KNOREX®

จุฬาลงกรณ์มหาวิทยาลัย
Chulalongkorn University
Pillar of the Kingdom

Plot Writing From ~~Scratch~~ Pre-Trained Language Models

Yiping Jin, Vishakha Kadam, Dittaya Wanvarie

@INLG 2022

Motivation

- **Pre-trained language models (PLMs)** fail to generate long-form narrative text; don't consider global structure
  - Generated texts - incohesive, repetitive, lack content

E.g. text generated by pretrained gpt2 model when provided with a prompt

"***On a hot summer day**, I would walk out of the small bar and get breakfast and then come back upstairs and drink another hot drink that morning, and that afternoon, I would drink another hot drink by myself every few days. If my husband and..*

*Generated using https://huggingface.co/tasks/text-generation demo

Motivation

- Recent work in long-form story generation reintroduced *explicit content planning* **(Reiter and Dale, 1997)**

- These content plans are not in natural language*
  - Prompts **(Fan et al., 2018)**
  - Keywords/Key phrases **(Xu et al., 2018; Yao et al., 2019)**
  - Semantic frames **(Fan et al., 2019)**
  - Summaries **(Sun et al., 2020)**

*Except for using summaries as the content plan

Motivation

- Story generation modelled as supervised learning task
  - Model learns common sense and frequently occurring actions

e.g. Generated action plan from **Fan et al., 2019**

Prompt: Everything you do in a day.
Generated SRL Sequence:
<V> woke up <A1> ent0 <A2> to the sound of my alarm blaring
<V> rolled over <A1> ent0
<V> looked <A0> ent0 <A1> at my alarm clock
<V> went <A0> ent0 <A4> to the bathroom
<V> wash <A0> ent0 <A2> myself <A1> of my clothes
<V> got <A0> ent0 <A1> dressed <A2> in my boxers
<V> dressed <A0> by my boxers <A1> ent0 <A2> as ent0 walked to the kitchen
<V> walked <A0> ent0
<V> turned on <A0> ent0 <A1> the water
<V> poured <A0> ent0 <A1> a cup of coffee <A3> in
<V> sat down <A1> ent0

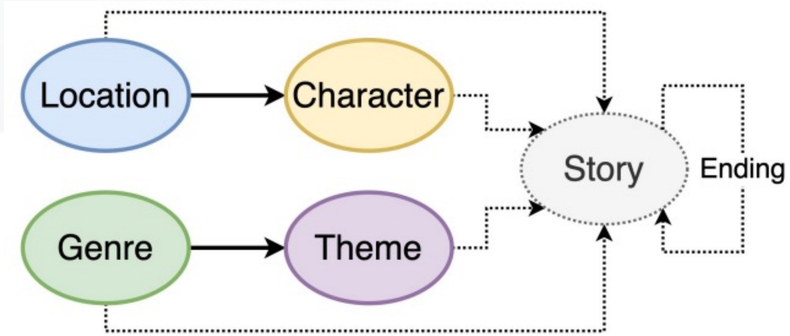
Here the model generates a common morning routine

- Lacks interestingness and level of surprise – two important characteristics of stories

We propose generating story plots using off the-shelf PLMs while maintaining the benefit of content planning to generate cohesive and contentful stories

Method

We propose ***SCRATCHPLOT***, a method to ***generate stories using off-the-shelf PLMs*** *without fine-tuning*.



Step 1: Progressive content planning

Task: Write a plot summary of a *{genre}* story featuring *{character1}* and *{character2}* in *{location}* with the main theme *{theme}*

Plot summary: “”

Step 2: Generate story body

Task: Write the ending of a *{genre}* story.

What happened earlier: *{story}*

What happens in the end: “”

Step 3: Generate story ending and select the top ranked ending

Figure 1 : Overview of SCRATCHPLOT

- **SCRATCHPLOT** outperformed the baselines on different story generation aspects - naturalness, interestingness, cohesiveness

Method

Step 1: Progressive Content Planning

- We use ***D**atasets from **In**structi**o**ns* (**DINO**) framework **(Schick and Schütze, 2021)** to generate plot elements

- It uses a pre-trained GPT-2 model to generate entire datasets of labeled text pairs from scratch by providing task descriptions/instructions

Task: Write two sentences that mean the same thing.
Sentence 1: "A man is playing a flute."
Sentence 2: "He's playing a flute."

Task: Write two sentences that are somewhat similar.
Sentence 1: "A man is playing a flute."
Sentence 2: "A woman has been playing the violin."

Task: Write two sentences that are on completely different topics.
Sentence 1: "A man is playing a flute."
Sentence 2: "A woman is walking down the street."

Figure 2 : Text pairs generated using DINO **(Schick and Schütze, 2021)**

- Uses two different unsupervised approaches
  1. The input sentence is given and only the continuation is generated
  2. Both input sentence and continuation are generated.

Method

Step 1: Progressive Content Planning

- We define four main plot elements: **location, characters, genre, and theme**. These elements are not entirely independent.
  e.g. the genre will influence the theme

Element	Task Description
Location	**Task:** Write the name of a country.\n **Country:** “ **Task:** Write the name of a province.\n **Province:** “ **Task:** Write the name of a city.\n **City:** “ **Task:** Write the name of a county.\n **County:** “
Cast	**Task:** Write the male character’s full name in a story that happened in <X1>. **Full name:** “ **Task:** Write the female character’s full name in a story that happened in <X1>. **Full name:** “
Genre	**Task:** Write a story genre.\n **Story genre:** “ **Task:** Write a literary genre.\n **Literary genre:** “ **Task:** Write a novel genre.\n **Novel genre:** “
Theme	**Task:** Write the main point from a <X1> story.\n **Main point:** “ **Task:** Write the twist in a <X1> story.\n **Twist:** “ **Task:** Write the lesson learned from a <X1> story.\n **Lesson learned:** “ **Task:** Write the spectacle of a <X1> story.\n **Spectacle:** “

Table 1. List of task descriptions to generate each element. <X1> denotes the previously generated element

Method

Step 2: Generate Story Body

- ❑ Sample one value for each plot element except for characters, generate separate male and female characters

- ❑ After sampling all plot elements → fuse them into a single task description to generate the story

> **Task:** Write a plot summary of a *{genre}* story featuring *{character1}* and *{character2}* in *{location}* with the main theme *{theme}*
>
> **Plot summary:** “”

- ❑ Generate the story with a fixed length and truncate it till the last complete sentence

Method

Step 3: Generate Story Ending

- Design a separate task description to write the story ending explicitly

> **Task:** Write the ending of a *{genre}* story.
>
> **What happened earlier:** *{generated story}*
>
> **What happens in the end**: “”

- Rank the story endings:
  - **Next Sentence Prediction** (NSP) task of BERT **(Devlin et al., 2019)**
  - **Perplexity Score (Lee et al., 2020, 2021)**

Method

Step 3: Generate Story Ending

❑ **Next Sentence Prediction**

Measure the coherence between the story and the ending

Calculate P_{NSP} *(b, e)* where *b* → story body ; *e* → story ending

❑ **Perplexity Score**

Concatenate the story body and ending to form input to PLM

$X = \{x_{b_0}, ..., x_{b_B}, x_{e_0}, ..., x_{e_E}\}$

where *B* and *E* denote the number of tokens in the story body and ending

Calculate the conditioned perplexity by $$PPL(X) = \sqrt[E]{\prod_{i=1}^{E} \frac{1}{p(x_{e_i}|x_{b_0}, ..., x_{b_B}, ..., x_{e_{i-1}})}}$$

❑ Sample multiple (story body, story ending) pairs and use NSP and PPL to rank them*

**NSP the higher, the better. PPL(X) the lower, the better.*

Experiments

Models used:

- Official implementation of **DINO*** **(Schick and Schütze, 2021)** with the default GPT2-XL language model
  (k=30 for top-k sampling and blocking repeating trigrams during generation)

- For story ending ranking :
  HuggingFace *bert-base-uncased* checkpoint to calculate the NSP probability
  gpt2 (base) to calculate the perplexity

- Generate plot elements offline in batches and store them. When generating stories, randomly sample each type of plot element and combine them to form a content plan

*https://github.com/timoschick/dino

Experiments

Compared with following baselines:

- **Fusion (Fan et al., 2018)** : A seq2seq model with a convolutional encoder and a self-attention decoder generating stories conditioned on a prompt.

- **Plan-and-write (Yao et al., 2019)** : A bidirectional gated recurrent unit (BiGRU) seq2seq model that first predicts the storyline keywords from the title and then generates the story conditioned on both the title and the storyline

- **ProGen (Tan et al., 2021)** : A multi-stage BART seq2seq model using salient keywords as intermediate representations

- **GPT2-XL without content planning**

Results

Evaluation of different story ending ranking techniques

1. Pairwise comparison of different story ending ranking techniques – NSP, PPL and random selection

Content Plan	Location: San Francisco Cast: John Jones and Evelynn Genre: dark fantasy Theme: The specter of the future is in the telling
Random	*The protagonist is a woman who has become a cyborg. She must discover who she really is while dealing with the problems of a modern metropolis. John Jones is a mysterious person who claims to be an assassin who is also an inventor and an expert on the cyborgs and other strange and mysterious beings. Evelynn Jones is an attractive and brilliant woman who finds herself caught between the two worlds, and the protagonist's own conflicted past.* ***[SEP]*** *He had to be the hero! It was a tragedy that he couldn't be. He was a killer, an assassin. He had to do it and he did it with a smile on his face.*
NSP	*As the year comes to a close, two people begin to lose their grip on reality. Evelynn's father has been dead for several years, but her mother never truly accepted her fate and continues to live a life of denial. John Jones was once a successful businessman, but his fortune was spent by a mysterious corporation that he believed was his own. His life is about to take a turn for the worse as he discovers that Evelynn is living with a life-long secret.* ***[SEP]*** *In order for her to see the future, she'll have to take the risk.*
PPL	*In the past, John Jones used to be a normal person who worked for the government. But after a strange accident, he was taken to a secret facility, where he met the girl he loved, Evelynn, and started a relationship with her. But as the years passed, his memories started to grow more and more vague, and he started to realize that he didn't really remember how he got into that facility.* ***[SEP]*** *After a few months, John's memory returned to normal. He and Evelynn had their own children, but the memories remained.*

Table 2. Story body and ending selected by different algorithms

Results

Evaluation of different story ending ranking techniques

1. Pairwise comparison of different story ending ranking techniques – NSP, PPL and random selection

Method	Win:Lose
NSP vs. **Random**	19:31
PPL vs. Random*	35:15
PPL vs. NSP *	32:18

Table 3. Pair wise comparison of story ending ranking methods

- PPL selects more favourable story endings than NSP or Random
- NSP performs worse than random story and ending pairs. (can be attributed to the weakness of the NSP pre-training task)

Results

Evaluation of different story ending ranking techniques

2. Evaluate the story ending ranking methods on the Story Cloze Test dataset **(Mostafazadeh et al., 2016)**

- ❑ The dataset is created using crowd-sourcing to test models' commonsense story understanding
- ❑ Each story contains four preceding sentences, a 'right ending' and a 'wrong ending'
- ❑ Requirements to be selected as the right ending:
  - ❑ share at least one of the characters of the story in the ending
  - ❑ the ending sentence is entirely realistic and sensible when read in isolation

Results

Evaluation of different story ending ranking techniques

2. Evaluate the story ending ranking methods on the Story Cloze Test dataset

Method	Accuracy
Word2Vec	0.539
Skip-thought	0.552
DSSM	0.585
NSP	0.580
PPL	**0.587**

Table 4. The accuracy of various models on the Story Cloze test dataset. First three baselines taken from **Mostafazadeh et al., 2016**

- **Word2Vec** and **Skip-thoughts** predict the story ending whose embedding is nearest to the preceding story's embedding using word2vec and sent2vec embeddings separately
- **Deep Structured Semantic Model (DSSM) (Huang et al., 2013)** employs two jointly trained deep neural networks to project the preceding story context and the story ending into the same semantic space
- **NSP** and **PPL performed comparably** with **DSSM**
- **NSP** and **PPL** do **not require any in-domain data** or task-specific training unlike **DSSM**

Results

Human crowd sourced evaluation results for generated story quality based on different aspects

Model	natur	inter	cohes
Fusion	2.13	2.31	1.89
Plan-and-write	2.79	1.70	2.66
ProGen	2.13	3.05	1.88
SCRATCHPLOT	**4.04***	**3.99***	**3.47**
w/o content plan	3.64	3.19	3.41

Table 5. Human evaluation results of various models on different aspects - naturalness, interestingness and cohesiveness. All scores are on a scale from 1(worst) to 5(best).

- **SCRATCHPLOT** performed best on all aspects, the improvement in interestingness being especially pronounced
- **SCRATCHPLOT** outperformed the baseline w/o content planning in all aspects, demonstrating the contribution of content plans in story generation even when they contain noise

Results

Stories generated by the baselines for a sample content plot

Content Plot	
Location	Alameda County
Cast	John *and* Cheryl
Genre	Literary fiction
Theme	The only problem with the house was the people.

1. Fusion model

"What?" she whispered, as she pulled a small piece from her coat. It was a small piece of paper, with a picture of a young man sitting on that couch. " What?" he said, as he flipped the picture around. "My name is John, and I 'm here to give you a story." he said, as the photo of his wife came out of his hand and he pulled out a book. The picture was of some kind of book, with the Briggs on it. "What" she said, as she looked at it . " I don't know," John said, as he read the book. "It was about to get pretty bad. I don't like it here." "It's not good," she said. "What did you tell me?" he asked, as the picture was on the couch. "I just can't believe it

➢ **Fusion model** tends to generate stories that consist primarily of dialogues

Results

Stories generated by the baselines for a sample content plot

Content Plot	
Location	Alameda County
Cast	John *and* Cheryl
Genre	Literary fiction
Theme	The only problem with the house was the people.

2. Plan and Write

Intermediate content plan: *wanted; decided; bought; started; happy*

I wanted to go to the store. I decided to buy a new car. I bought a new car. I started to look for a new car. I was happy to have a new car.

- **Plan-and-write** generates short common-sense stories (like *ROCStories*), lacks diversity in sentence structure, do not have rich plots and characters

Results

Stories generated by the baselines for a sample content plot

Content Plot	
Location	Alameda County
Cast	John *and* Cheryl
Genre	Literary fiction
Theme	The only problem with the house was the people.

3. ProGen

Intermediate content plan: *object; active; hemperforming; fail; blinked; first; son; joined; wife; bedroom; boys; excited; boy; jerry;*

The first son I held inside my wife skin and body, I rested my head on her chest and prayed to the waiting Forbiddone, loved by the boy and girl alike, and that the house would bring smiles to all. Jerry Fairey, anxious and excited, pushed his face into my naked flesh. Jerry sounded just like his daddy and chirped, "Grr!" and they laughed together. Jerry took a deep breath, thought of again, "Fuck this. This place is empty, people." Jerry slept.

➢ **ProGen** trained on the *WRITINGPROMPTS* dataset generates stories that are often not logical

Results

Stories generated by the baselines for a sample content plot

Content Plot	
Location	Alameda County
Cast	John *and* Cheryl
Genre	Literary fiction
Theme	The only problem with the house was the people.

4. **S**CRATCH**P**LOT

John and Cheryl had been divorced for years, and while the divorce was final, there were still legal problems. John had bought a house in a nice part of town, and now, as it was approaching its five-year anniversary, his friends began asking questions about the house. The house was nice, and had nice rooms and a pool and all that. But the people who lived there had some problems. They lived with their mother and her family of five, with no money and no job. On the first day of the week of September, 1995, John and his friends got in their cars and drove back to the house they had been renting in the woods.

5. **S**CRATCH**P**LOT without content plan

An interesting tale of murder, betrayal, and the power of the human mind. (Sum- mary by Ransom, written by L. Ron Hubbard.) Ransom: "In the year 2392, an entire city has been destroyed and the entire population of a neighboring planet is now on the run. In this story, an ordinary man is caught up in a conspiracy that puts him on the edge of a world-changing discovery." (Plot summary by L Ron Hubbard, written to a friend.) Plot Summary: A plot summary of one of Hubbard's books. A summary from Hubbard's biography, Rise to Power.

Results

Automatic evaluation results for generated story quality based on diversity

Model	self-BLEU		distinct-n	
	n=1	n=2	n=1	n=2
Fusion	.805	.483	.079	.316
Plan-and-write	.960	.900	.056	.150
ProGen	.805	**.347**	.191	.683
SCRATCHPLOT	**.782**	.353	**.228**	**.705**
w/o content plan	.799	.380	.204	.650

Table 6. Result of self-BLEU scores to measure intra-story diversity (the lower the better) and distinct-n scores to measure repetitions (the higher the better).

- We measure *intra-story* lexical diversity using self-BLEU **(Zhu et al., 2018)** and *within-story* lexical diversity (or repetition) using *distinct-n*
- The baseline without explicit content planning generates less diverse stories because they are sampled by conditioning on the same instruction
- Plan-and-write often ignores the input theme completely and predicts common storylines similar to its training data, Therefore, it has much worse inter-/intra-story diversity
- **SCRATCHPLOT** generates stories with natural progression

Take-aways

- We introduced **SCRATCHPLOT**, a framework to perform unsupervised content planning for story generation using only pretrained language models (PLM)
- **SCRATCHPLOT** achieved strong results compared to supervised baselines fine-tuned on large parallel corpora and a PLM without access to content plans
- Code and data available at: https://github.com/YipingNUS/scratchplot-story-generation

THANK YOU !