is among the most important representations of human creativity. Humans use stories for entertainment and practical purposes, such as teaching lessons and creating advertisements. Stories are deeply rooted in our lives.

In computer science, understanding how humans read and create a story, and imitating these activities with a computer, is a major challenge. Mostafazadeh et al. (2016) proposed *Story Cloze Test* (SCT) as a reading comprehension task and released a large-scale corpus *ROCStories*. SCT presents four sentences, where the last sentence is excluded from a story comprising five sentences. A system must select an appropriate sentence from two choices that complement the missing 5th sentence. Among the two options is "right ending", i.e., the appropriate one to complete the story, and the other is "wrong ending".

Herein, we consider *story ending generation* (SEG) (Guan et al., 2019; Li et al., 2018; Zhao et al., 2018). This is a relatively new task inspired by SCT, and it is designed to be generation-oriented. In SEG, the last sentence of a given incomplete story is generated automatically. This is challenging because the system should comprehend the context to generate an appropriate ending.

Despite the importance of this task, no clear evaluation metric has been established thus far. To serve as a reference for future proposals of the evaluation metrics, we conduct human evaluations and study the various elements involved in evaluating an automatic SEG method.

The main contributions of this paper are:

- In order to show how well a baseline method performs and what drawbacks it has for SEG, we conducted a pairwise comparison against human-written right endings.
- Besides a quantitative evaluation, we conducted a qualitative evaluation by asking annotators to specify the reason for their choice. From the collected reasons, we explored the elements that the evaluation should focus on, to thus propose effective metrics for SEG.

## 2 Related Work

Automatic evaluation metrics that measure word matching are not effective in text generation, espe-



*Proceedings of The 12th International Conference on Natural Language Generation*, pages 383–390,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. 

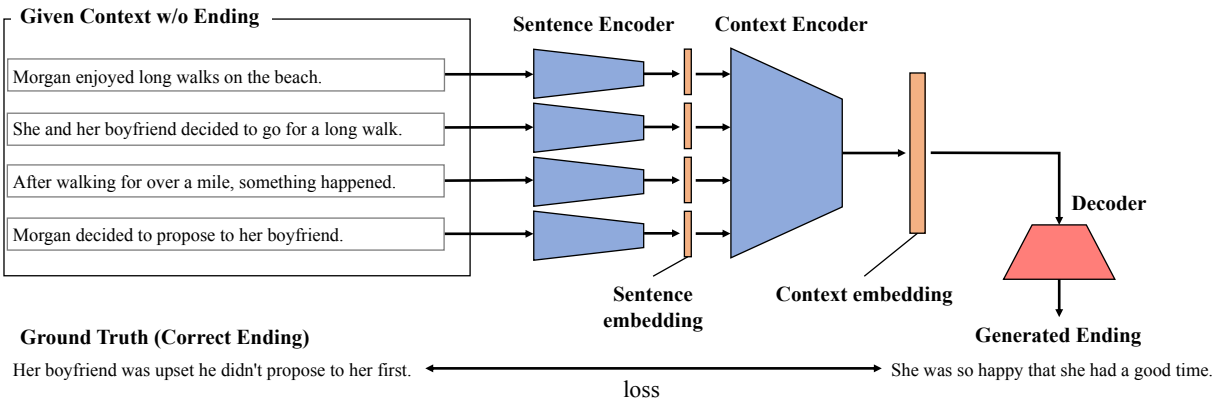


Figure 1: Given stories where the last sentence has been excluded, a method is required to generate an appropriate ending to complete the story. Our baseline method has two steps: sentence encoder and context encoder. The first encoder processes each sentence and generates corresponding sentence embeddings. These sentence embeddings are input to the second encoder, which calculates a representation of the context. The recurrent neural network (RNN) decoder receives the context embedding and generates a sentence to complete the story.

cially in dialog generation (Liu et al., 2016). Further, in story generation, it seems difficult to evaluate text generation methods with conventional automatic evaluation metrics.

As SEG is a relatively new task, metrics for human evaluation have also not been established. Zhao et al. (2018) defined two criteria, *Consistency* and *Readability*, to implement human evaluation. For each criterion, human assessors rated endings on a scale of 0 to 5. Li et al. (2018) assigned four levels to each ending (Bad (0), Relevant (1), Good (2), and Perfect (3)). Three judgement criteria were provided to annotators: *Grammar and Fluency*, *Context Relevance*, and *Logic Consistency*. They also conducted a direct comparison of the story endings generated by their baseline and their proposed approach. Guan et al. (2019) defined two metrics, *Grammar* and *Logicality*, for human evaluation. For each metric, the score 0/1/2 was applied.

In order to measure the distance from the goal of "a system writing story endings like humans", it is useful to compare the generated endings directly with human-written endings. Therefore, we conducted a pairwise comparison against human-written "right endings". To show what elements of stories humans focus on, we conducted a qualitative evaluation by asking annotators to specify the reasons for their choices.

## 3 Baseline Method for SEG

We define $S = \{s_1, s_2, ..., s_n\}$ as a story consisting of $n$ sentences. In SEG, $S' = \{s_1, s_2, ..., s_{n-1}\}$ is given as an input. Then, a method is required to generate an appropriate ending $s_n$. We refer to $S'$ as "context".

A hierarchical approach is useful for generating a long text (Liu et al., 2018), and also effective in story generation (Fan et al., 2018; Ravi et al., 2018). Using the "sequence-to-sequence" (Seq2seq) model (Sutskever et al., 2014) as a point of departure, we introduce a baseline method that handles input text hierarchically. This refers to the conventional method using a hierarchical structure for document modeling (Li et al., 2015) and query suggestion (Sordoni et al., 2015).

To be more precise, we use a two-step encoder. The first encoder receives $\{s_1, s_2, ..., s_{n-1}\}$ as a word-level input and outputs the sentence embeddings $\{v_1, v_2, ..., v_{n-1}\}$. Then, the second encoder receives the sentence embeddings as a sentence-level input and generates a distributed representation of the entire context. We named the first encoder "sentence encoder", and the second encoder "context encoder". We refer to this method as "Hierarchical Seq2seq" (H-Seq2seq). An overview of the method is shown in Figure 1.

**Sentence Encoder:** As a sentence encoder, we apply the pre-trained "InferSent" model, a bi-directional long-short term memory (Bi-LSTM) network with max pooling, trained with a natural language inference task (Conneau et al., 2017). InferSent was devised as a supervised universal sentence embedding model and demonstrated good performance with various tasks.

**Context Encoder:** Using the sentence embeddings obtained with the sentence encoder, we applied another embedding layer for context em-

bedding $v_{context}$ for the entire input sequence $S'$. We use a gated recurrent unit (GRU) (Cho et al., 2014), for the context encoder to consider the sentences as a time series. A batch normalization layer (Ioffe and Szegedy, 2015) followed.

Then, we input $v_{context}$ to the RNN decoder.

Compared with tasks like translation, current datasets for story generation are relatively small. We believe that techniques for avoiding overfitting become more effective in such a situation. We use "word dropout", which drops words from input sentences according to a Bernoulli distribution (Iyyer et al., 2015). As an association from word dropout, we also introduced dropout at the sentence-level. When obtaining sentence embeddings, sentence-level dropout drops some elements randomly according to a given probability ratio and scales the remaining elements.

# 4 Experiment

## 4.1 Methods

**H-Seq2seq** As explained in Section 3, pre-trained InferSent was applied as the sentence encoder.

**Seq2seq** As a particularly simple method, we used basic Seq2seq for comparison. To examine the strength of the hierarchical approach, non-hierarchical basic Seq2seq is useful. The series of input words was handled collectively without considering sentence-level information.

**Human Right Ending** We used the human-written "right ending" in SCT as the ground truth. Two candidates in SCT are written by a person that did not write the original story (Mostafazadeh et al., 2016). Therefore, we can consider the "right ending" as the answer if SEG is solved by humans.

Note that while H-Seq2seq uses pre-trained embeddings, Seq2seq does not. For Seq2seq, we randomly initialized the word embeddings because we intended to simplify the implementation of Seq2seq. The results with and without the pre-trained embeddings should be compared for more accurate evaluation. Different parameters should also be examined. However, in story generation, it is difficult to evaluate methods with conventional automatic evaluation metrics. On the other hand, conducting all evaluations with humans is unrealistic. Therefore, we focused on investigating how much a baseline model can solve SEG and discuss how to conduct human evaluation. Although there are more sophisticated methods for SEG already proposed (Guan et al., 2019; Li et al., 2018; Zhao et al., 2018), they are beyond the scope of this study. We leave it as a future work to apply the evaluation method discussed in this paper to more advanced models.

| | |
|---|---|
| *ROCStories* (training data) | 98,161 |
| *Story Cloze* validation set, Spring 2016 | 1,871 |
| *Story Cloze* test set, Spring 2016 | 1,871 |

Table 1: The size of the dataset for our experiment.

## 4.2 Dataset

Refer to the setting of SCT competition in SemEval-2017 (Mostafazadeh et al., 2017), we used "Spring 2016 release" and "Winter 2017 release" from *ROCStories* for training and "Spring 2016 release" validation and test sets from SCT for validation and testing (Table 1).

## 4.3 Quantitative Evaluation with MTurk

As a story is created on the premise that a human will read it, evaluation by human readers is considered to be the most accurate evaluation. We conducted human evaluation with help from Amazon Mechanical Turk (MTurk) workers. We evaluated the performance of the model depending on whether the generated ending correctly considers the context and properly completes the story. MTurk workers were given a four-sentence incomplete story and two options for an ending (5th sentence), and they were asked to indicate the best ending among them. We instructed workers that the given stories originally consisted of five sentences but the 5th sentence is lost, and they are required to choose the 5th sentence to complete each story. The workers were given four choices: option A is more appropriate (A), option B is more appropriate (B), both options are equally appropriate (both A and B), and neither options are suitable (neither A nor B). For each pair, we used 200 stories from the SCT test set for comparison. Five MTurk workers evaluated each story and its corresponding candidate endings. The most popular answer among the five workers was considered as agreement among the workers. The results are shown in Table 2.

## 4.4 Qualitative Evaluation with MTurk

We conducted another experiment similar to that in Section 4.3, where workers were required to write the reason they chose the answer. We focused on comparing H-Seq2seq against humans

**H-Seq2seq** VS. Human Right Ending

| **H-Seq2seq** | Human | both | neither |
|---|---|---|---|
| 3 | 180 | 16 | 1 |

Seq2seq VS. Human Right Ending

| Seq2seq | Human | both | neither |
|---|---|---|---|
| 1 | 194 | 5 | 0 |

**H-Seq2seq** VS. Seq2seq

| **H-Seq2seq** | Seq2seq | both | neither |
|---|---|---|---|
| 30 | 18 | 5 | 147 |

Table 2: Human evaluation results in a pair-wise experiment. The most frequently chosen answers were considered as an agreement among the five workers.

and used 50 stories for evaluation. Five workers provided responses to each question. Examples from the results are shown in Table 3. Similar to Table 2, the agreement among five workers was counted: H-Seq2seq = 0, Human = 42, both = 5, neither = 3 (total of 50 stories). The collected reasons are publicly available[1].

### 4.5 Sentiment Analysis

Regarding SCT, when crowdsourced workers write the "right ending" and "wrong ending" without constraints, the "right ending" tended to be more positive (Sharma et al., 2018). Referring to this finding, we analyzed the endings in SCT, *ROCStories*, and 1,871 endings generated by H-Seq2seq. To calculate sentiment, we used the VADER sentiment analyzer (Hutto and Gilbert, 2014). The results are shown in Table 4.

Focusing on the difference of sentiment between "right ending" and "wrong ending" in SCT, Sharma et al. (2018) aim to improve SCT as a reading comprehension task. In order to eliminate the bias, they apply constraints when they have crowdsourced workers write new "right ending" and "wrong ending". On the other hand, our goal is to clarify what sentiment bias exists when humans freely write the story (SCT "right ending" and ROCStories), and how this sentiment bias is reproduced in SEG as a generation task. As the task of story generation aims to imitate the story that humans write freely, our focus is not on setting constraints when humans write.

[1] https://github.com/mil-tokyo/SEG_HumanEvaluationReasons

## 5 Discussion

In quantitative evaluation with MTurk, H-Seq2seq beat Seq2seq in 30 stories out of 200. As this exceeds the number of stories in which Seq2seq beat H-Seq2seq (18 stories), it can be concluded that H-Seq2seq performs better than Seq2seq. Comparing H-Seq2seq with Human Right Ending, H-Seq2seq is far from generating endings that mirror those written by humans. However, 20 stories were evaluated to be equal to or better than an ending written by humans.

To clarify the characteristic of the endings generated with H-Seq2seq, we analyzed the qualitative evaluation results. Table 3 shows that endings containing positive emotions are frequently generated. This tendency is also supported by sentiment analysis (Table 4). Considering the human-written endings, the mean score of the endings from *ROCStories* is 0.119. This value is significantly different from 0 ($p < 0.05$). Hence, if we have crowdsourced workers write short stories on everyday life, they tend to write stories with happy endings.

We then analyzed the 250 reasons for the choices (five answers for each of the 50 stories). Some examples are shown in Table 3. To identify important elements of the reasons, we tried topic modeling with latent Dirichlet allocation (LDA) (Blei et al., 2003). However, the elements that characterize the topics depended on the content of the story (such as eating or going somewhere), and it was not clear what elements of the reasons were important for the choice. Therefore, we instead used word frequency to analyze 250 reasons. The most frequent 20 words among the 250 reasons are shown in Figure 2. Using this word frequency as a reference, we checked all 250 reasons to determine what was the important factor. "Logical" is a frequently used word; some reasons insisted on the importance of logicality. "Make" and "sense" are both frequently-used words because the idiom "make sense" was commonly used. When a word unrelated to the context was generated, annotators evaluated the generated ending as bad, saying "no mention". As mentioned earlier, an ending is often emotionally biased toward being positive. Therefore, the reasons also included references to emotions, such as "happy". The example in Table 3 shows that an immoral story seems to be disliked. Even the human-written ending was considered as inappropriate. Moreover, there were cases where a choice was made based on common sense, such

| | | |
|---|---|---|
| **Context** | Howard is a senior. He feels a lot of bittersweet thoughts. He holds a senior party with all of his friends. They all enjoyed it and drank a lot. | |
| **Human** | Howard liked socializing. | |
| **H-Seq2seq** | He is happy that he has a good time. | |
| | **Answers with Reasons (A: Human, B: H-Seq2seq)** | |
| | Human (A) | he was sad about leaving his friends |
| | both | Both make sense, even if B has a tad more detail. |
| | both | Either ending will work for the story. A might be a bit better. |
| | both | Both fit, he might have a bittersweet feeling but he would likely be happy at the end of the party, esp if they all enjoyed themselves. |
| | both | He wanted to interact with his friends, and they "all enjoyed it", so he was happy. |
| **Context** | Lily and Pam were popular girls in school. They invited Joy to a diner after school. Joy was not popular and Lily and Pam knew it. They invited her just to bully her when they got there! | |
| **Human** | Joy had brought a gun and shot both the bullies in the face. | |
| **H-Seq2seq** | They had a great time at the party. | |
| | **Answers with Reasons (A: H-Seq2seq, B: Human)** | |
| | Human (B) | B is morbid, but it fits. |
| | Human (B) | People who are bullied sometimes use guns on others. |
| | neither | I don't think she would have a good time with the bully there |
| | Human (B) | a terrible story which shows that bullying is risky; sometimes very risky. |
| | neither | Neither, they certainly didn't have a great time, and why would she shoot them? |

Table 3: Examples of contexts and endings, followed by answers and reasoning provided by MTurk workers.

| | mean | # of Positive | # of Negative |
|---|---|---|---|
| Right | 0.146 | 1,652 (44.1%) | 715 (19.1%) |
| Wrong | 0.011 | 997 (26.6%) | 1,016 (27.2%) |
| *ROCStories* | 0.119 | 39,368 (40.1%) | 20,558 (20.9%) |
| **H-Seq2seq** | 0.457 | 1,337 (71.46%) | 81 (4.33%) |

Table 4: Sentiment of endings calculated with VADER.

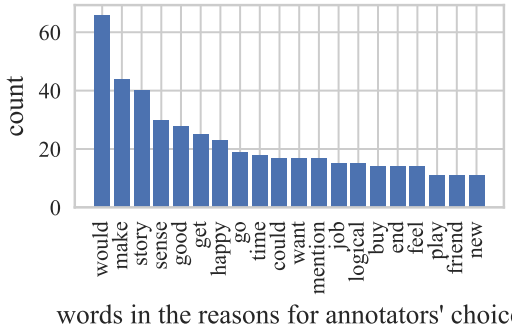


Figure 2: Frequency of words appearing in the reasons written by annotators for their choice.

as "dogs do love to play in the snow." There was a "neither" case involving grammar. For example, an annotator explained that "trying to buy a car" implies that it was not successful; therefore, the human-written ending implying buying a car is considered inappropriate. Thus, it would be desirable that the evaluation metric be designed by considering that annotators are conscious of emotions, morals, and common sense, in addition to logic and grammar.

## 6 Conclusion

We undertook an SEG task, and examined how to make manual evaluation more effective. As a baseline method, we introduced a hierarchical sequence-to-sequence model. Our focus is not on proposing a better model, but on discussing how to conduct human evaluation. Through quantitative and qualitative evaluations, we showed how well a baseline model performs and what drawbacks it has. To examine the qualities of the generated endings, we asked crowdsourced workers to provide reasons for their choice. This qualitative evaluation illustrates the characteristics of our baseline method. The analysis indicated that the evaluation metric should be designed by considering that workers are conscious of emotions, morals, and common sense when they evaluate story endings. Although the amount of analyzed data is limited, we believe that the findings obtained by human-reasoned evaluation would contribute to suggest metrics for story generation in future research.

## Acknowledgements

This work was partially supported by JST CREST Grant Number JPMJCR1403, Japan.

## References


Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, volume abs/1808.10113.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *International AAAI Conference on Web and Social Media*, pages 216–225. The AAAI Press.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1033–1043, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hareesh Ravi, Lezi Wang, Carlos Muniz, Leonid Sigal, Dimitris Metaxas, and Mubbasir Kapadia. 2018.

Show me a story: Towards coherent neural story illustration. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7613–7621.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 553–562, New York, NY, USA. ACM.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Yan Zhao, Lu Liu, Chunhua Liu, Ruoyao Yang, and Dong Yu. 2018. From plots to endings: A reinforced pointer generator for story ending generation. In *Proceedings of Natural Language Processing and Chinese Computing (NLPCC)*, volume abs/1901.03459.

# A Supplemental Material

## A.1 Training details

**H-Seq2seq** Pre-trained InferSent was applied as the sentence encoder. Pre-trained word embeddings "GloVe" (Pennington et al., 2014) is used in InferSent. The sentence encoder is not fine-tuned during training. We used a 1-layer GRU for each context encoder and decoder and set the number of hidden layer units to 256. We set the dropout ratio to 0.3 for word-dropout and to 0.5 for sentence-level dropout. We used Adam optimization with parameters $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-08$. The results obtained from 15 epochs were used for evaluation.

We implemented the method with Chainer, a Python-based deep learning framework (Tokui et al., 2015).

**Seq2seq** The series of input words were handled collectively by Seq2seq without considering sentence-level information. We used a 2-layer LSTM encoder and 2-layer LSTM decoder. Each hidden layer contained 512 units. We used the results from 15 epochs.

## A.2 Human Evaluation with MTurk

We conducted the quantitative evaluation and the qualitative evaluation with help from MTurk workers. In Figure A.1, we show the snippet of the instruction and the question we used in the qualitative evaluation. In the quantitative evaluation, the reason for the choice was not asked.

Each story was evaluated by five workers. Workers chose their answer from "A", "B", "both A and B", and "neither A nor B". Among the five answers obtained for each story, The most frequently chosen answers were considered as an agreement among the workers. We should note that we had to handle exceptions if multiple answers were most popular. If "A" and "B" received two votes each, we considered the agreement among workers to be "both". Similarly, if "A" and "both" received two votes each, we considered the agreement among workers to be "A". Similarly, if "B" and "both', then "B"; if "A" and "neither', then "A"; if "B" and "neither', then "B"; if "both" and "neither", then "both".

## A.3 Sentiment Analysis

According to the original paper on VADER, We count an ending as positive if the score $\geq 0.05$ and count it as negative if $\leq -0.05$.

## A.4 Pre-processing for Word Frequency

Pre-processing of word frequency was done as below. First, we used Gensim (Řehůřek and Sojka, 2010) for converting sentences into lists of lowercase tokens. Second, we removed stop words with NLTK (Bird et al., 2009). Then, we lemmatized the words with spaCy (Honnibal and Montani, 2017). We used the words with part-of-speech (pos) tags 'NOUN,' 'ADJ,' 'VERB,' and 'ADV.' After the pre-processing process, we counted the number of tokens.

**Pick the right endings to complete unfinished stories.**

For each question in this task, you are given an unfinished short story and two options, A and B, to complete the story.

Each story originally consists of 5 sentences, but the 5th sentence is lost.

Please answer which of A and B do you think more appropriate as the 5th sentence (in other words, "ending") to complete each story .

- If you think A is more appropriate, please choose "A".
- If you think B is more appropriate, please choose "B".
- If you think both A and B are equally appropriate, please choose "both A and B".
- If you think neither A nor B are suitable, please choose "neither A nor B".

You are also required to write the reason why you choose the answer.

# Question 1:

Last week a group of friends decided to go on a hike. They went to a mountain nearby. They followed a popular trail and set out for the day. The walk was exhausting but worth it.

**Which is the right ending following this story?**

- A: They had a great time at the beach.
- B: They felt good.

| |
|---|
| A |
| B |
| both A and B |
| neither A nor B |

**Please write the reason why you choose the answer.:**

Figure A.1: The instruction and the question snippet we used in the qualitative evaluation.