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[1] 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**[2] sentences.

[1] Zhang et al. (2019) [2] 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

SP-10K word pairs + ukWaC sentences
⇓
Mask dependent word in sentence
⇓
Encode sequence (+ Attention mask)
⇓
BERT-base Masked Language Model
⇓
Probability of dependent word in [MASK]

Number of sentences per category

Type	Word pairs	Sents	Avg. plaus. score
nsubj	958 / 2,000	30,526	6.64
dobj	980 / 2,000	56,777	7.39
amod	1,030 / 2,000	23,110	7.62
nsubj_amod	956 / 2,061	12,911	5.75
dobj_amod	922 / 2,063	21,839	6.32
Total	4,846 / 10,124	181,867	145,163

Attention mask

Sentence:		the	film	tells	the	story	
standard	[CLS]	the	film	tells	the	[MASK]	[SEP]
head	[CLS]	the	film	■■■	the	[MASK]	[SEP]
context	[CLS]	■■■	■■■	tells	■■■	[MASK]	[SEP]
control	[CLS]	■■■	■■■	■■■	■■■	[MASK]	[SEP]

Results

Type	standard	head	context	control
nsubj	0.03	-0.02	0.16	-0.01
dobj	0.05	-0.07	0.05	-0.05
amod	0.04	-0.06	0.24	-0.04
nsubj_amod	-0.01	-0.13	0.29	-0.00
dobj_amod	0.06	0.01	-0.03	0.02

Micro-averaged

Type	standard	head	context	control
nsubj	0.19	0.15	0.29	0.08
dobj	0.16	0.04	0.27	0.05
amod	0.15	0.03	0.35	0.03
nsubj_amod	0.01	-0.04	0.22	0.06
dobj_amod	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

Eleni Metheniti [♣♦], **Tim Van de Cruys** [♠], **Nabil Hathout** [♣]

firstname.lastname@{univ-tlse2.fr[♣], irit.fr[♦], kuleuven.be[♠]}

CLLE-CNRS [♣], IRIT (CNRS)[♦], KU Leuven [♠]