Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models

2024.04.17

서승배

LMs stand their Ground: Investigating the Effect of Embodiment in Figurative Language Interpretation by Language Models

Philipp Wicke
Ludwig-Maximilians-University (LMU)
Institute for Information and Language Processing (CIS)
Munich Center for Machine Learning (MCML)
pwicke@cis.uni-muenchen.de

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

비유적 표현을 이해하기

Examples

The dataset is formatted as a Winograd schema. This means that sentences with the same beginning, but opposite meaning are paired together. This formatting was designed to reduce shortcut learning. Accuracy is calculated over all examples though, rather than for pairs.

Sentence	Correct Answer
The future is as bright as the sun	The future is bright
The future is as bright as ink	The future is not bright
The concert was as crowded as a rush-hour train	The concert was crowded
The concert was as crowded as a mausoleum	The concert was not crowded
Sleeping over at his house is like spending a night at the Waldorf Astoria	He has very nice accommodations
Sleeping over at his house is like spending a night at the Motel 6	He has below average accommodations

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Introduction

- 배경
  - 심리학에 따르면, 인간은 육체적 경험을 통해 획득한 비유적 표현을 사용한다.
  - 예를 들면, “She dances likes a turtle”라는 문장은 인간의 머릿속에서 거북이처럼 춤추는 이미지를 그리게 하고 “She dances poorly”라는 결론을 쉽게 낼 수 있음
  - 그러나 LLM은 육체적 경험을 통해 언어를 학습하지 않기 때문에 이는 큰 도전이 될 수 있음
  - 따라서 이 연구에서는 비유적 표현에 사용되는 신체화(embodiment) 정도에 따라 LLM의 비유적 표현 이해가 얼마나 영향을 받는지를 알아보려고 하였음
- 목표
  - 비유적 표현에 사용되는 단어(동사)의 신체화 정도에 LLM의 비유적 표현 이해가 영향을 받음을 보이고자 함
- 방법론
  - 통계 분석으로 다양한 파라미터 개수를 가지는 LLM에 대하여 신체화 정도에 따라 비유적 표현에 대한 이해의 성능(accuracy)에의 상관관계를 조사함
  - 통계 분석으로 비유적 표현에 사용되는 동사에서 신체화 외의 다른 변수(concreteness, 단어 습득 시기, 단어 빈도, 단어 길이) 등등의 성능에의 상관관계를 조사함
- 결과
  - Embodiment score는 1B이 넘는 파라미터를 가지는 LLM에서 유의미한 상관관계를 보였으며, 나머지 변수는 유의미한 상관관계를 보이지 않았음

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Method: Model Selection

- GPT-2, GPT-3, GPT-Neo, OPT 등등의 다양한 수의 파라미터를 가지는 모델을 선정하였음

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Method: Dataset

Examples

The dataset is formatted as a Winograd schema. This means that sentences with the same beginning, but opposite meaning are paired together. This formatting was designed to reduce shortcut learning. Accuracy is calculated over all examples though, rather than for pairs.

Sentence	Correct Answer
The future is as bright as the sun	The future is bright
The future is as bright as ink	The future is not bright
The concert was as crowded as a rush-hour train	The concert was crowded
The concert was as crowded as a mausoleum	The concert was not crowded
Sleeping over at his house is like spending a night at the Waldorf Astoria	He has very nice accommodations
Sleeping over at his house is like spending a night at the Motel 6	He has below average accommodations

- Fig-QA 데이터셋

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Method: Embodiment Scoring

(A) *The pants were as faded as ...*

Embodiment Rating: 2.36

(B) *She dances like a ...*

Embodiment Score: 6.50

- Sidhu et al.(2014)에서 687개의 영어 동사에 대하여 실험 참여자들에게 각각의 동사가 얼마나 신체와 관련되어 있는지를 평가하게 하여 embodiment score를 매김
- Fig-QA 데이터셋 중에서 동사가 Sidhu et al.의 687개의 동사 안에 포함되는 데이터를 선택해서 데이터셋의 각각의 문장에 대하여 embodiment score를 매김
- 이렇게 embodiment 점수를 매긴 데이터를 LLM에 대하여 성능을 측정하고, 그 성능과 embodiment score의 상관관계를 조사함

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Method: Concreteness Scoring

- Rotaru et al.(2020)의 연구에 기반하여 concreteness-in-context predictor를 구성하여 embodiment score와 같이 각각의 동사에 대해 concreteness score를 매김

- 마찬가지로 concreteness score와 LLM의 비유적 표현에 대한 이해 성능(accuracy) 사이의 상관관계를 조사함

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Method: Statistical Tests

- Concreteness에 추가로 단어의 습득 연령(AoA), 단어 빈도, 단어 길이 등의 다른 변수들도 측정하고, embodiment score를 포함한 모든 변수에 대하여 성능에의 회귀분석을 진행함

- 각 feature들의 VIF를 조사하여 다중공선성을 고려해야 하는지를 확인함

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Results: Embodiment Correlation

Model	Accuracy on C_{Emb}	p-Value	Correlation coefficient	* p <.05 ** p <.01
GPT-3 (small)	0.594	0.018	0.062	*
GPT-3 (large)	0.667	0.034	0.056	*
OPT (small)	0.561	0.206	0.033	
OPT (medium)	0.627	0.034	0.056	*
GPT-Neo (small)	0.535	0.399	0.022	
GPT-NeoX (medium)	0.648	0.005	0.073	**
GPT-2 (small)	0.597	0.158	0.037	
GPT-2 XL (medium)	0.606	0.009	0.069	**

Table 3: Experimental results of all model pairs (small and larger versions) on the C_{Emb} corpus. The last column marks significant results of the point biserial correlation between embodiment score and model performance for $p < 0.05$ with * and for $p < 0.01$ with **.

- 큰 모델이 작은 모델보다 항상 나은 성능을 보임
- 1B를 넘는 모든 모델에서 embodiment score와 p값 사이에 유의미한 상관관계가 나타났으며(p<0.05) 몇몇 모델에서는 p<0.01을 나타냄
- 이는 큰 모델에서 embodiment score가 LLM의 비유적 표현 이해에 관련이 있음을 나타냄.

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Results: Concreteness

Model	GPT-3	GPT-3	OPT	OPT	GPT-Neo	GPT-NeoX	GPT-2	GPT-2 XL
Parameters	350M	175B	350M	13B	125M	20B	355M	1.5B
Correlation	-0.016	-0.039	-0.037	-0.032	-0.020	-0.026	-0.036	-0.003
p-value	0.554	0.139	0.163	0.227	0.448	0.318	0.176	0.921

Table 6: Results of the point biserial correlation between the concreteness of action in context and performance of LM on the interpretation of figurative language phrases (C_{Emb}). None of the LMs shows a statistically significant correlation between the variables ($\alpha < 0.05$).

- 모든 모델에 대하여 concreteness는 유의미한 상관관계를 보이지 않았다.

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Results: Regression, Variance Inflation

AoA	Word Frequency	Embodiment Rating	Word Length	Constant
1.610	1.345	1.326	1.017	86.387

Table 4: VIF from the four features. Constant denotes the intercept provided for the VIF. A factor close to 1 indicates no correlation with values above 4 regarded as moderate correlation.

- 모든 모델에서 embodiment 변수가 포함되었을 때 R^2 값이 일관적으로 증가
- VIF 값은 다중공선성을 고려하지 않아도 됨을 나타내고 있음

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Results: Interpretation

- 1B 파라미터 이상의 모든 상대적으로 큰 모델에서 embodiment에 대해 유의미한 상관관계가 나타나며, 모델의 크기가 클 수록 좋은 작업 성능과 embodiment의 효과가 크게 나타났다.

- Concreteness, AoA(언어의 습득 시기), 단어 길이, 단어 빈도에 대해서는 유의미한 상관관계가 나타나지 않았으며, 이는 embodiment가 단어(동사)의 여러 feature들 중에 비유적 표현을 해석하는 데에 있어서 중요함을 나타냄

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

Conclusion and Discussion

- LLM의 비유적 표현 해석에 있어서 embodiment가 중요함을 알아내었다.
- 인간의 자연어 이해는 환경과의 물리적 상호작용에 의해 실체화되므로, 신체화된 행동에 의존하는 은유적 언어의 해석에서는 LLM이 어려움을 겪을 것으로 예상할 수 있지만, 오히려 좀더 신체화된 어휘일수록 비유적 표현의 해석에 있어서 성능이 높아짐.
- Fig-QA에만 실험하였기 때문에, 후에는 BIG-bench 등 다른 언어 해석 작업에서도 embodiment가 유효한지 검증하고자 함
- Embodiment의 영향을 검증하기 위해 좀더 많은 연구가 필요하다.

Lms stand their ground: Investigating the effect of embodiment in figurative language interpretation by language models (2023.05)

HCC LAB SNU

QnA