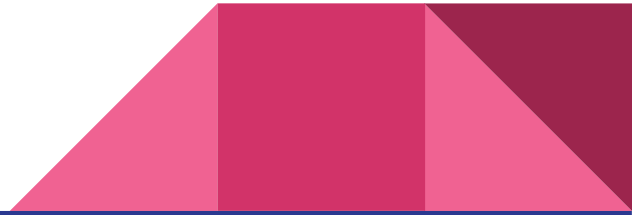
What do Large Language Models Learn about Scripts?

Abhilasha Sancheti, Rachel Rudinger

University of Maryland, College Park

Agenda

1. Summary
2. Methodology
3. Results
4. Strengths
5. Weaknesses
6. How it relates to Interactive Fiction/Story Generation?
7. Questions



Summary

- **Focus**: The paper explores the presence of explicit script knowledge in pre-trained generative language models like GPT-2, BART, and T5 to generate full **event sequence descriptions (ESDs)** with minimal prompting.
  .
- **Problem**: Through zero-shot probing, it is found that generative LMs produce poor event sequence descriptions (ESDs).

- **Solution**: The proposed **Script Induction Framework (SIF)**, yields substantial improvements over a fine-tuned LM, showing potential for inducing script knowledge.

Probing for Script Knowledge

Zero-shot Probing Experiment:
Designed to evaluate PLMs' ability to generate ESDs using carefully selected natural language prompts.
Experiment:
16 manually crafted prompts were used to probe GPT2, BART, and T5 for script knowledge.

Prompt Beginnings		Continuations
here is a sequence of events that happen while baking a cake:		None
these are the things that happen when you bake a cake:	⊕	1.
describe baking a cake in small sequences of short sentences:		1. get a cake mix
here is an ordered sequence of events that occur when you bake a cake:		1. get a cake mix 2. gather together other ingredients

Figure 2: Different prompt formulations for BAKING A CAKE scenario for probing. 16 prompts are created by combining a prompt beginning with a continuation.

Results:

- **BART and T5**: Unable to generate anything except the input tokens.
- **GPT-2 (GP)**: Generated some scenario-relevant events, but the ESDs were often incomplete, with auxiliary details and incorrect event ordering.

Conclusion:
A Script Induction Framework (SIF) is proposed due to poor quality ESDs generated in the zero-shot setting.

SIF: Script Induction Framework

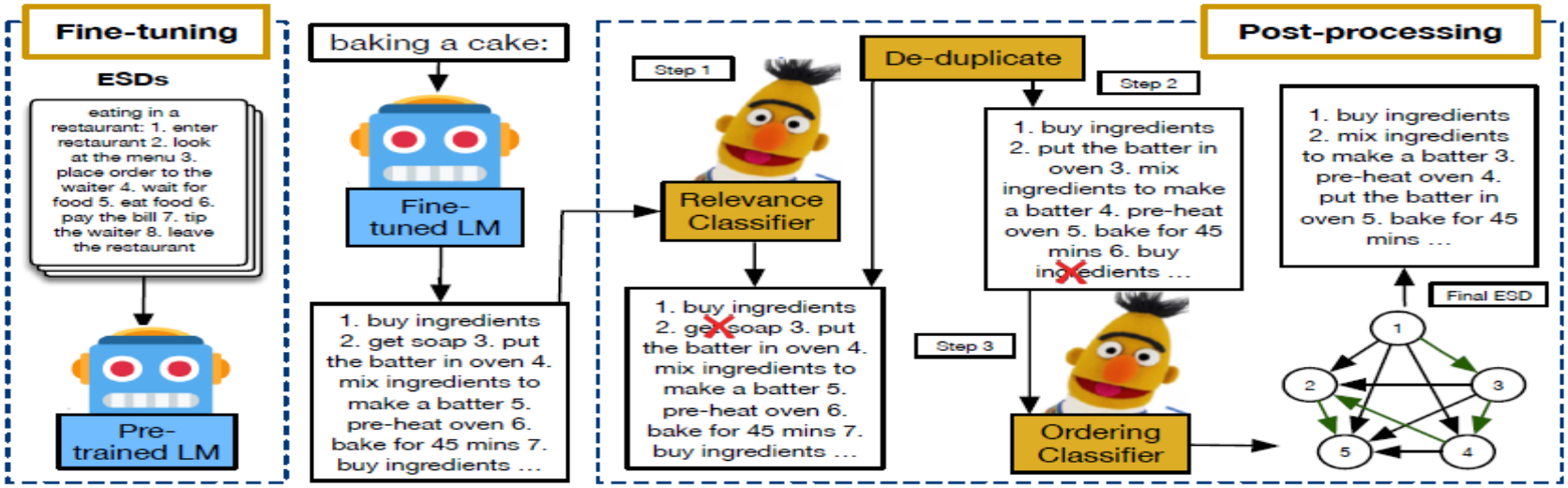


Figure 3: SIF: Pre-trained LM is fine-tuned on DeScript (Wanzare et al., 2016). Generated scripts are then post-processed with RoBERTa-based classifiers to correct for event relevance (Step 1), repetition (Step 2), and temporal ordering (Step 3).

Implementation Details:

- **DeScript Dataset**: Contains 100 event sequences (ESDs) for 40 scenarios.
- The data is split into **8 parts** for training and validation.
- Each ESD is lowercase and marked with:
  - **⟨BOS⟩**: Begin of scenario
  - **⟨EOS⟩**: End of scenario
- Input to classifiers:
  - **Relevance classifier**: scenario ⟨/s⟩ e (e.g., "baking a cake ⟨/s⟩ mix ingredients").
  - **Temporal classifier**: scenario name ⟨/s⟩ e1 ⟨/s⟩ e2 (e.g., "baking a cake ⟨/s⟩ mix ingredients ⟨/s⟩ preheat oven").

SEQUENCE here is a sequence of events that happen while baking a cake: 1. e_1 2. e_2
EXPECT these are the things that happen when you bake a cake: 1. e_1 2. e_2
ORDERED here is an ordered sequence of events that occur when you bake a cake: 1. e_1 2. e_2
DESCRIBE describe baking a cake in small sequences of short sentences: 1. e_1 2. e_2
DIRECT baking a cake: 1. e_1 2. e_2
TOKENS ⟨SCR⟩ baking a cake ⟨ESCR⟩: 1. e_1 2. e_2
ALLTOKENS ⟨SCR⟩ baking a cake ⟨ESCR⟩: ⟨BEVENT⟩ e_1 ⟨EEVENT⟩ ⟨BEVENT⟩ e_2 ⟨EEVENT⟩

Table 2: Different prompt formulations for BAKING A CAKE scenario with two events (e_1 and e_2).

Evaluation using BLEU Scores

Models	TOKENS	EXPECT	SEQUENCE	ALLTOKENS	DESCRIBE	DIRECT	ORDERED
(1) Zero-shot	03.1 (5.2)	03.6 (5.5)	05.4 (2.8)	03.1 (5.2)	03.2 (3.6)	03.9 (5.1)	06.2 (6.6)
(2) GPT2-$L_{SCRATCH}$	17.2 (3.1)	19.3 (3.7)	16.8 (2.9)	18.6 (4.5)	17.6 (2.6)	14.4 (3.9)	17.7 (3.2)
(3) BART-FT	15.5 (6.0)	20.8 (3.5)	19.6 (3.5)	19.7 (9.2)	19.2 (3.9)	18.0 (6.6)	11.7 (4.8)
(4) GPT2-FT	30.7 (5.1)	31.3 (5.5)	32.4 (6.3)	30.7 (6.6)	32.3 (5.9)	31.4 (5.8)	31.0 (4.8)
(5) BART-**SIF**	16.8 (5.1)	21.1 (4.2)	19.9 (3.7)	20.5 (11.1)	20.0 (3.8)	19.6 (7.2)	13.7 (5.0)
(6) GPT2-**SIF**	**33.6** (5.4)	**33.9** (5.6)	**35.2** (6.9)	**32.5** (6.9)	**34.2** (5.3)	**33.6** (5.7)	**33.2** (5.5)

Table 3: Automatic evaluation results: Mean BLEU scores (and std. dev.) over 8 folds of held-out scenarios are reported. (1) is pre-trained GPT2 (no fine-tuning or post-processing); (2) is randomly initialized GPT2 with fine-tuning; (3-4) are fine-tuned BART and GPT2; (5-6) are **SIF** applied to BART and GPT2.

Models	TOKENS	EXPECT	SEQUENCE	ALLTOKENS	DESCRIBE	DIRECT	ORDERED
(1) GPT2-FT	30.7 (5.1)	31.3 (5.5)	32.4 (6.3)	30.7 (6.6)	32.3 (5.9)	31.4 (5.8)	31.0 (4.8)
(2) GPT2-FT+Relevance (R)	33.1 (5.1)	33.1 (4.9)	34.7 (6.9)	31.9 (6.7)	33.7 (5.0)	32.6 (5.8)	33.2 (5.2)
(3) GPT2-FT+R+De-duplicate (D)	33.5 (5.2)	33.6 (5.2)	35.1 (6.9)	32.1 (6.7)	**34.3** (5.0)	32.9 (5.7)	**33.6** (5.5)
(4) GPT2-FT+R+D+Reorder (GPT2-SIF)	**33.6** (5.4)	**33.9** (5.6)	**35.2** (6.9)	**32.5** (6.9)	34.2 (5.3)	**33.6** (5.7)	33.2 (5.5)

Table 4: Ablation analysis of each step in the proposed pipeline for GPT2. Mean BLEU scores (and std. dev.) over 8 folds of held-out scenarios are reported. (1) fine-tuned GPT2; (2-4) are fine-tuned GPT2 with successive post-processing steps.

Manual Evaluation

Variants	BLEU↑	Manual Evaluation		
		R↑	O↑	M↓
TOKENS	19.2/**22.8**	77.2/**84.3**	72.3/**89.3**	**2.6**/**2.6**
EXPECT	22.8/**26.0**	81.9/**82.7**	74.5/**86.5**	**3.0**/**3.0**
SEQUENCE	27.8/**33.4**	73.3/**83.2**	74.0/**87.5**	**2.5**/**2.5**
ALLTOKENS	**33.5**/**35.0**	83.5/**85.7**	82.7/**89.5**	**2.6**/**2.6**
DESCRIBE	27.1/**28.6**	80.7/**86.3**	83.9/**85.9**	**2.8**/**2.8**
DIRECT	30.9/**34.1**	81.2/**84.2**	**88.5**/86.1	**2.6**/**2.6**
ORDERED	**31.9**/31.5	**84.9**/**86.2**	78.6/**86.8**	**2.6**/**2.6**

Table 5: Manual and BLEU scores on fine-tuned GPT2 (GPT2-FT) SIF applied to GPT2 (FT/SIF), computed for a stratified sample of outputs (one ESD per scenario across two folds). Mean scores across two annotators are reported. Annotator agreement is measured with Cohen's Kappa (Cohen, 1960) (κ=0.61 for **O**, κ=0.56 for **R**) and Spearman's correlation (ρ=0.64 for **M**). Underline and **bold** denotes the best across variants, and between FT and Ours, respectively. O scores are calculated only when both the events are marked as relevant by the two annotators.

Scenario	R↑	O↑	M↓
Order fastfood online	81.5	84.6	2.6
Cook in a microwave	89.5	92.0	2.4
Answer telephone	65.5	91.7	2.0
Buy from vending machine	77.1	81.3	3.4
Tie shoe laces	65.8	66.7	3.6
Brush teeth	75.9	71.4	2.6
Make ginger paste	41.5	85.7	3.4
Attend a wedding	71.9	100.0	2.4
Wash a car	85.7	90.0	3.0
Take out trash	88.5	92.3	2.2
Take a taxi	85.7	76.2	2.0
Surf the internet	73.3	62.5	2.8
Watch television	77.4	73.7	3.0
Go to a club to dance	100.0	93.5	1.4
Average Score	77.1	83.0	2.6

Table 6: Manual evaluation of ESDs for novel scenarios. Averaged across 5 sampled ESDs per scenario generated using the best performing SEQUENCE variant of GPT2-SIF as per automatic measure.

Results

		Good quality generations
GOING ON A TRAIN	FT	1. get dressed 2. go to station 3. buy ticket 4. get on train 5. wait for train 6. get on train 7. sit in seat 8. read newspaper 9. wait for train 10. get off train 11. get dressed 12. leave station
	SIF	1. go to station 2. buy ticket 3. wait for train 4. get on train 5. sit in seat 6. get off train 7. leave station

		Poor quality generations
CLEANING UP A FLAT	FT	1. get the flat 2. take the flat to the bathroom 3. take some water 4. take the flat to the sink 5. turn on the water 6. let it wash the flat 7. then put in some dryer sheet 8. get a sponge 9. scrub the flat 10. put the flat in the dryer 11. when done dry the flat with a towel 12. put away the flat
	SIF	1. get the flat 2. take the flat to the bathroom 3. take some water 4. take the flat to the sink 5. turn on the water 6. let it wash the flat 7. then put in some dryer sheet 8. get a sponge 9. scrub the flat 10. put the flat in the dryer 11. when done dry the flat with a towel 12. put away the flat

Table 7: Scripts generated using SEQUENCE variant of GPT2 for held-out scenarios. FT denotes output from the fine-tuned model and SIF refers to outputs from our framework applied to GPT2.

Results

		Good quality generations
BRUSHING TEETH	FT	1. go into bathroom 2. turn on faucet 3. remove any dirt or debris 4. grab a brush 5. scrub and floss the teeth 6. leave the bathroom
	SIF	1. go into bathroom 2. grab a brush 3. scrub and floss the teeth 4. leave the bathroom
		Poor quality generations
MAKING GINGER PASTE	FT	1. get your hot water 2. get your bowl 3. turn on the hot water 4. whisk a bowl of sugar into a paste 5. put the bowl on the stove 6. turn on the hot water 7. boil the paste 8. add salt to the paste 9. turn off the water 10. put the bowl on a rack 11. pour the hot water into a saucepan 12. put some salt and sugar in the saucepan 13. turn the heat on 14. pour the sauce on to the bowl 15. eat the paste
	SIF	1. get your hot water 2. get your bowl 3. turn on the hot water 4. whisk a bowl of sugar into a paste 5. put the bowl on the stove 6. boil the paste 7. add salt to the paste 8. put the bowl on a rack 9. pour the hot water into a saucepan 10. put some salt and sugar in the saucepan 11. turn the heat on 12. pour the sauce on to the bowl 13. eat the paste

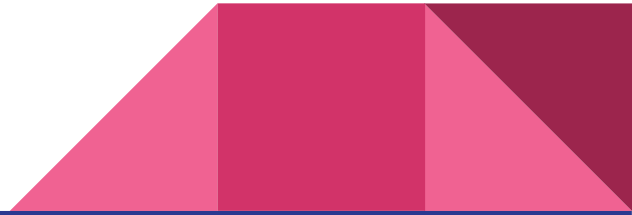
Table 8: Scripts generated using SEQUENCE variant of GPT2 for novel scenarios. FT denotes output from the fine-tuned model and SIF refers to outputs from our framework applied to GPT2.

Strengths

- **The use of Script Induction Framework (SIF)**
- **Language Model -Agnostic**
- **Comprehensive Evaluation**

Weakness

- **BLEU Metric Limitations**
- **Granularity Issues**
- **Generalization**
- **Paraphrase Handling**



How it relates to Interactive Fiction/Story Generation

Story Generation:

- SIF can be used to create scripts for stories or games where events need to happen in a specific order. For example, in a game, the designer can input a scenario like "character exploring a dungeon," and SIF can generate events like "find a key," "unlock a door," "fight a monster."

Interactive Fiction:

- In interactive fiction, SIF can create event sequences based on player decisions, where players make choices that lead to different outcomes.

THANK YOU!

QUESTIONS?