StoryAnalogy: Deriving Story-level Analogies from Large Language Models to Unlock Analogical Understanding

Cheng Jiayang[1], Lin Qiu[3], Tsz Ho Chan[1], Tianqing Fang[1], Weiqi Wang[1], Chunkit Chan[1], Dongyu Ru[3], Qipeng Guo[3], Hongming Zhang[1], Yangqiu Song[1], Yue Zhang[2], Zheng Zhang[3]

1 The Hong Kong University of Science and Technology

2 Westlake University 3 Amazon AWS AI

{jchengaj, yqsong}@cse.ust.hk zhangyue@westlake.edu.cn zhaz@amazon.com

Author's homepage

Introduction

S1: The *virus* invades *cells*. As a result, their *DNA* is damaged.

S2: The *burglar* breaks into the *house*. As a result, the *valuables* inside are smashed

Motivation

- Word-level analogies are well-studied. (e.g., *king* to *man* is like *queen* to *woman*)
- Understanding and reasoning on narrative level analogies is a crucial ability for intelligent agents, there has been limited research on this direction.

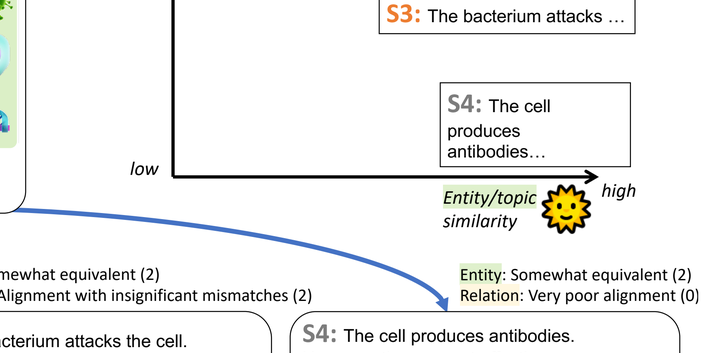
Key takeaways

- [Release of the StoryAnalogy dataset] We constructed a story analogies collection with 24k story pairs annotated on two dimensions of similarities.
- [LLMs are not so good at identifying story analogies]
  - Even the best-performing LLM still falls short of human performance by 37.7% in the correlation test.
  - LLMs can be easily distracted by negative choices with similar entities. This indicates that the models prioritize surface similarity over structural similarity, despite the latter being more important in identifying analogies.
- [Relational feature aware Encoder LMs are better at analogy search]
  - Encoder LMs like SimCSE or OpenAI-ada produce weak embeddings for analogy search.
  - In contrast, models aware of relational features perform better, such as RelBERT, GloVe-Verb, and Discourse Marker Representation.
  - Finetuning can help align the produced embeddings for analogy search.

How to evaluate the analogy level between a pair of stories?

Inspiration from cognitive psychology

- The Structure-Mapping Theory (SMT; Gentner, 1983: analogies between objects occur when
  - they have similar relational structures, but
  - have different attributes (e.g., *the hydrogen atom* v.s. *the Solar system*).



Extension on stories

Attributes => Entity similarity (EntSim): The similarity of entity and topics in stories.

Relational structures => Relation similarity (RelSim): The similarity of relational structures in stories.

- E.g. *predicates; logical connections between events*;

Modeling the analogy score (α)

According to SMT, analogy happens when the RelSim is high and the EntSim is low. So, a single score α that is proportional to the level of analogy between stories can be defined as

- α = RelSim / EntSim
- α = RelSim – EntSim

We include this score in the later experiments.

Dataset Overview

Collect seed data → Generate candidate story pairs → Crowd annotation

From four domains:

- **Scientific scripts:** ProPara
- **Social commonsense**: ROCStories
- **Word analogies**: SAT, U2, U4, SCAN
- **Knowledge graph**: ConceptNet

Seed data are fed into LLMs to generate story analogy candidate pairs

Recruit workers from AMT to assign score estimations.

Source story	Target story	Scores
The stream becomes a river. The river continues to flow along the same path for a long time.	A person grows from a child into an adult. As time passes, the person experiences ongoing growth and maturation.	: 0.6 : 2.8
They left him the key to the entrance. When Tom went over he realized it was the wrong key.	They gave her the password to the website. When Jane logged in, she realized it was the wrong password.	: 1.0 : 2.7
Foundations are poured to support the walls and roofs of buildings. The structure of the building is only as strong as it's foundation.	Reasons are formulated to make theories. The conclusions of theories are only as dependable as their initial premises.	: 0.6 : 1.8
His memory has broken into fragmented pieces. He can recall flashes and images of the past, but nothing concrete or clear.	His memories remain a confused mess. Nothing holds together and what he remembers don't make sense.	: 2.7 : 3.0

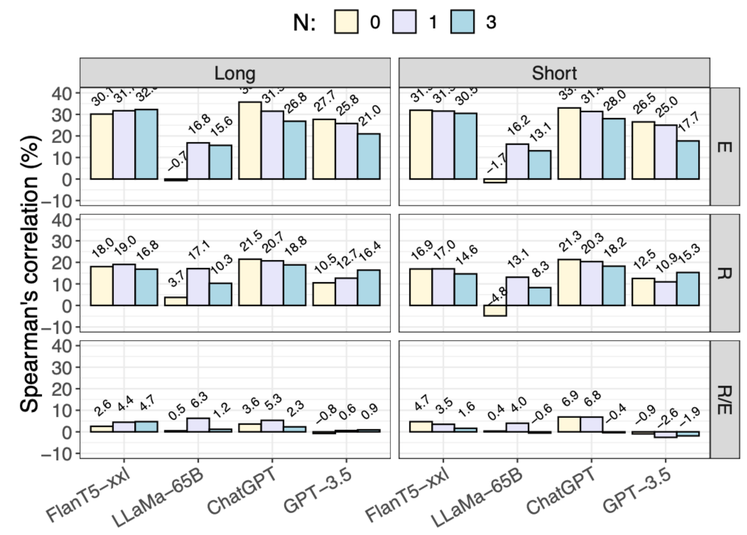
Experiment A: Story Analogy Identification

A1. Correlation test:

- Evaluate how well the score predictions correlate with the scores (EntSim, RelSim, α).
- Baselines:
  - Encoder models, including (1) general encoder LMs like RoBERTa, SimCSE, and (2) relational feature-aware models, e.g., RelBERT, GloVe-Verb, and Discourse Marker Representation (DMR)
  - LLMs, including FlanT5-xxl, LLaMa-65B, GPT-3.5, and ChatGPT.

	Mean		
	E	R	α
Random	0.0	0.0	0.0
Human	64.7	69.9	62.5
Encoder models			
RoBERTa	33.9	29.4	6.1
SimCSE	34.1	24.8	1.5
OpenAI-ada	37.2	26.9	0.7
DMR	23.5	28.7	11.4
RelBERT	27.2	24.9	5.3
GloVe-Noun	25.5	13.2	-3.3
GloVe-Verb	18.1	21.7	8.2
GloVe-All	27.4	18.0	-2.3
LLMs			
FlanT5-xxl	32.3	16.8	4.7
LLaMa-65B	15.6	10.3	1.2
GPT-3.5	21.0	16.4	0.9
ChatGPT	26.8	18.8	2.3
Finetuned models			
RoBERTa-Reg	24.2	30.5	17.0
RoBERTa-CL	28.7	43.2	25.0

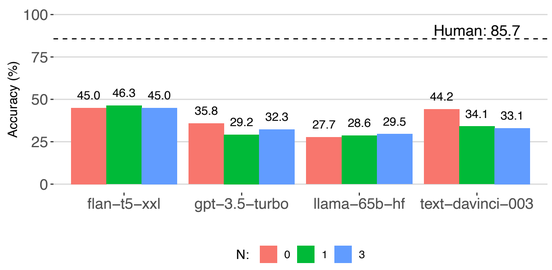
Left: The Spearman's rho correlation (%) between model predictions and scores from our dataset. E, R: EntSim, RelSim.



Right: The Spearman's rho (%) of LLMs, under different numbers (N=0, 1, 3) or types (Long/Short) of demonstrations.

A2. Multiple-choice test:

- The dataset is reframed as a multiple choice questions test.
  - Hard negatives are selected by sampling stories with similar entities (high EntSim), and easy negatives are chosen by random sampling.



	Target	Hard	Easy
Random	25.0	25.0	50.0
(Sultan and Shahaf, 2023)	44.9	17.8	37.2
FlanT5-xxl	45.4	37.2	17.4
LLaMa-65B	28.6	59.7	11.7
ChatGPT	32.4	59.5	8.1
GPT-3.5	37.1	55.8	7.1

Experiment B: Story Analogy Generation

- Examine whether the dataset can enhance the ability of analogy generation. The model generations are judged by human annotators to check their quality.

Setting	Model	Generation quality		
		Analogy	Novelty	Plausibility
Zero	FlanT5-xl	52.5	48.3	92.5
	FlanT5-xxl	46.7	49.2	92.5
	LLaMa-65B	38.3	39.2	**93.3**
	ChatGPT	70.0	72.5	90.8
	GPT-3.5	75.8	81.7	87.5
Few	FlanT5-xl	48.3	50.0	91.7
	FlanT5-xxl	40.0	43.3	85.0
	LLaMa-65B	66.7	66.7	92.5
	ChatGPT	**78.3**	**83.3**	86.7
	GPT-3.5	77.5	79.2	88.3
Tuned	FlanT5-xl	65.8	79.2	88.3
	FlanT5-xxl	72.5	81.7	86.7

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.