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

Ting-Rui Chiang   Dani Yogatama

tamagotchi lab @USC

The Distributional Hypothesis

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

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

the distributional property

This property is nice

- It connects the pretraining objective to word semantics.
- It has been used to explain the efficacy of word embeddings.

Theoretical Analyses

○ apple tastes good | ○ sushi is delicious | × soba tastes bad | × orange is bland

A task in your mind

Semantic relationship can improve sample efficiency:

○ 2 3 5 | ○ 11 4 6 | × 12 3 7 | × 1 4 8

What a machine sees

○ 2 **3 5** | ○ 11 **4 6** | × 12 **3 7** | × 1 **4 8**

Utilizing the semantic

Semantic relationships can help generalization:

Test set: It is tasty. | 0 4 9 | 0 **4 9**

But these analyses assume that we use a pretrained model as a ***static*** feature extractors.

Experimenting with Real-world Data

Research question:

Dose *the distributional property explain* generalization?

The premise of the experiment:

If a fine-tuned model f generalizes, e.g. knowing

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

because of the relationship encoded in the distribution property

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

then the pretrained model f_0 should model this distributional property well

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

Thus, we measure the correlation between (1) and (2).

Step 1: Perturb features in examples

(noisy) paraphrase	
feature1 → feature2	is delicious → tastes good
	tastes bad → distasteful
	…

Step 2: Measure (1) by inferring the fine-tuned model f

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

Step 3: Measure (2) by inferring the pretrained model f_0

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

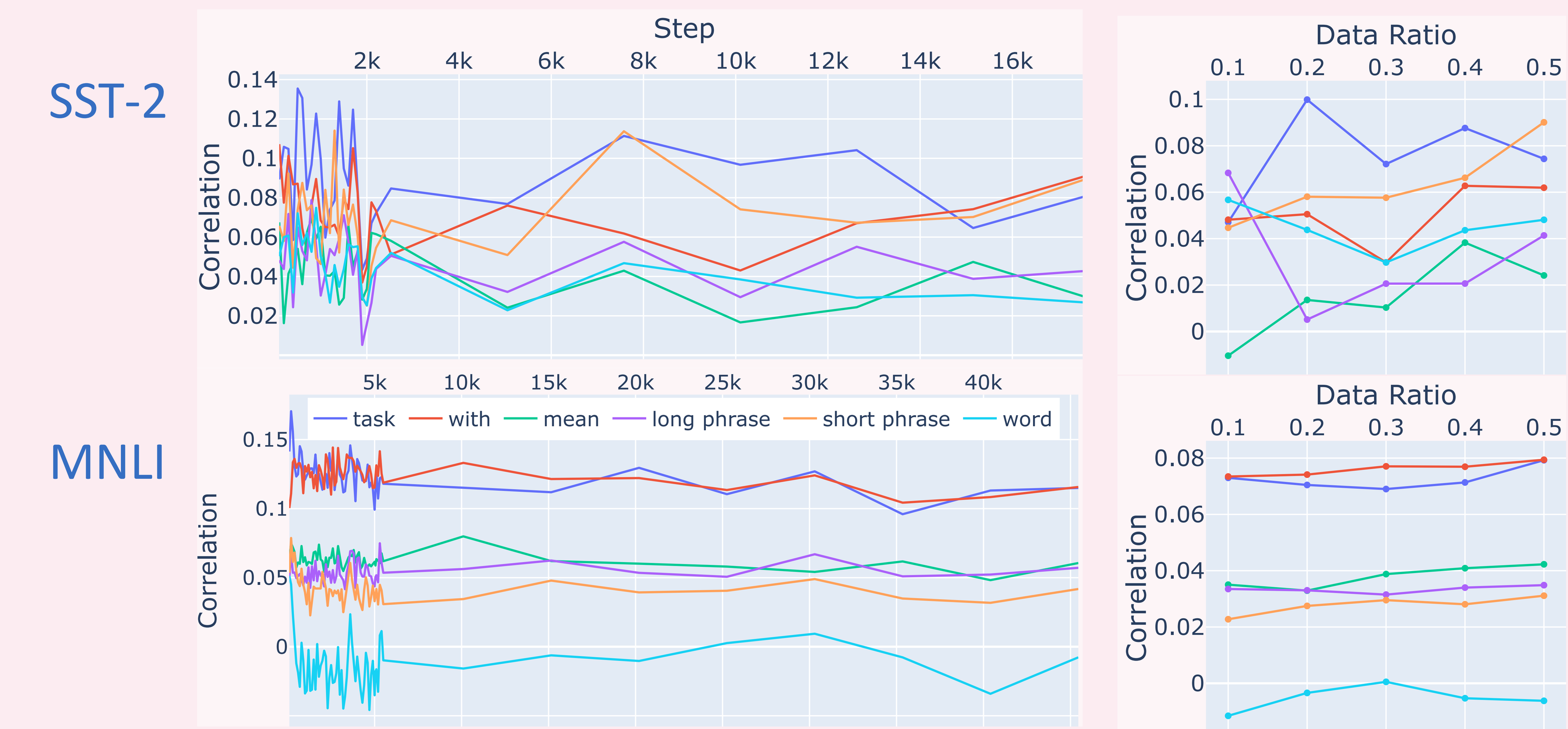
For word-level and phrase-level features: query with POS-dependent templates

{NP} [MASK]	[MASK] {VP}	[MASK] is {ADJP}
e.g. *a running car* [MASK]	e.g. [MASK] *is chased by a dog.*	e.g. [MASK] *is well-made and lovely.*

For sentence-level features:

{sentence} with [MASK]	“{sentence}” means [MASK]
e.g. *This is a novel paper with* [MASK]	e.g. *“This is a novel paper” means* [MASK]

Step 4: Compute the correlation



The distribution property does not explain generalization.

Experimenting with Synthetic Data

Research question:

Dose *the distributional property* helps fine-tuning?

Step 1: Define a pseudo-language.

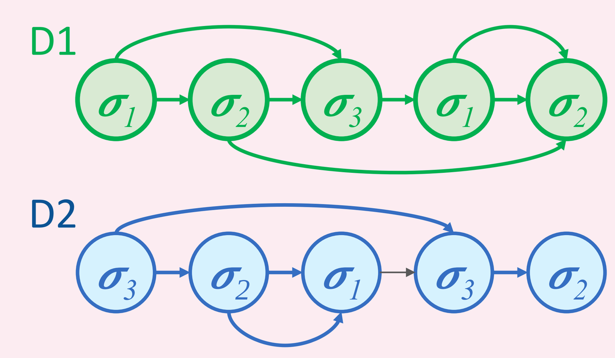
synsets (sets of synonyms)

$$\sigma_1 = \{a_1 \leftrightarrow b_1\}$$
$$\sigma_2 = \{a_2 \leftrightarrow b_2\}$$
$$\sigma_3 = \{a_3 \leftrightarrow b_3\}$$

A B

Two isomorphic vocabulary sets

Two distributions of synset sequences



Step 2: Generate data for pretraining

$\sigma_1, \sigma_2, \sigma_3, \sigma_1, \sigma_2, \ldots$ | $\sigma_3, \sigma_2, \sigma_1, \sigma_3, \sigma_1 \ldots$

map synsets to sequences without the distributional property

$a_1, a_2, a_3, a_1, a_2, \ldots$ | $b_3, b_2, b_1, b_3, b_2, \ldots$

map synsets to sequences with the distributional property

$a_1, b_2, a_3, b_1, a_2, \ldots$ | $a_3, b_2, a_1, b_3, b_2, \ldots$

Step 3: Define a downstream task

The label is ***True*** if the underlying synsets matches some predefined patterns such as

$$\sigma_1 * * \sigma_2 * * * \sigma_6$$

otherwise, the label is ***False***.

Step 4: Pretrain and fine-tune models

Fine-tune with the mixture of two vocabulary sets.

Pretrained with the distribution property (w/DH) improve sample efficiency!

Fine-tune with only one vocabulary set.

Pretrained with the distribution property (w/DH) does not help generalization.

Conclusion:

- **The Distributional Hypothesis explains pretrained models’ better sample efficiency.**
- **But it does not explain the generalization ability.**

follow me on X   paper on arxiv

USC NLP

USC University of Southern California