[1] McGill
[2] Microsoft Research
[3] Mila
[4] 香 港 大 學 THE UNIVERSITY OF HONG KONG

Linguistic Dependencies and Statistical Dependence

Jacob Louis Hoover[1,3] **Alessandro Sordoni**[2] **Wenyu Du**[4] and **Timothy J. O'Donnell**[1,3]

Overview

Two kinds of relationships between words in natural language

1. **Linguistic dependencies**: tree structure over sentence, representing compositional structure
2. **Statistical dependence**: how words affect probability of other words

➡ *Do words that are statistically dependent tend to be those in linguistic dependencies?*

We use **pretrained LMs** to estimate statistical dependence as **CPMI** (defined below), which we can compare to linguistic dependencies. We find:

1. CPMI dependency accuracy is **only as high as** a **simple baseline that connects adjacent words**
   - **across languages**
   - **even for syntactically-aware LMs**
   - even **between POS tags** instead of wordforms
2. CPMI-dependencies **differ substantially between LMs**.

Statistical dependence: CPMI

Pointwise mutual information (PMI) quantifies statistical dependence between words. In context:

$$\text{pmi}(x; y \mid c) \equiv \log \frac{p(x, y \mid c)}{p(x \mid c)p(y \mid c)} = \log \frac{p(x \mid y, c)}{p(x \mid c)}.$$

We define **Contextualized Pointwise Mutual Information** (**CPMI**), an estimate of PMI between two words in a sentence $W = w_{1:n}$ using a pretrained language model M:

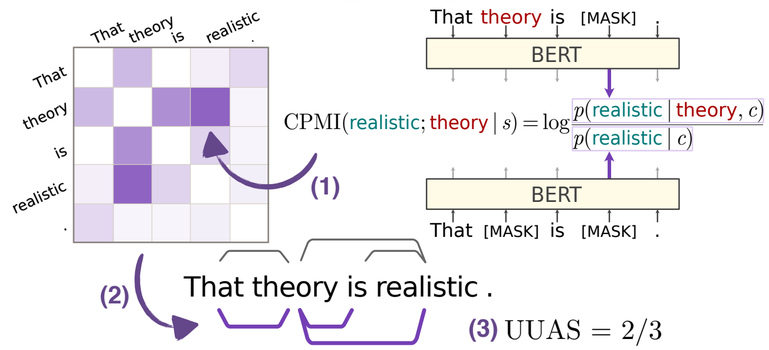
$$\text{CPMI}_M(w_i; w_j) \equiv \log \frac{p_M(w_i \mid W_{-i})}{p_M(w_i \mid W_{-i,j})}.$$

$p_M(w_i \mid W_{-i})$: probability of w_i given sentence without w_i

$p_M(w_i \mid W_{-i,j})$: as above, but also without w_j.

Extracting CPMI-dependencies

(1) Compute of matrix of CPMI values per sentence
(2) Extract max-CPMI spanning tree per sentence
(3) Compute accuracy against gold dependencies



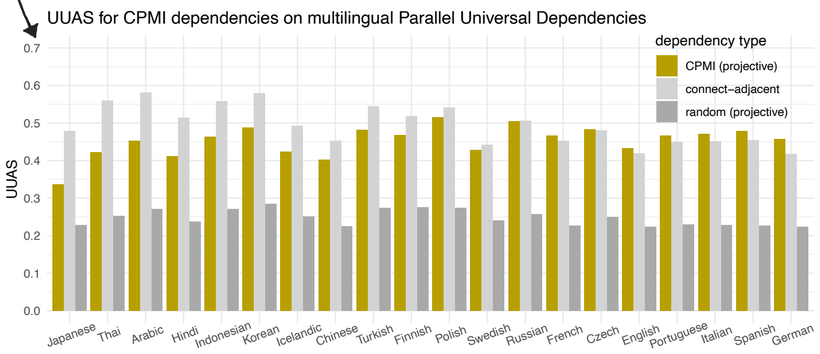
Accuracy not higher than baseline

We compute unlabeled undirected attachment score (**UUAS**; the proportion of edges in common) between LM's **CPMI-dependencies** and gold dependencies.

random	.22
connect-adjacent	**.49**
Word2Vec	.39
BERT base	.46
BERT large	.47
DistilBERT	.48
Bart large	.38
XLM	.42
XLNet base	.45
XLNet large	.41
vanilla LSTM	.44
ONLSTM	.44
ONLSTM-SYD	.45

Accuracy of CPMI trees is **only as good as the connect-adjacent baseline**, at best.

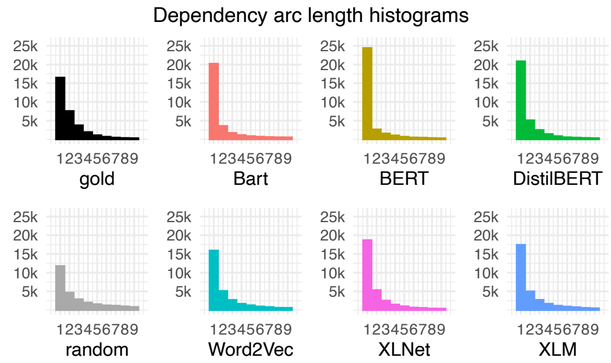
- across **multiple languages**
- also for **syntactically-aware LMs**.
- also for dependence between as **POS tags**.



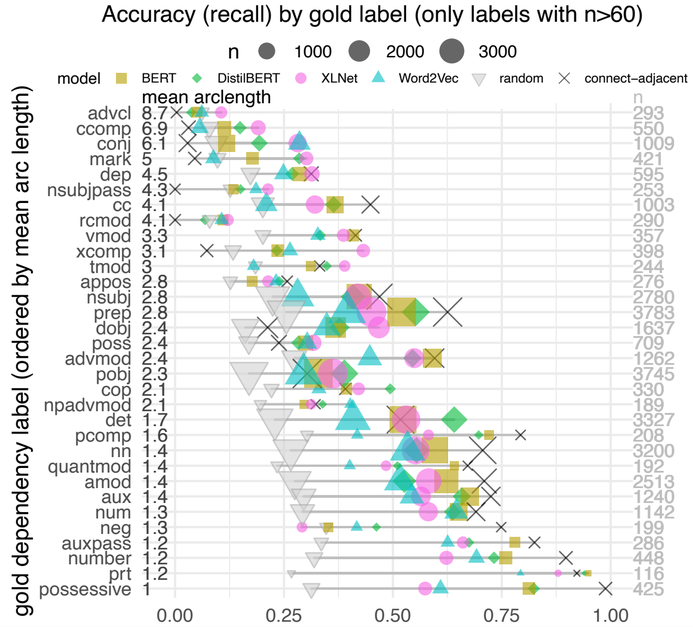
Analysis of CPMI-dependencies

Looking closer at CPMI-dependencies from pretrained LMs, we find that

- accuracy is **not correlated with LM performance**
- CPMI-dependencies **over-predict connecting adjacent words** more often than gold (esp. BERT's):



- no **types of dependency** have very high accuracy beyond connecting adjacent words:



- **CPMI-dependencies differ substantially between LMs**

edited: 2021-10-11 paper: https://arxiv.org/abs/2104.08685