

# How Relevant Are Selectional Preferences for Transformer-based Language Models?

## Previous Work

BERT's linguistic abilities (via learned embeddings):

- ▶ Syntactic: knowledge of parts-of-speech & roles, dependencies, hierarchical structure
- ▶ Semantic: knowledge of roles, entity types, relations, but can't generalize!
- ▶ World knowledge: bad at inference, biases

But is this **profound knowledge** or **frequency-based**?

## Selectional Preferences

- ▶ "The athlete runs a marathon" = **felicitous** (run + athlete) + (run + marathon)
- ▶ "The bassoon runs a banana" = **infelicitous** (run + bassoon) + (run + banana)

## Our corpus

**SP-10K<sup>1</sup> corpus:** 2,5K freq. words → 10K dependency word pairs + **plausibility score:** degree of felicity

One-hop syntactic dependencies:

- ▶ **nsubj:** head/verb + dep./noun/subject
- ▶ **dobj:** head/verb + dep./noun/direct object
- ▶ **amod:** head/noun + dep./adjective/modifier

Two-hop syntactic dependencies:

- ▶ **nsubj\_amod:** head/verb + dep. to nsubj/adj./mod.
- ▶ **dobj\_amod:** head/verb + dep. to dobj/adj./mod.

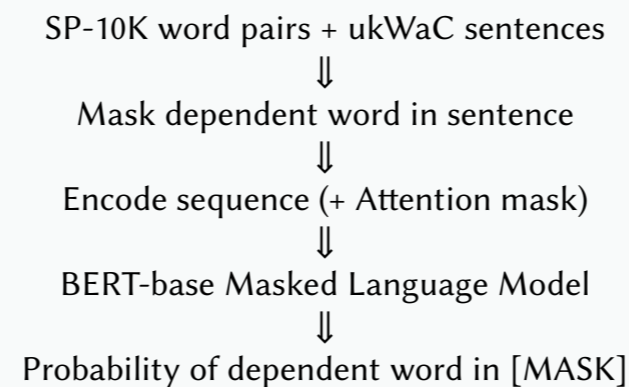
Find the word pairs in parsed **ukWaC<sup>2</sup>** sentences.

<sup>1</sup> Zhang et al. (2019) <sup>2</sup> Ferraresi et al. (2008)

## Our research question

Do BERT encodings capture the **selectional preferences** of a head word for its dependents?  
**Correlation probability-plausibility**

## Methodology



## Number of sentences per category

TYPE	WORD PAIRS	SENTS	AVG. PLAUS. SCORE
<b>nsubj</b>	958 / 2,000	30,526	6.64
<b>dobj</b>	980 / 2,000	56,777	7.39
<b>amod</b>	1,030 / 2,000	23,110	7.62
<b>nsubj_amod</b>	956 / 2,061	12,911	5.75
<b>dobj_amod</b>	922 / 2,063	21,839	6.32
TOTAL	4,846 / 10,124	181,867	145,163

## Attention mask

SENTENCE:	the	film	tells	the	story		
<b>standard</b>	[CLS]	the	film	tells	the	[MASK]	[SEP]
<b>head</b>	[CLS]	the	film	■■■	the	[MASK]	[SEP]
<b>context</b>	[CLS]	■■■	■■■	tells	■■■	[MASK]	[SEP]
<b>control</b>	[CLS]	■■■	■■■	■■■	■■■	[MASK]	[SEP]

## Results

TYPE	STANDARD	HEAD	CONTEXT	CONTROL
<b>nsubj</b>	0.03	-0.02	0.16	-0.01
<b>dobj</b>	0.05	-0.07	0.05	-0.05
<b>amod</b>	0.04	-0.06	0.24	-0.04
<b>nsubj_amod</b>	-0.01	-0.13	0.29	-0.00
<b>dobj_amod</b>	0.06	0.01	-0.03	0.02

Micro-averaged

TYPE	STANDARD	HEAD	CONTEXT	CONTROL
<b>nsubj</b>	0.19	0.15	0.29	0.08
<b>dobj</b>	0.16	0.04	0.27	0.05
<b>amod</b>	0.15	0.03	0.35	0.03
<b>nsubj_amod</b>	0.01	-0.04	0.22	0.06
<b>dobj_amod</b>	0.14	0.10	0.20	0.07

Macro-averaged

## Findings

- ▶ **No strong correlation!** (<-0.4 or >0.4)
- ▶ **nsubj, amod** strongest, two-hop lowest
- ▶ **Context mask > No mask > Head mask** ⇒ Head word strongly influences probability of dependent word, context dilutes probability ⇒ **Selectional preferences are present!**
- ▶ Head word also affected **two-hop relations!**
- ▶ Head word categories/classes? Not discernible.
- ▶ Dependent word categories/classes? Unclear.
- ▶ BERT: **high frequency** = high probability
- ▶ **Problems:** implausible word pairs never found, problematic SP-10K annotation, BERT tokenization