The Distributional Hypothesis Does Not Fully Explain the Benefits of Masked Language Model Pretraining

Ting-Rui Chiang, Dani Yogatama
University of Southern California
{tingruic,yogatama}@usc.edu

tamagotchi lab
@USC

The Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings (Harris, 1954)

- i.e. words semantically similar have similar distribution of neighbor words

$$P(x_1, x_2, \ldots, x_n \mid \text{“delicious”}) = P(x_1, x_2, \ldots, x_n \mid \text{“tasty”})$$

(*the distributional property*)

Historically

- It has been used to explain the efficacy of word embedding training.
- and also [1]

[1] Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

The distributional property encodes semantics

It connects semantics to data distribution:

$$\text{“delicious”} = \text{“tasty”}$$

$$P(x_1, x_2, \ldots, x_n \mid \text{“delicious”}) = P(x_1, x_2, \ldots, x_n \mid \text{“tasty”})$$

The distributional property infuses semantic relationships in the pretrained model.

Theoretically

Knowledge about semantic relationships improves

- sample efficiency
- generalization capability

It’s nice but…

- It assumes we use pretrained models as static models. (and so do many existing probing works)
- Does the distributional property really help the fine-tuning process? 🤔

Validating with Synthetic Data

Experimental Design

Hypothesis: $\boxed{\text{the distributional property}}$ in the pretraining data

1. improves the sample efficiency of downstream tasks
2. helps fine-tuned models generalize better

The experimental variable we want to manipulate.

With this pseudo language:

dataset **w/** the property → pretrained with MLM → fine-tuned & test

dataset **w/o** the property → pretrained with MLM → fine-tuned & test

Results

- Sample efficiency: yes
- Generalization capability: no

Experimenting with Real-world Data

Premise of the Experiment (informal)

If the distributional hypothesis is the explanation, then whether a fine-tuned model f generalizes, e.g. knowing

$$f(\text{“It is delicious”}) = f(\text{“It tastes good”}), \quad (1)$$

should be related to whether the pretrained model f_0 models this distributional property well

$$f_0(x_1, x_2, \ldots, x_n \mid \text{“is delicious”}) = f_0(x_1, x_2, \ldots, x_n \mid \text{“tastes good”}). \quad (2)$$

Inspecting the correlation between whether (1) and (2) are true.

Experimental Design

paraphrase

feature 1 → feature 2

is delicious → tastes good

tastes bad → distasteful

…

fine-tuned model f

$\mathbf{D}[\,f(\mathrm{y} \mid \text{“It is delicious”}) \;\|\; f(\mathrm{y} \mid \text{“It tastes good”})\,]$

correlation

pretrained model f_0

e.g. bert-based-uncased

$\mathbf{D}[\,f_0(\texttt{[mask]} \mid \text{“is delicious”}) \;\|\; f_0(\texttt{[mask]} \mid \text{“tastes good”})\,]$

Inferring the semantic relationship in an MLM - word & phrase

$$f_0(\text{ctx} \mid \textit{feature 1}) = f_0(\text{ctx} \mid \textit{feature 2})$$

For words and phrases: query with POS-dependent templates

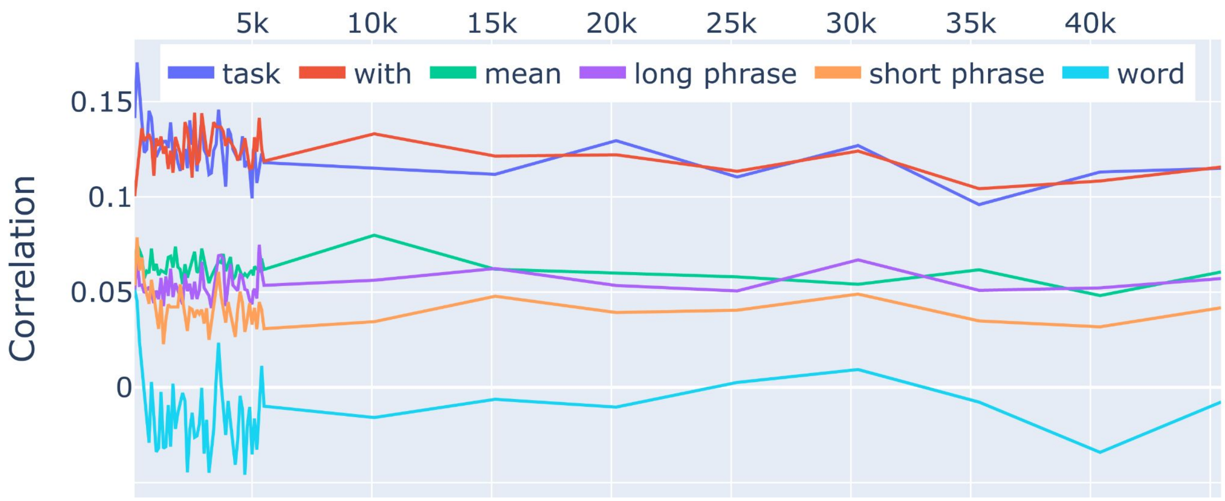
`{NP} [MASK]`
e.g. *a running car* [MASK]

`[MASK] {VP}`
e.g. [MASK] *is chased by a dog*.

`[MASK] is {ADJP}`
e.g. [MASK] *is well-made and lovely*.

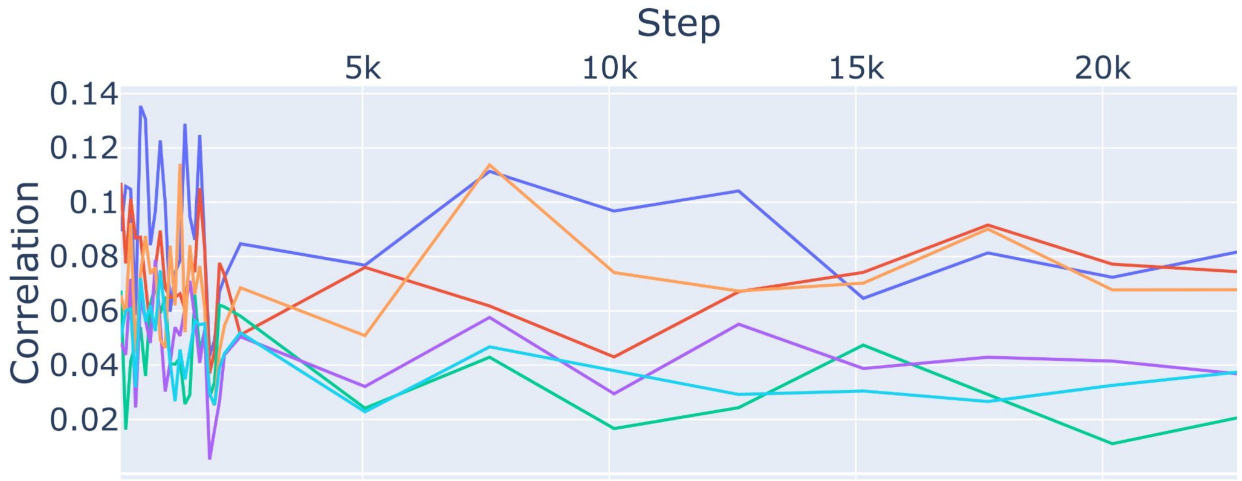
This should be a natural way to query the relationship from an MLM

MNLI



5k
10k
15k
20k
25k
30k
35k
40k
task
with
mean
long phrase
short phrase
word
0.15
0.1
0.05
0
Correlation

SST2



Step
5k
10k
15k
20k
0.14
0.12
0.1
0.08
0.06
0.04
0.02
Correlation

Conclusion

- The distributional property contributes to better sample efficiency.
- But it doesn't explain the generalization capability.

tamagotchi lab
@USC