Visualizing and Measuring the Geometry of BERT

Accepted at NIPS

Coenen et al, 2019 from Google Brain

Ganwoo Kim

Data Mining & Information Systems Lab.
Department of Computer Science and Engineering,
College of Informatics, Korea University

0 Introduction

How does BERT represent useful linguistic information internally?

Three main explorations – Probing tasks

Syntactic

1. Attention matrices contain grammatical representations

2. Relations with parse tree and hidden representations
   + Visualization

Semantic

3. BERT representation also has the information of word sense
   + Visualization
   + Measurement

1 Attn. with grammar

How does BERT represent useful linguistic information internally?

Three main explorations

Syntactic

1. Attention matrices contain grammatical information

2. Relations with parse tree and hidden representations
   + Visualization

Semantic

3. BERT representation also has the information of word sense
   + Visualization
   + Measurement

1 Attn. with grammar

1. Attention matrices contain grammatical information

An attention probe

- Goal : to classify a given **dependency relation** between two tokens
- Model : a linear model
- Input : *a model-wide attention vector*, formed by concatenating entries in every attention matrix from every head and layer.

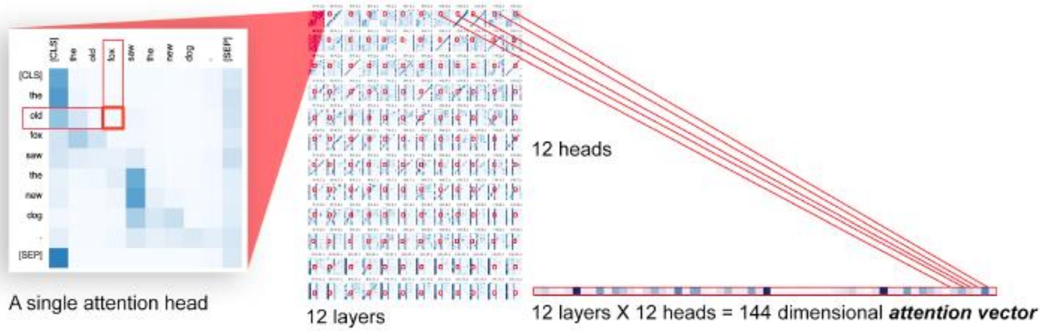


Figure 1: A *model-wide attention vector* for an ordered pair of tokens contains the scalar attention values for that pair in all attention heads and layers. Shown: BERT-base.

1 Attn. with grammar

6.5 Dependency relation performance

Dependency	precision	recall	n
advcl	0.34	0.08	1381
advmod	0.32	0.32	6653
amod	0.68	0.48	10830
aux	0.64	0.08	6914
auxpass	0.68	0.50	1501
cc	0.84	0.77	5041
ccomp	0.67	0.78	2792
conj	0.64	0.85	5146
cop	0.49	0.16	2053
det	0.81	0.95	15322
dobj	0.74	0.66	7957
mark	0.58	0.67	2160
neg	0.83	0.17	1265
nn	0.67	0.82	11650
npadvmod	0.53	0.23	580
nsubj	0.72	0.83	14084
nsubjpass	0.30	0.14	1255
num	0.82	0.55	3464
number	0.77	0.74	1182
pcomp	0.14	0.01	957
pobj	0.78	0.97	17146
poss	0.74	0.54	3567
possessive	0.83	0.86	1449
prep	0.79	0.92	17797
prt	0.67	0.33	593
rcmod	0.55	0.30	1516
tmod	0.55	0.15	672
vmod	0.84	0.07	1705
xcomp	0.72	0.40	2203
all	**0.72**	**0.72**	**150000**

Table 2: Per-dependency results of multiclass linear classifier trained on attention vectors, with 300,000 training examples and 150,000 test examples.

Detail

- … run each sentence through BERT-base and obtained the *model-wide attention vector*
- First Probe : Binary classification about the existence of a dependency relation
  Second probe : Multiclass classification for predicting which type of dependency relation

Results

- The first probe achieved an accuracy of 85.8%
- The second probe achieved an accuracy of 71.9%

1. **Attention matrices contain grammatical information => Proved!**

2 BERT rep. with parse tree

How does BERT represent useful linguistic information internally?

Three main explorations

Syntactic

1. Attention matrices contain grammatical information

2. Relations with parse tree and hidden representations
   + Visualization

Semantic

3. BERT representation also has the information of word sense
   + Visualization
   + Measurement

2 BERT rep. with parse tree

2. Relations with parse tree and hidden representations

An structural probe (J. Hewitt and C. Manning, ACL 2019)

- Goal : to approximate the parse tress distance with the distance of hidden rep.
- Method : a linear transformation of the Euclidean distance between two rep.

$$d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 = \left(B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell)\right)^T \left(B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell)\right) \quad (1)$$

, where l is the sentence and B is a learnable parameter

$$\min_B \sum_\ell \frac{1}{|s^\ell|^2} \sum_{i,j} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 \right|$$

, where $|s^l|$ is the sentence length

$d_T(u, v) = 1$, if u, v are neighbors in parse tree

2 BERT rep. with parse tree

An structural probe (J. Hewitt and C. Manning, ACL 2019)

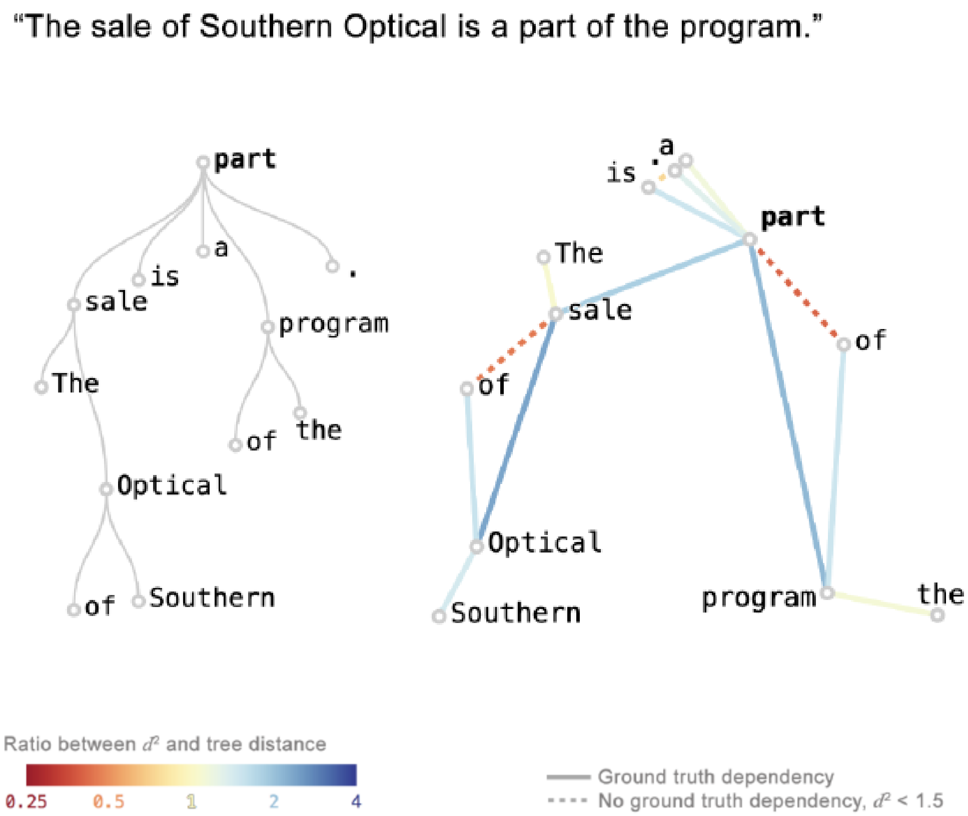


Figure 2: Visualizing embeddings of two sentences after applying the Hewitt-Manning probe. We compare the parse tree (left images) with a PCA projection of context embeddings (right images).

$$d_T(u,v) = 1$$

$$d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 = \left(B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell)\right)^T \left(B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell)\right) \quad (1)$$

Preposition “of”

2 BERT rep. with parse tree

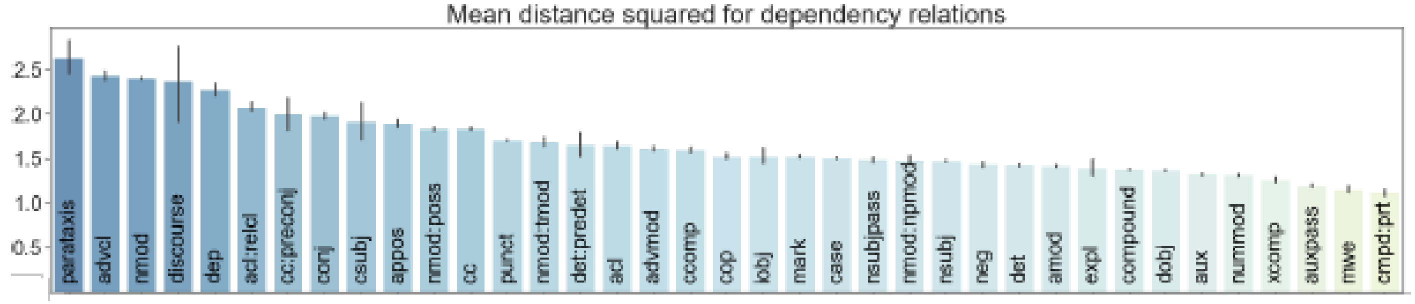


Figure 3: The average squared edge length between two words with a given dependency.

2. Relations with parse tree and hidden representations=> Proved!

Discussion

- Is the difference between these projected trees and the canonical ones is merely noise, or an additional quantitative aspect.

3 Word sense with BERT rep.

How does BERT represent useful linguistic information internally?

Three main explorations

Syntactic

1. Attention matrices contain grammatical representations

2. Relations with parse tree and hidden representations
   + Visualization

Semantic

3. BERT representation also has the information of word sense
   + Visualization
   + Measurement

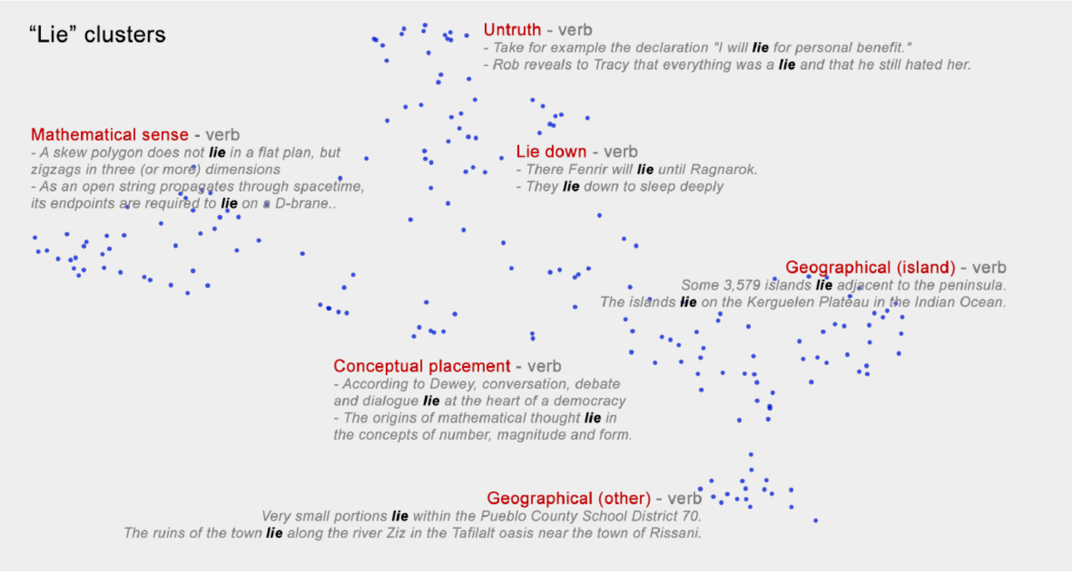
3 Word sense with BERT rep.

3. BERT representation also has the information of word sense

It is natural to speculate this, but the visualization and measurement…

Qualitative Analysis

- The system retrieves 1K sentences containing that word
- … visualize these context embeddings using UMAP



3 Word sense with BERT rep.

3. BERT representation also has the information of word sense

Quantitative Analysis : Word Sense Disambiguation(WSD)

- Goal : to classify *n* word senses of polysemy(다의어)
- Dataset : SemCor
- Baseline : a nearest-neighbor classifier where each neighbor is the centroid of a given BERT embeddings
- Suggested method : a bilinear model similar with (J. Hewitt and C. Manning, ACL 2019)

$$\min_B \sum_{\ell} \frac{1}{|s^\ell|^2} \sum_{i,j} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell)^2 \right|$$

Method	F1 score
Baseline (most frequent sense)	64.8
ELMo [20]	70.1
BERT	**71.1**
BERT (w/ probe)	**71.5**

m	Trained probe	Random probe
768 (full)	71.26	70.74
512	71.52	70.51
256	71.29	69.92
128	71.21	69.56
64	70.19	68.00
32	68.01	64.62
16	65.34	61.01

Semantic probe % accuracy on final-layer BERT-base

3 Word sense with BERT rep.

+ Concatenation Experiments

- Hypothesis : Can sentence concat. Influence context embedding
- Method
  1. define a *matching* and an *opposing* sense centroid
  2. Record cosine similarity for each pairs (*key.*, *matching*), (*key.*, *opposing*)
  3. Concatenate random sentence
  4. Repeat [2]

Similarity ratio = $\dfrac{\text{(similarity with } \textit{matching}\text{)[2]}}{\text{(similarity with } \textit{Opposing}\text{)[4]}}$

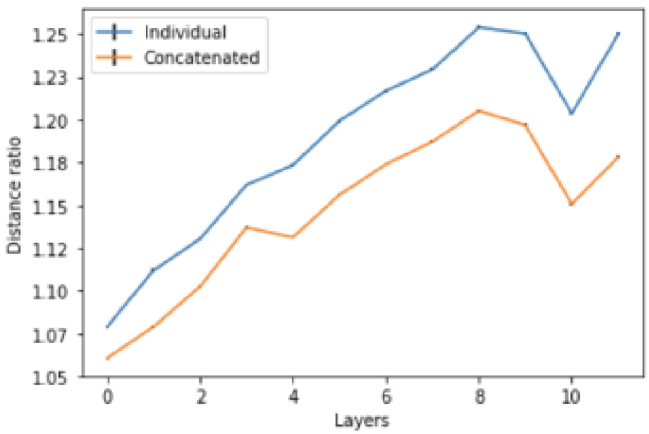


Figure 5: Average similarity ratio: senses A vs. B.

Subspace for syntactic-semantic

We compared our word sense disambiguation probe (A) to Hewitt and Manning's syntax probe (B).

Idea

a vector encodes both syntax and semantics, but in separate complementary subspaces