The Nectar of Missing Position Prediction for Story Completion

Yusuke Mori[1] Hiroaki Yamane[2,1] Yusuke Mukuta[1,2] Tatsuya Harada[1,2]
[1]The University of Tokyo, [2]RIKEN

Code

Overview

- This paper is the **nectar** of
  "***Finding and Generating a Missing Part for Story Completion***,"
  we proposed in LaTeCH-CLfL 2020 (COLING 2020 workshop).
- We summarize the major points of the previous paper and add some more insights and discussions.

Main Contributions of The Summarized Paper

- **Propose “Missing Position Prediction (MPP)” task**
  - Predicts the position of a missing part of an incomplete story; this has significance in the context of support for the creation of stories.
- **Propose a novel method for Missing Position Prediction**
  - An analysis shows that highly accurate predictions can be obtained when the missing part is the beginning or end.
- **Story Completion with Missing Position Prediction**
  - It is possible to restore a story comparable with the original human-written story in 26% of cases.

Main Contributions of This Nectar Paper

Summarize [Mori et al., 2020]

- **Propose "Missing Position Prediction (MPP)" task**
  - Predicts the position of a missing part of an incomplete story; this has significance in the context of support for the creation of stories.
- **Propose a novel method for Missing Position Prediction**
  - An analysis shows that highly accurate predictions can be obtained when the missing part is the beginning or end.

Adding more Insights and Discussions

- **Conduct Further Analysis**

Background – Story Understanding/Generation

- Today, thanks to the Internet, anybody can freely publish their original stories.
- However, writing a story is not easy.
- **To write/generate a good story, a human/an AI model must know what a good story is.**
  - Contrary, writing a story makes us understand more about the secret of creating a story.

Background – Human Story Writing Assistance

- Story writers have various problems…
  - can’t complete one’s story
  - works don't sell well
  - balance with other work (if they are part-time writers)
  - etc.
- Recent progress in natural language processing makes it feasible to support human creative endeavors.
- To assist writers in creating stories, it is essential to train computers to understand and generate stories.

Background – Creative Process (Example)

Genre

Mystery, SF,
Romance, …

Proposal
- Theme
- Target
- Related works
- Character settings
- Plot

About the purpose of living,
For teenagers,
…

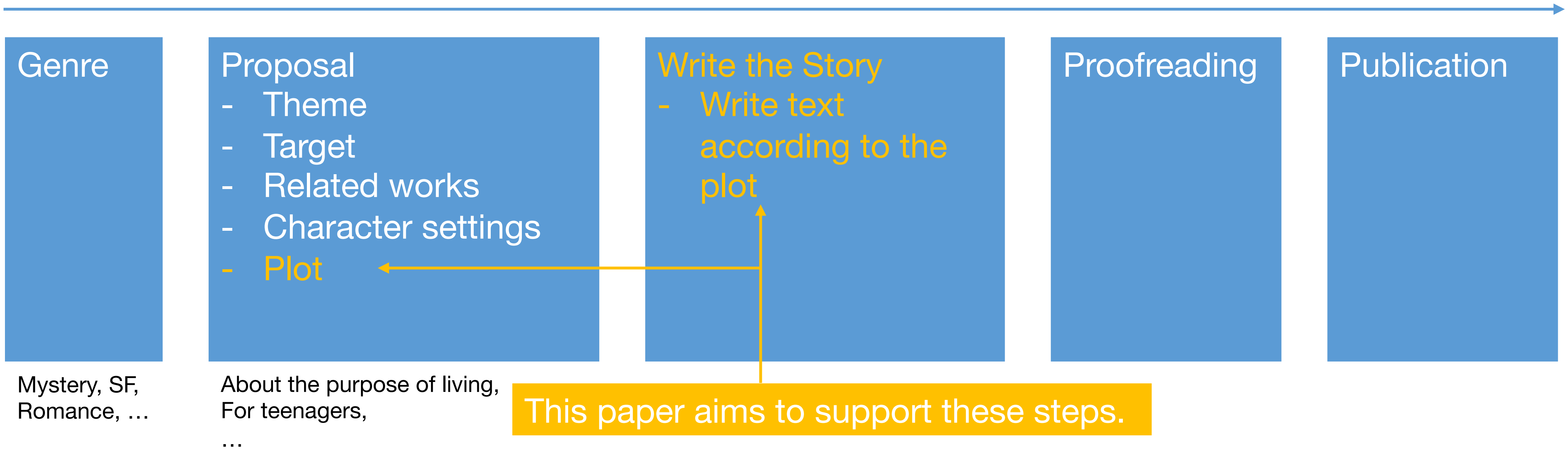
Write the Story
- Write text according to the plot

Proofreading

Publication

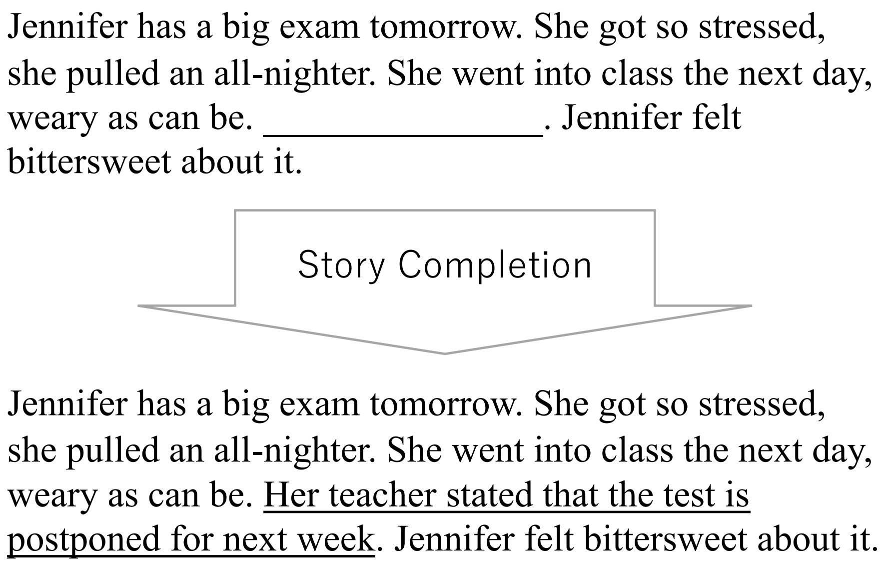
- Creation method is different for media, genre, and even for each creators. Above is just an example.
- Sometimes it goes back to the previous step and starts over.

Background – Creative Process (Example)



- Creation method is different for media, genre, and even for each creators. Above is just an example.
- Sometimes it goes back to the previous step and starts over.

Related Work: Story Completion



- Wang and Wan [2019] proposed "**Story Completion (SC)**" task in the field of generating and understanding stories.
- Given any four sentences of a five-sentence story, the objective of SC is to generate the sentence that is not given.

Related Work: Cloze Test for Story Understanding

- Cloze Test [Taylor, 1953] is a well-known assessment for readability.
  - Some words are lost from a text, and a test taker fills them.
- Close Test for Story/Narrative
  - **Narrative Cloze Test** [Chambers & Jurafsky, 2008]
  - **Story Cloze Test** (SCT) [Mostafazadeh et al., 2016]
    - From an original five-sentence story, the last sentence is excluded. The 1st to 4th sentences are presented and the objective is to select an appropriate sentence from two options that complement the missing last sentence.

"Cloze Procedure": A New Tool for Measuring Readability" [Taylor, 1953]
"Unsupervised Learning of Narrative Event Chains" [Chambers & Jurafsky, 2008]
"A corpus and cloze evaluation for deeper understanding of commonsense stories" [Mostafazadeh et al., 2016]

Related Work: Story Generation inspired by SCT

- Subtasks of Story Generation inspired by SCT.
- Tasks of completing incomplete stories:
  in other words, cloze tasks that "generate" rather than "select".
  - **Story Ending Generation** [Zhao et al., 2018]
  - **Story Completion** [Wang & Wan, 2019]

"From Plots to Endings: A Reinforced Pointer Generator for Story Ending Generation" [Zhao et al., 2018]
"Transformer-Based Conditioned VAE for Story Completion" [Wang & Wan, 2019]

Challenge

- Previous sentence-level story cloze tasks require a user to have prior knowledge of the missing parts.
  - **SEG**: the last sentence is lost.
  - **SC**: the k-th sentence is lost. k *is given.*
- The case where the missing parts is not known was remain untouched.
- Our **Missing Position Prediction** aims to fill this gap.

Objective

- **Finding a missing part in the flow of a story** to complete an incomplete story
  - For Story Understanding, Story Generation, and Creative Support

  - Note that the incompleteness of a story is not only “missing”, but also there are other kind of incompleteness.
  - **Writing a story and supporting it is a complex and difficult task. We tackled it first by focusing on one important step.**

Main Contributions of The Summarized Paper

- **Propose “Missing Position Prediction (MPP)” task**
  - Predicts the position of a missing part of an incomplete story; this has significance in the context of support for the creation of stories.
- **Propose a novel method for Missing Position Prediction**
  - An analysis shows that highly accurate predictions can be obtained when the missing part is the beginning or end.
- **Story Completion with Missing Position Prediction**
  - It is possible to restore a story comparable with the original human-written story in 26% of cases.

Main Contributions of This Nectar Paper

Summarize [Mori et al., 2020]

- **Propose “Missing Position Prediction (MPP)” task**
  - Predicts the position of a missing part of an incomplete story; this has significance in the context of support for the creation of stories.
- **Propose a novel method for Missing Position Prediction**
  - An analysis shows that highly accurate predictions can be obtained when the missing part is the beginning or end.

Adding more Insights or Discussions

- **Conduct Further Analysis**

Related Task: Story Completion

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. ________________. Jennifer felt bittersweet about it.

Story Completion

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. <u>Her teacher stated that the test is postponed for next week</u>. Jennifer felt bittersweet about it.

- Wang and Wan [2019] proposed “Story Completion (SC)” task in the field of generating and understanding stories.
- Given any four sentences of a five-sentence story, the objective of SC is to generate the sentence that is not given.

Proposal: “Missing Position Prediction (MPP)”

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Jennifer felt bittersweet about it.

Missing Position Prediction

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. ________________. Jennifer felt bittersweet about it.

Story Completion

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it.

- Existing “Story Completion”
  - The position of a missing part is given.

- **Our New Task**
  - Predicting the position of a missing part from a remaining context.

How are these tasks applied to the real situation?

- Advise a writer where and how to brash up his/her story.

How do you think **this story**?

Writer

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Jennifer felt bittersweet about it.

Editor

I recommend you add information as "Her teacher stated the test was postponed" after the 3rd sentence!

How are these tasks applied to the real situation?

- Advise a writer where and how to brash up his/her story.

How do you think **this story**?

Writer

Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated the test was postponed. Jennifer felt bittersweet about it.

Editor

I recommend you add information as "Her teacher stated the test was postponed" after the 3rd sentence!

Formulation of the MPP

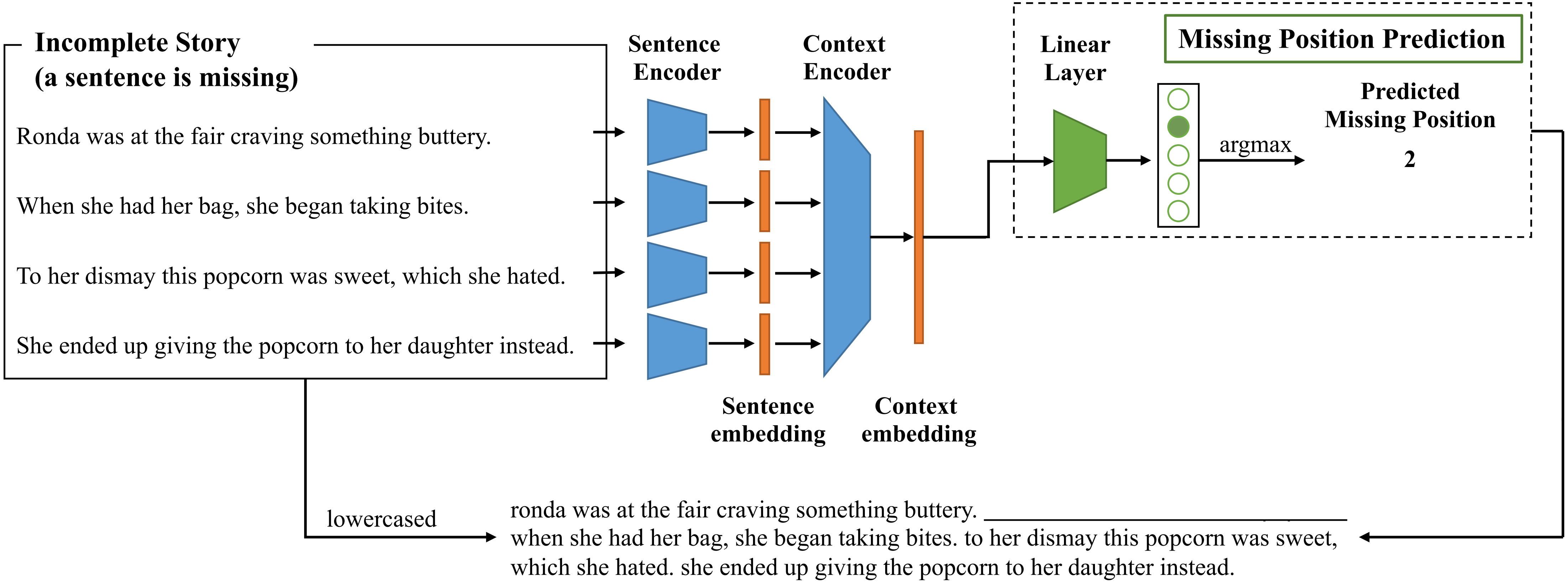
- We define $S = \{s_1, s_2, \dots, s_n\}$ as a story comprising n sentences.
- Input: an incomplete story consists of $n - 1$ sentences

$$S' = \{s_1, \dots, s_{k-1}, s_{k+1}, \dots, s_n\}$$

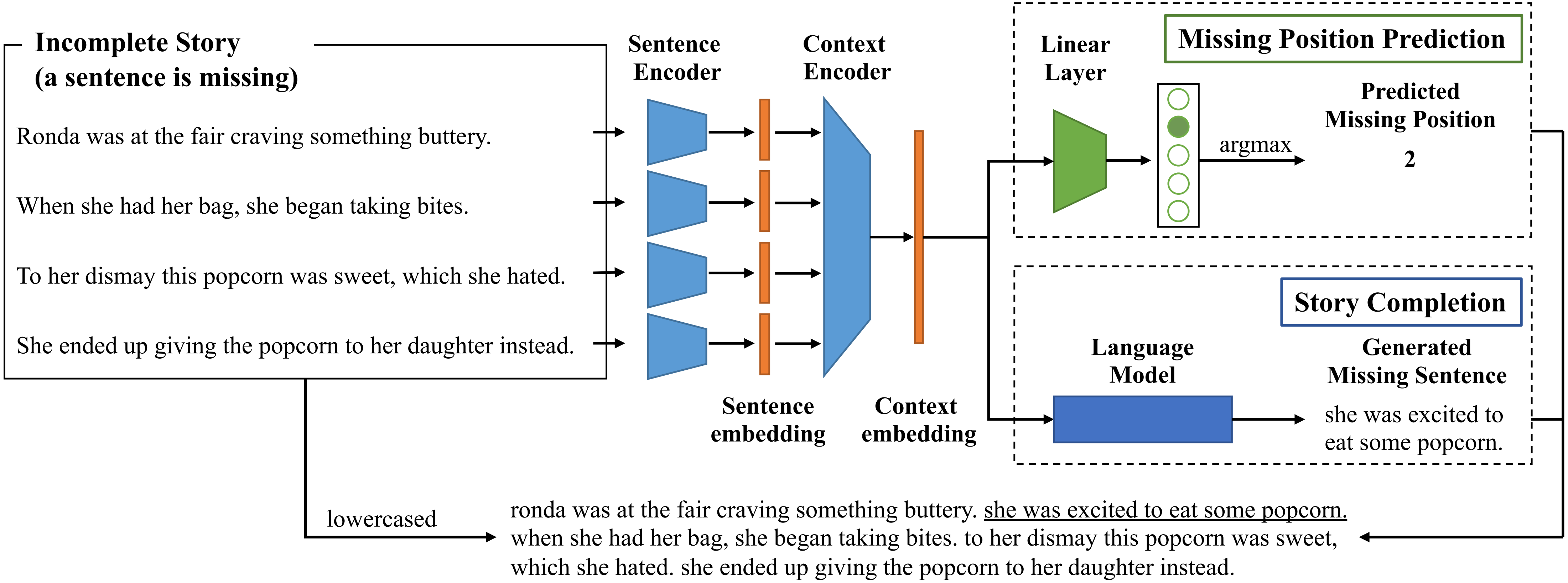
  where k represents the position of the missing sentence.
  - Any information regarding k is not given.
- Objective: predict k from S'

- (Reference)
  - The objective of SC is to generate s_k from S' and given k.

Proposed Method for MPP



Proposed Method for MPP and SC



Experiments

- We conducted two experiments in [Mori et al., 2020], but in this study we focused on Experiment 1.
- **Experiment 1:**
  Missing Position Prediction
  - Investigated the part of the proposed method that excludes LM, to show the desired Context Encoder.
- **Experiment 2:**
  Missing Position Prediction + Story Completion
  - Based on the results of Experiment 1, we tackled the combined task of Missing Position Prediction and Story Completion.

Experiments

- We conducted two experiments in [Mori et al., 2020], but in this study we focused on Experiment 1.
- **Experiment 1:**
  Missing Position Prediction
  - Investigated the part of the proposed method that excludes LM, to show the desired Context Encoder.
- **Experiment 2:**
  Missing Position Prediction + Story Completion
  - Based on the results of Experiment 1, we tackled the combined task of Missing Position Prediction and Story Completion.

Experiment 1: Dataset

- ROCStories [Mostafazadeh et al., 2016]
  - A large-scale story corpus (about 10K stories) proposed with SCT.
  - It is a collection of non-fictional daily-life stories written by hundreds of workers belonging to Amazon Mechanical Turk.

  - Not only for SCT, but it is also widely used in story generation tasks:
    - Story Ending Generation [Zhao et al., 2018; Li et al., 2018; Guan et al., 2019]
    - Story Completion [Wang and Wan, 2019]

Generating reasonable and diversified story ending using sequence to sequence model with adversarial training” [Li et al., 2018]
“Story ending generation with incremental encoding and commonsense knowledge” [Guan et al., 2019]

Experiment 1: Dataset

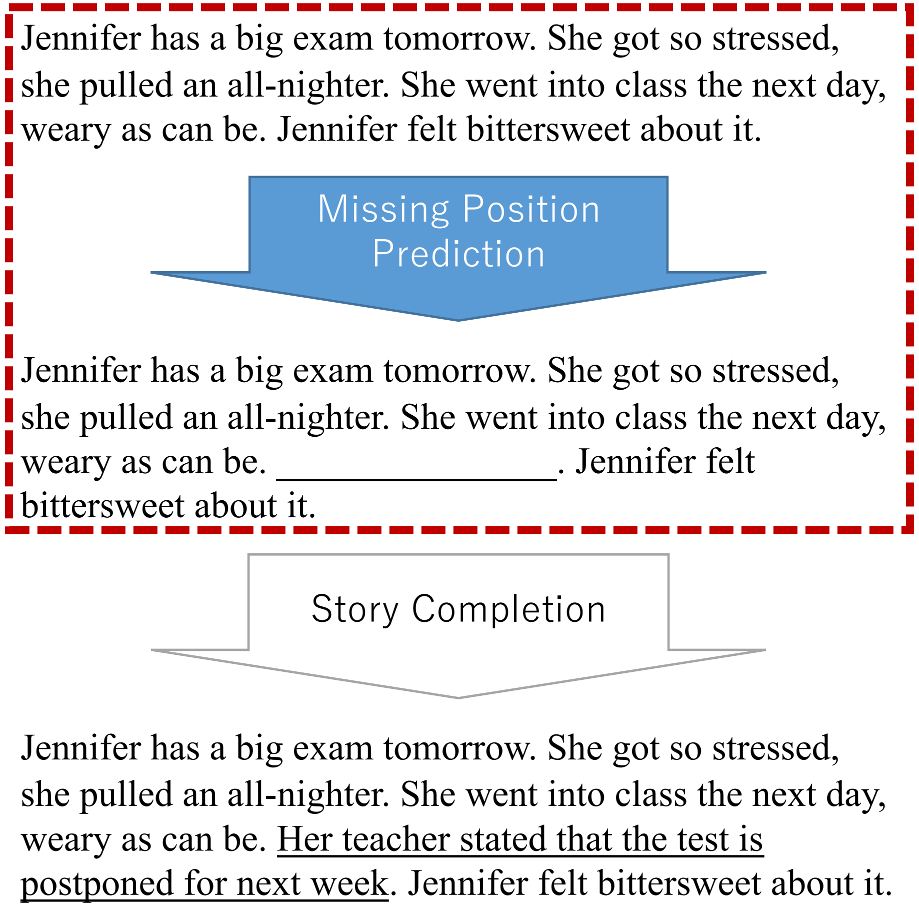
- ROCStories [Mostafazadeh et al., 2016]

Table 1: Overview of the Dataset Used

set	#stories	missing position
train	78,528	Given randomly during training
validation	9,816	Given when creating dataset
test	9,817	Given when creating dataset
total	98,161	

  - For our proposed MPP task, we randomly split the dataset in the ratio of 8:1:1 to obtain the train/validation/test sets.
  - For each story, one sentence was randomly excluded to create an incomplete story

Experiment 1: Settings



- Model
  - Sentence Encoder
    - Sentence-BERT (SBERT) [Reimers & Gurevych, 2019]
  - Context Encoder
    - GRU Context
      - Treat sentences as a sequence
    - Max-pool Context (comparison method)

“Sentence-BERT: Sentence embeddings using Siamese BERT-networks” [Reimers & Gurevych, 2019]

Experiment 1: Settings

- Training epochs: 30
    - The state with the smallest validation loss was used for further tests.
- Optimizer: Adam [Kingma & Ba, 2015]
    - a learning rate of 0.001, $\beta 1 = 0.9$, $\beta 2 = 0.999$, and a weight decay of 0. Gradient clipping with a value of five was used.
- Batch size: 256
- Among the trained SBERTs, we used “bert-base-nli-mean-tokens,” where the output dimension was 768. The Context Encoder consists of a GRU with 256 hidden units, and a linear layer with 256 dimensions for both the input and output.
- To obtain the five-class prediction, we use another linear layer to receive the Context Encoder’s output with 256 dimensions and subsequently outputs five dimensions.

“Adam: A Method for Stochastic Optimization” [Kingma & Ba, 2015]

Experiment 1: Results

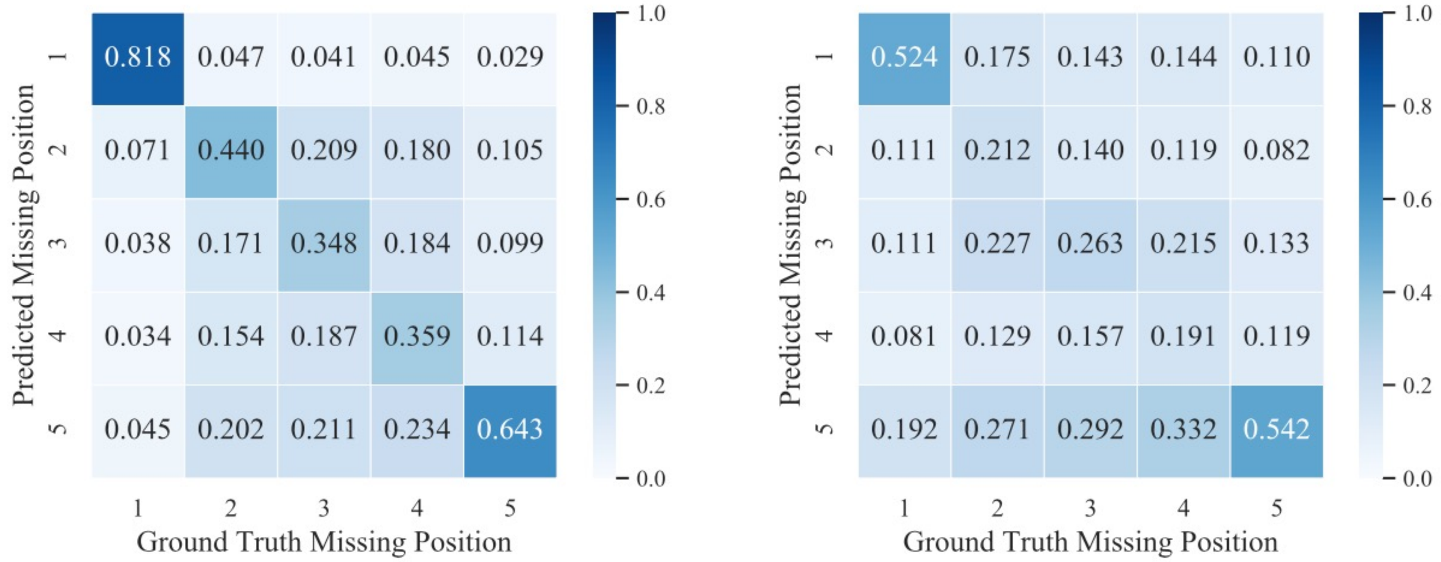
- Overall Accuracy

Methods	Accuracy (%)
Max-pool Context	35.0±0.334
GRU Context	**52.2±0.220**

  - Prediction accuracy, shown as $mean \pm std$. It is a five-class classification task, so the chance rate is 20%.

Experiment 1: Results

- Accuracy for each position
  - The performance is lower when $k = 2,3,4$ than when $k = 1,5$.

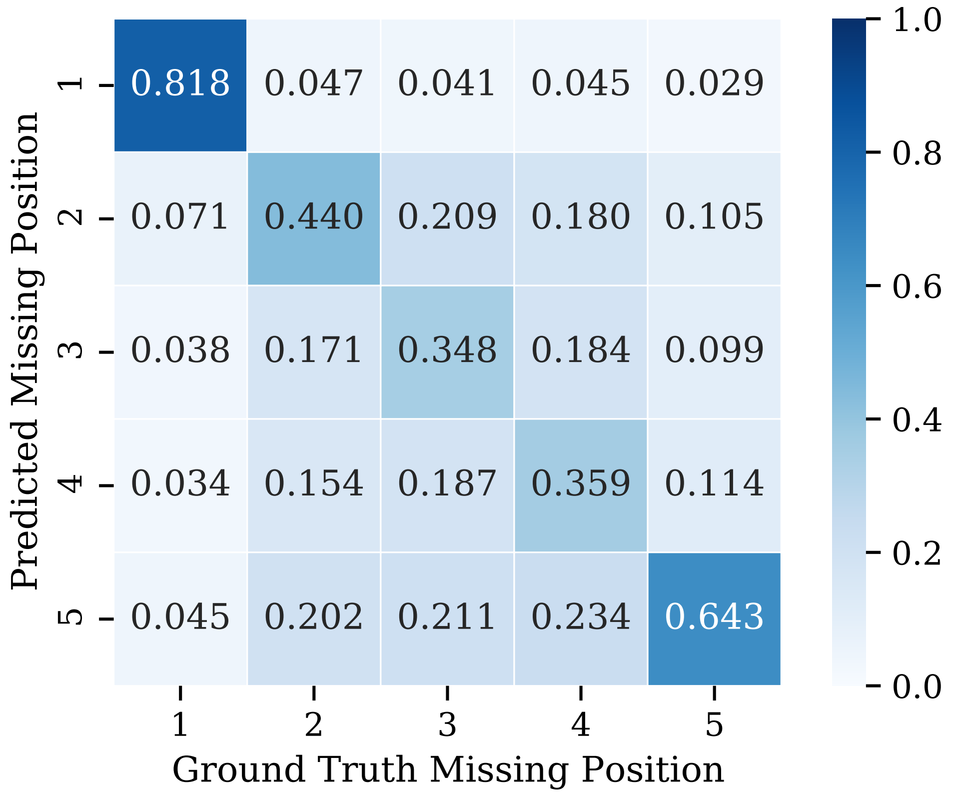


(a) GRU context (b) Max-pool context

Figure 3: Heat maps showing the results of the (a) GRU context and (b) Max-pool context. The ground truth (GT) label is shown on the x-axis and the predicted label is on the y-axis. The squares on the diagonal line denote correct cases. The ratios of the predicted label to the GT label are shown numerically.

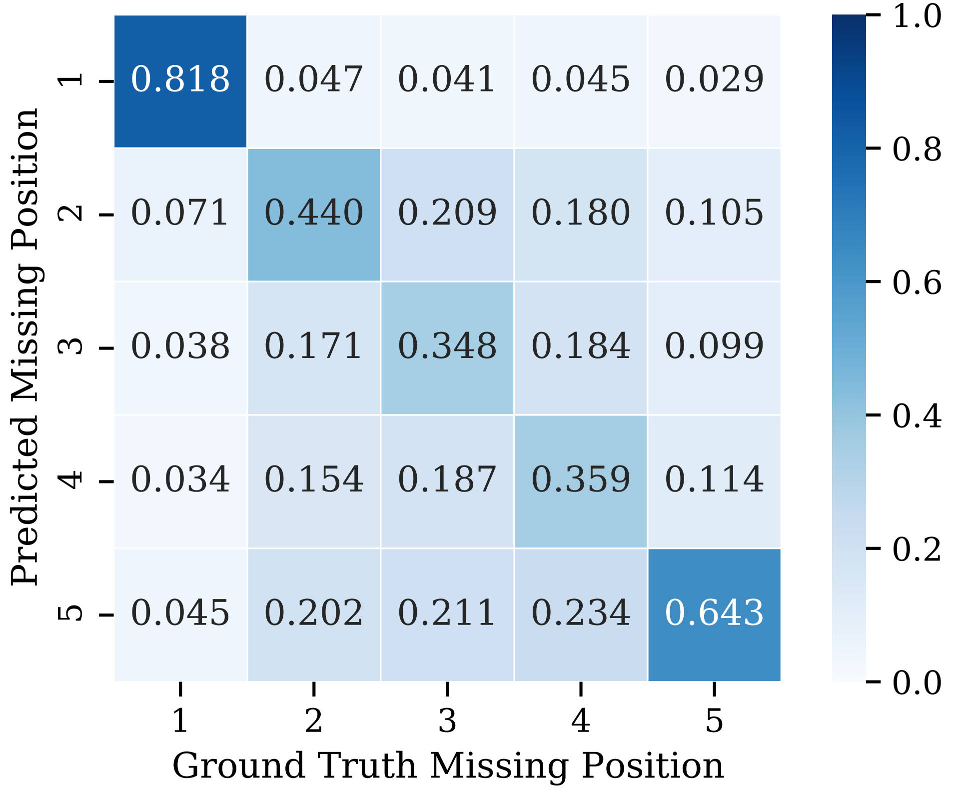
Discussions about Experiment 1

- When the first or fifth sentence was missing, the accuracy was higher than when the second, third, or fourth sentence was missing.
- In other words, the beginning or the ending of a story can be easily predicted when they are the lost sentence.



Discussions about Experiment 1

- This appears to be related to how ROCStories was collected: "the story should read like a coherent story, with a specific beginning and ending, where something happens in between."
- Thus, it is likely that if the beginning or the ending is missing, our method can interpret it as unnatural.



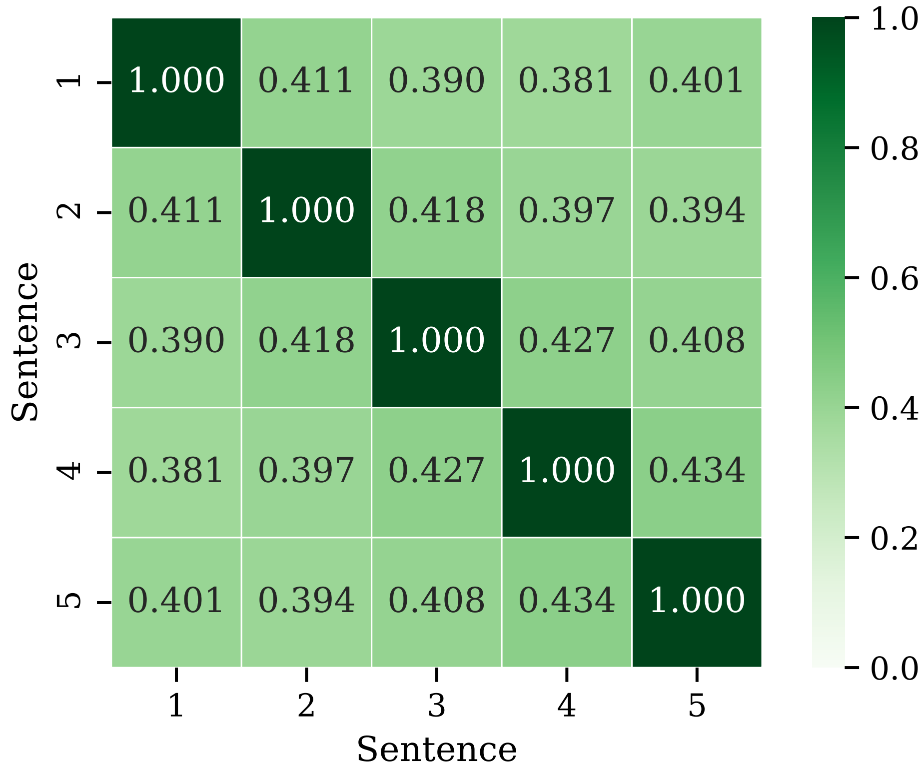
Further Analysis

Further Analysis

- We conducted further analysis on the finding: the beginning or the ending is easy to predict as missing.
- Two hypotheses:
  - The second to fourth sentences in a five-sentence story are represented by a similar sentence embedding.
  - If part-of-speech (POS) tagging is applied for each sentence, a different trend may be observed.

Hypothesis 1: The Similarity of Embeddings in Middle Sentences

- The similarity of Embeddings in training set
  - Sentence Embeddings by SBERT



Hypothesis 1: The Similarity of Embeddings in Middle Sentences

- For each story in the training set, we calculated the cosine similarity among the sentence embeddings of five sentences.
- Contrary to our hypothesis, the sentence embeddings of the second to fourth sentences were not more like each other than to the first and fifth sentences.

Hypothesis 2: POS tagging

- POS tagging of words in each sentence was performed for all stories in the training set.
- For each sentence number, we took the average number of times each tag appeared in that sentence.

Hypothesis 2: POS tagging

- The focus is on whether there is anything special about the distribution of POS tags in the first and last sentences.
- In the first sentence, the appearance of PROPN (proper noun) is remarkable, and in the fifth sentence, although not as prominent, ADV (adverb) and ADJ (adjective) occur frequently.

Conclusion

- In the summarized paper
  - To overcome the limitation of the conventional SC task, we proposed “Missing Position Prediction” to predict the position of the missing part based on the given incomplete story. We found that a prediction is easier if the beginning or the end of a story is missing.
- In this study (nectar paper)
  - The further analysis suggested that the distribution of POS tags may play a significant role in prediction accuracy.
  - Whether the prediction models pay attention to words with these tags is a subject for future analysis.

Supplement

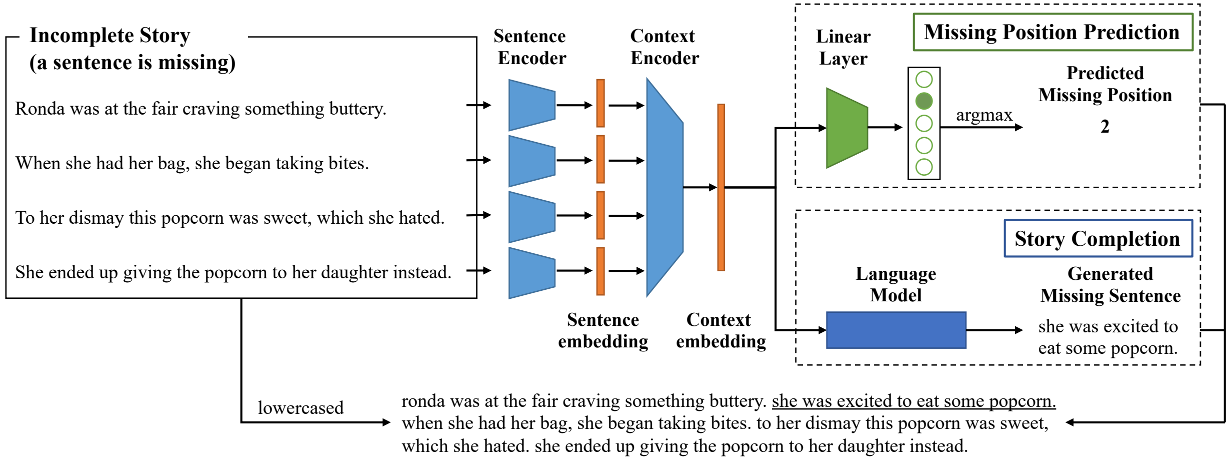
Experiment 2 of the summarized paper

Experiments

- **Experiment 1:**
  Missing Position Prediction
  - First, worked only on the proposed task.
  - Investigated the part of the proposed method that excludes LM, to show the desired Context Encoder.
- **Experiment 2:**
  Missing Position Prediction + Story Completion
  - Based on the results of Experiment 1, we tackled both Missing Position Prediction and Story Completion.
  - Conducted Human Evaluation with MTurk.

Experiment 2: Settings

- Dataset
  - ROCStories (same as Experiment 1)
- Model
  - Sentence Encoder
    - Sentence-BERT
  - Context Encoder
    - GRU Context
  - Language Model
    - BERT [Devlin et al., 2019]



"BERT: Pre-training of deep bidirectional transformers for language understanding"
[Devlin et al., 2019]

Experiment 2: Human Evaluation

- Qualification Test
  - Choose workers with the ability in evaluating stories.
  - Used randomly selected ten questions from Story Cloze Test.
- Pair-wise Evaluation Task
  - The qualified workers were given two similar short stories and asked to choose which story gave the impression of being a complete story.
  - Five MTurk workers evaluated each story pair. The most popular answer of five workers was considered as an agreement.

Experiment 2: Human Evaluation

Proposed	GT	both	neither
8	148	44	0

- We used 200 stories. The most frequently chosen answers by five workers were considered as their agreement.
- **Our proposed method can generate a story that is as good as or better than a GT story with 26% probability.**
    - Regarding a simpler task, SEG, we had conducted a similar pairwise evaluation [Mori et al, 2019]. The generated endings were as good as or better than GT with a 10.5% ratio then. Thus, we believe that the 26% for this more difficult task is noteworthy.

“Toward a better story end: Collecting human evaluation with reasons” [Mori et al., 2019]

Experiment 2: Generation Example

Context	since the questions were complicated, i was extremely nervous. despite believing that i've failed, i turned the exam in. the teacher handed the exams back to us the next day. i ended up receiving a b.
GT	i took my class final in math today. since the questions were complicated, i was extremely nervous. despite believing that i've failed, i turned the exam in. the teacher handed the exams back to us the next day. i ended up receiving a b.
Ours	my teacher gave us a test. since the questions were complicated, i was extremely nervous. despite believing that i've failed, i turned the exam in. the teacher handed the exams back to us the next day. i ended up receiving a b.

Answers with Reasons (A: GT, B: Ours)

both	Whether it is a class final or a given test, both stories are the same and therefore both complete.
neither	both doesn't make sense
Ours	A is jumbled and does not make sense. B is logically arranged as a story.
Ours	In "A," it wouldn't make sense that a final exam was handed back in class the next day.
Ours	B was more appropriate since it is having a continuous flow than A

Context	tom was at a local park. there was an egg hunt for the kids. tom decided to pick some eggs up. he enjoyed the treats in them.
GT	tom was at a local park. it was easter. there was an egg hunt for the kids. tom decided to pick some eggs up. he enjoyed the treats in them.
Ours	tom was at a local park. there was an egg hunt for the kids. tom decided to pick some eggs up. tom was able to get many eggs. he enjoyed the treats in them.

Answers with Reasons (A: GT, B: Ours)

both	both are complete sentences
Ours	Option B is complete as it says that tom was able to get some eggs in the hunt.
both	Both of them can be considered complete. Story A tells us it is Easter (and story B doesn't) while Story B tells us Tom picked many eggs (and story A doesn't). Both of those details could be removed and the stories would still be the same.
Ours	The fact that he was able to gather some eggs was more complete than just deciding to pick up some eggs. Story A Easter gave a better time context but did not really add as much to the story since traditionally an egg hunt is held on Easter so the omission of that in Story B was made up for Tom being able to gather some eggs.
both	Both stories have a starting, content and ending.

Experiment 2: Generation Example

Context	timothy loved to dance. timothy didn't have much confidence in himself. it took everything he had to dance with all of his self doubt. everyone loved his dancing and he won the contest.
GT	timothy loved to dance. there was a dance contest that was coming up soon. timothy didn't have much confidence in himself. it took everything he had to dance with all of his self doubt. everyone loved his dancing and he won the contest.
Ours	timothy loved to dance. he decided to take dance lessons. timothy didn't have much confidence in himself. it took everything he had to dance with all of his self doubt. everyone loved his dancing and he won the contest.

Answers with Reasons (A: Ours, B: GT)	
GT	Only B makes sense and a complete story.
GT	A is more correct and arranged
GT	Story A doesn't mention the contest which Timothy ends up winning, therefore misses an important piece of the story.
GT	Story B mentions that there was a dance contest at the start and that he won it at the end. Story A only mentions a contest abruptly at the end making it seem out of place.
GT	B is more good

Experiment 2: Conclusion

- Tackled the combined task of missing position prediction and story completion. In human evaluation, our proposed method showed promising results.

Thank you!

- Our code for MPP is publicly available at https://github.com/mil-tokyo/missing-position-prediction

- Contact
  - Yusuke Mori (corresponding author)
    - mori [at] mi.t.u-tokyo.ac.jp