Understanding Story-Level Analogy

Contact: Dr. Daniel Bauer
bauer@cs.columbia.edu

1 Introduction

Analogy is a defining feature of human cognition (Gentner, 1989; Hofstadter and Sander, 2013), allowing us to make sense of new stimuli and ideas by mapping them to past experiences. For example, a young child might say "Look, I undressed the Banana". In NLP, analogical reasoning has mostly been explored at the level of simple word level analogies (Mikolov et al., 2013; Gladkova et al., 2016; Ushio et al., 2021).

Apple : Tree :: Grape : Vine
(Apple is to Tree, as Grape is to Vine).

A more challenging example are cross-domain (*far*) analogies:

Key : Lock :: Solution : Riddle

More complex analogies have received less attention. We are interested in **system analogies** between pairs of short narratives. For example, consider the following narratives:

Story 1: *Bob designed a well-organized project management spreadsheet and shared it with his colleagues. Jack changed the formulas and redesigned the charts. When Bob looked at the spreadsheet again, it was unusable.*

Story 2: *Marsha was cooking dinner for the night. She seasoned the food and left the room. Her husband walked into the kitchen and saw the food on the stove. He tasted it and added salt. When Marsha walked back into the room, she remembered that she forgot to season the food, so she added salt. The food was too salty and they had to order pizza that night.*

The two stories both describe a similar scenario in which multiple characters manipulate something, each with the intention of improving it, and end up ruining the result (matching the proverb "Too many cooks spoil the broth" – another analogy). Notably, the two stories take place in a different domain, but the characters and other elements map to each other to form the analogy (Bob ↔ Marsha, colleague ↔ husband, spreadsheet ↔ food, changing formulas and redesigning charts ↔ adding salt, etc.). The various elements participate in a *system* of interactions, which is mapped between the two stories.

Large language model excel at detecting and generating word level analogies,but even the best models seem to struggle with far (cross-domain) system analogies. In this project, we will explore different approaches to understanding and possibly generating story level analogy.

2 Data

Jiayang et al. (2023) develop the StoryAnalogy dataset, which includes 24.000 story pairs ("*The virus invades cells. As a result, their DNA is damaged.*" ↔ "*The burglar breaks into the house. As a result, the valuables inside are smashed.*"). The data is generated using LLMs on existing data sets (either stories or word analogies) and a number of manually created seed stories. Crowdsource workers filter the strength of the analogy and rate the strength of the analogy (on two axes: attribute similarity and relational similarity. They compute a final analogy score from these two values). Using this dataset, Jiayang et al. propose two different tasks: Predicting the analogy score and selecting the best analogy from a set of candidates (a multiple choice problem).

Sourati et al. (2024) introduce the Analogical Reasoning on Narratives (ARN) benchmark, which contains 1096 triples of the form (Q, A, N), where Q is a narrative, A is an analogous narrative, and N is a non-analogous narrative (distractor). The task is to classify which of the two candidate narratives

is the analogy. The stories tend to be longer and more complex than in the STORYANALOGY data. Consider, for example, the following additional story.

Story 3: *The woman worked as a teacher but took out loans to buy an expensive mansion and two fancy cars for herself. She bought designer clothing and fancy things. Eventually she went bankrupt because she couldn't afford the things she was buying on her teachers wage, which was only 40,000 a year.*

The benchmark task would require a model to select which of the two candidate stories (Story 2 or Story 3) is analogous to Story 1. Both A and N can be *near* or *far* analogies, resulting in four settings (nn, nf, fn, ff). The data is balanced between these settings.

3 Prior Results

Webb et al. (2023) evaluate the zero-shot performance of GPT-3 on a small set of narratives from (Gentner et al., 1993). While they identify analogical reasoning as a potential emergent property of LLMs, they find that humans outperform LLMs by a large margin, especially on far analogies.

Jiayang et al. (2023) evaluate a number of LLMs, as well as sentence encoders on their dataset. For the regression task (predicting analogy scores), they find that LLMs do not perform well. Pre-trained sentence encoders (using cosine similarity between the encoded stories) are unable to capture analogy even if they are relatively good at capturing textual similarity. However, fine-tuning is useful on this task. On the multiple-choice problem, they find that LLMs struggle with the task of selecting analogous stories (a 50% performance gab between human and LLM performance), especially in the presence of challenging distractors.

Similarly, Sourati et al. (2024) experiment with various LLM baselines on their benchmark (see above), using zero shot, few shot, and chain-of-thought prompting. They find that humans achieve accuracies of above 95%, while zero-shot and few-shot prompting achieved below 70% (even with the best available model, which was GPT-4.0). For the chain-of-thought experiment, they instruct the model to first extract the proverbial message from the source story *"Too many cooks spoil the broth"*. They report up to 30% improvement on far analogies in this setting, though still below human performance, and show limited benefit to near analogies compared to the zero-shot setting.

4 Project Outline

The goal of this project is to explore various techniques for analogy identification and possibly analogy generation. Following prior work you will initially explore one of three approaches:

- fine-tuning sentence encoders using contrastive learning. The learning objective would be to maximize the similarity between query Q and analogous narrative A, while minimizing the similarity between Q and a non-analogous distractor narrative N.
- Prompting based approaches, using LLMs with chain-of-thought reasoning. One approach here might be to ask the model to first identify the different elements of a narrative (objects/participants, their attributes, and relations between them), and then judge analogy based on this information.
- An approach based on explicit graph-based representations of the narrative, such as in Abstract Meaning Representation (AMR). We can also experiment with fine-tuning LLMs.

Additionally, we are interested in approaches to generating new analogies and potentially using such artificial data to boost training. Finally, we are especially interested in approaches that can *explain* analogies.

5 Future work / Applications

Analogy plays an important role in science and engineering, including computing, and maybe especially CS education. Consider common analogies like "a function is like a machine that converts arguments into return values", "algorithms are like cooking recipes", or "recursive functions are like a set of Russian nesting dolls". Future work could explore generating analogies to better explain CS concepts and provide student feedback as part of a tutoring system or interpreting analogies used by instructors and students. Analogy could also be used trying to understand non-literal speech, such as metaphor. Finally, analogy models could be used to investigate similarities between folk tales from different cultures (similar to the classifications in the Aarne–Thompson–Uther Index of folk tales)

6 Requirements

You should have completed COMS 4705 Natural Language Processing with a good grade. You should have excellent Python programming skills and be comfortable working with NLP data sets. You should have some experience fine-tuning transformer models and/or working with LLMs.

If you are interested, please reach out to bauer@cs.columbia.edu and mention relevant coursework, past research experience, and professional experience (if applicable).

References

D. Gentner, M.J. Rattermann, and K.D. Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, 25(4):524–575.

Dedre Gentner. 1989. The mechanism of analogical learning. In Stella Vosniadou and Andrew Ortony, editors, *Similarity and analogical reasoning*. Cambridge University Press.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *NAACL Student Research Workshop*, pages 8–15.

Douglas R. Hofstadter and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books, New York.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. 2024. Arn: Analogical reasoning on narratives. *Transactions of the Association for Computational Linguistics*, 12:1063–1086.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. Bert is to nlp what alexnet is to cv: Can pre-trained language models identify analogies? In *Proceedings of ACL*, pages 3609–3624.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.