Finding Syntax with Structural Probes

Jun Li

Representation learned by NN

The chef who ran to the store was out of food.

Representation learned by NN

The chef who ran to the store was out of food.

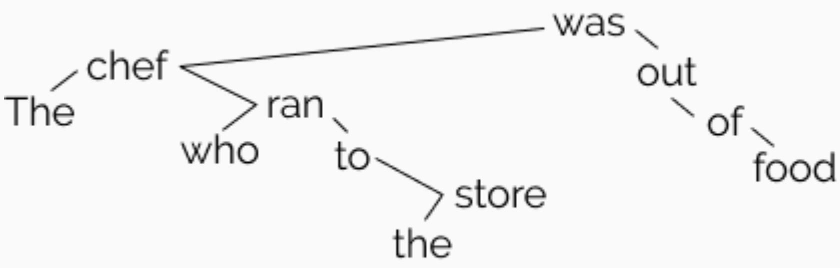
The	chef	who	ran	to	the	store	was	out	of	food
.4	.1	.3	.7	.4	.1	.3	.1	.3	-.8	0
-.2	.9	-.4	-.4	0	-.6	.1	.9	.1	.3	.7
.3	-.2	.2	0	-.5	.2	-.6	-.8	.8	-.6	-.9

Representation learned by NN

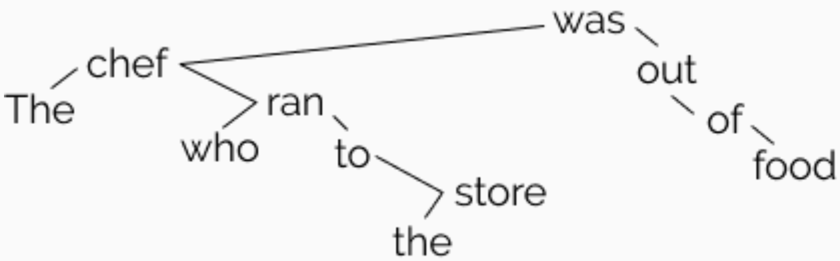
The chef who ran to the store was out of food.

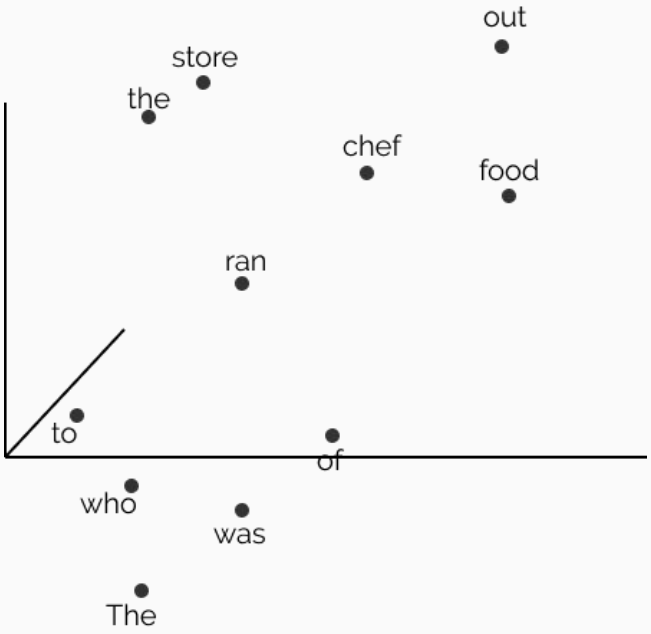
Learned by `human’

The chef who ran to the store was out of food.



Are they ***reconcilable***?





Previous works

- We can use these vectors to predict depth of the parse tree
- But can we do better?
- Can we recover a parse tree?

The	chef	who	ran	to	the	store	was	out	of	food
$\begin{bmatrix} .4 \\ -.2 \\ .3 \end{bmatrix}$	$\begin{bmatrix} .1 \\ .9 \\ -.2 \end{bmatrix}$	$\begin{bmatrix} .3 \\ -.4 \\ .2 \end{bmatrix}$	$\begin{bmatrix} .7 \\ -.4 \\ 0 \end{bmatrix}$	$\begin{bmatrix} .4 \\ 0 \\ -.5 \end{bmatrix}$	$\begin{bmatrix} .1 \\ -.6 \\ .2 \end{bmatrix}$	$\begin{bmatrix} .3 \\ .1 \\ -.6 \end{bmatrix}$	$\begin{bmatrix} .1 \\ .9 \\ -.8 \end{bmatrix}$	$\begin{bmatrix} .3 \\ .1 \\ .8 \end{bmatrix}$	$\begin{bmatrix} -.8 \\ .3 \\ -.6 \end{bmatrix}$	$\begin{bmatrix} 0 \\ .7 \\ -.9 \end{bmatrix}$

Question

Q: What do `vector space` and `parse tree space` have in common?

A: Geometry between words

Vector vs. Parse Tree

- Words are represented as vectors

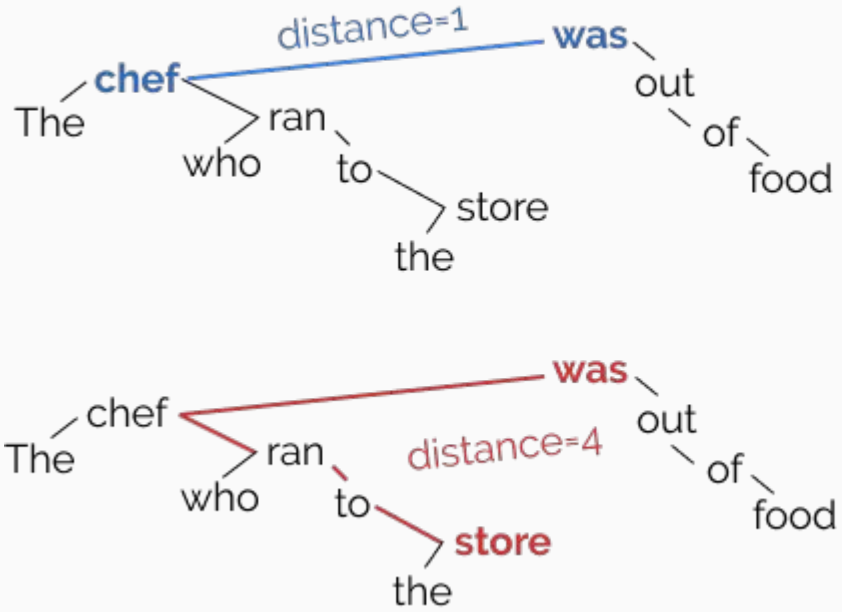
- Words are represented as words

Vector vs. Parse Tree

- Words are represented as vectors
- Distance between words can be obtained by calculating distance between vectors

- Words are represented as words
- Distance between words is the length of the tree path

Distance on tree



Vector vs. Parse Tree

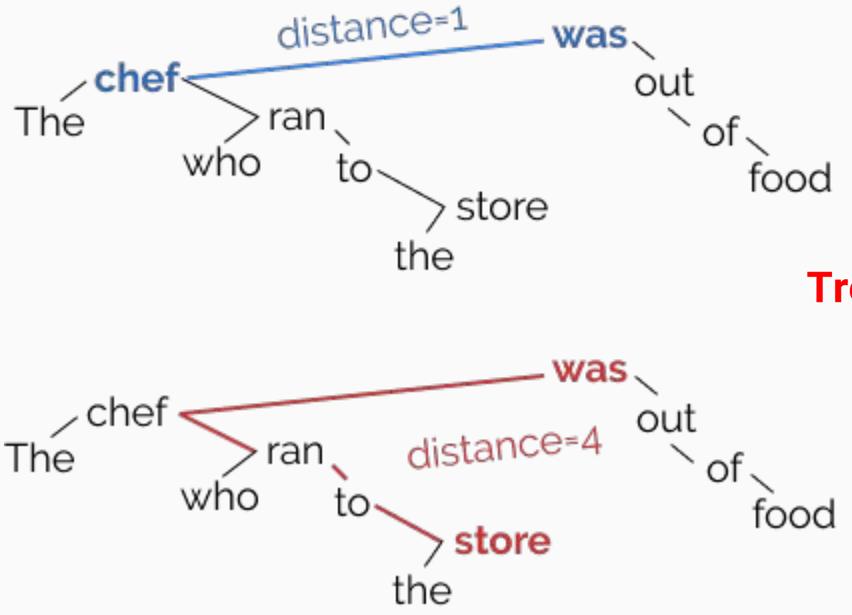
- Words are represented as vectors
- Distance between words can be obtained by calculating distance between vectors
- Norm of a vector represents how far it is from the origin

- Words are represented as words
- Distance between words is the length of the tree path
- Depth of a node represents how far it is from the root node

Why do we mention this?

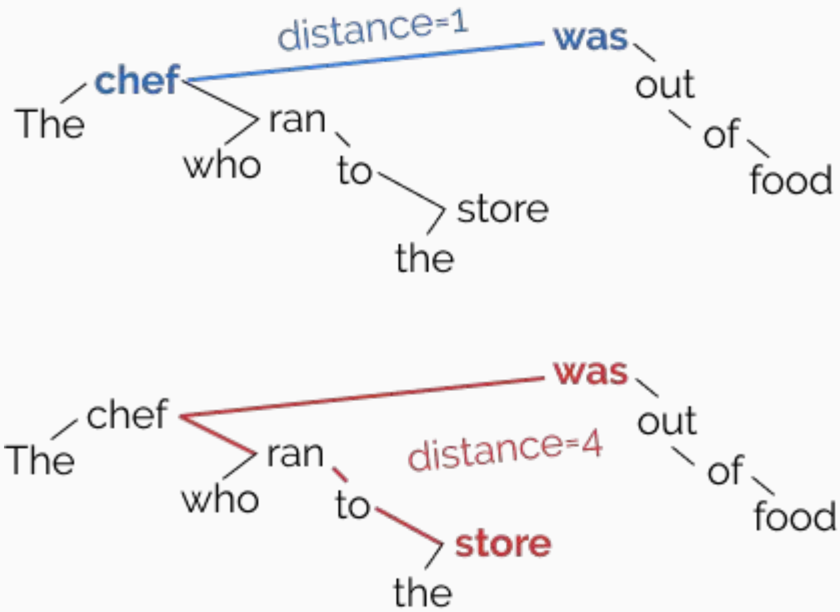
king - man = queen - woman

Hypothesis



Tree distance should be encoded in vectors!

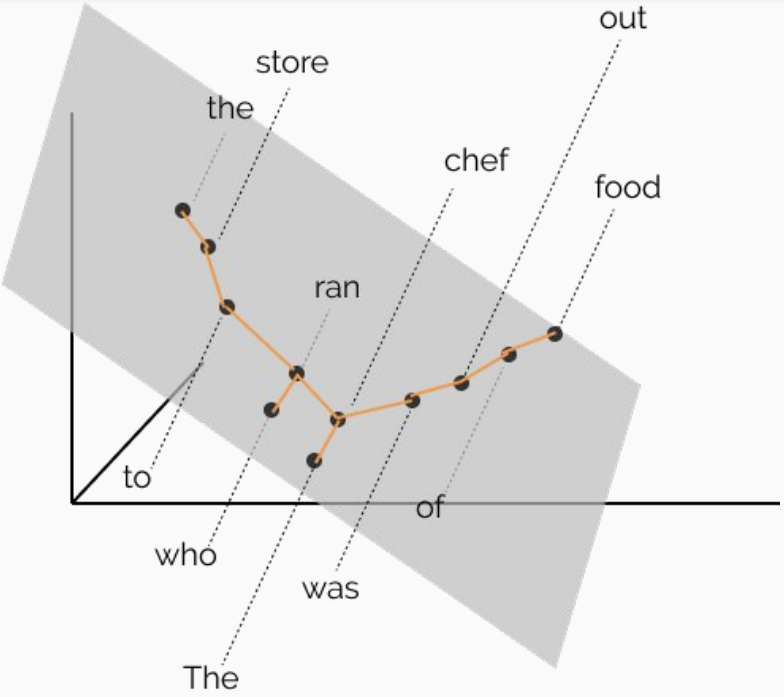
Unfortunately...



Unfortunately...

Probably the vectors encode too many information, not just syntax

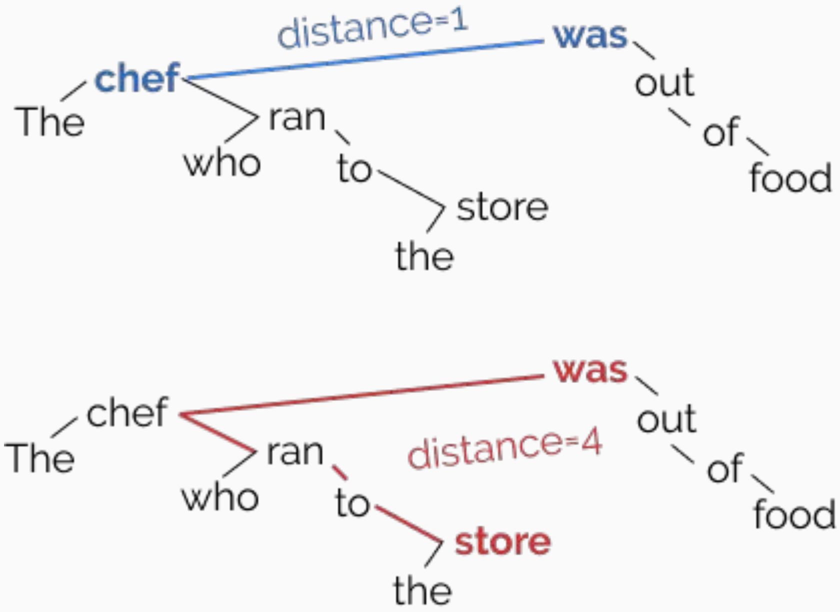
Maybe...from a different perspective?



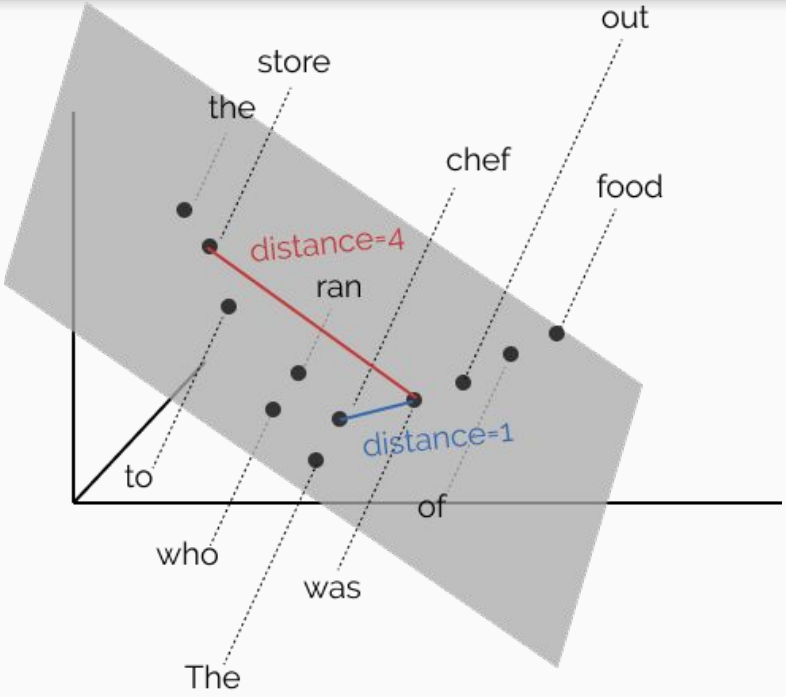
Relaxed version

Tree distance should be encoded in vectors(after a linear transformation**)!**

We hope



distance=1
was
chef
The
out
ran
of
who
to
food
store
the
was
chef
The
out
ran
distance=4
of
who
to
food
store
the



out
store
the
chef
food
distance=4
ran
distance=1
to
of
who
was
The

Formal definition

We can define a new distance:

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

$$d(h_i, h_j) = (h_i - h_j)^T (h_i - h_j)$$

And the objective is:

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

Finding syntax with structural probes

What is a probe by the way?

A supervised model for finding information in a representation

Observation Evidence: Whether a given desired behaviour is observed(S-V)

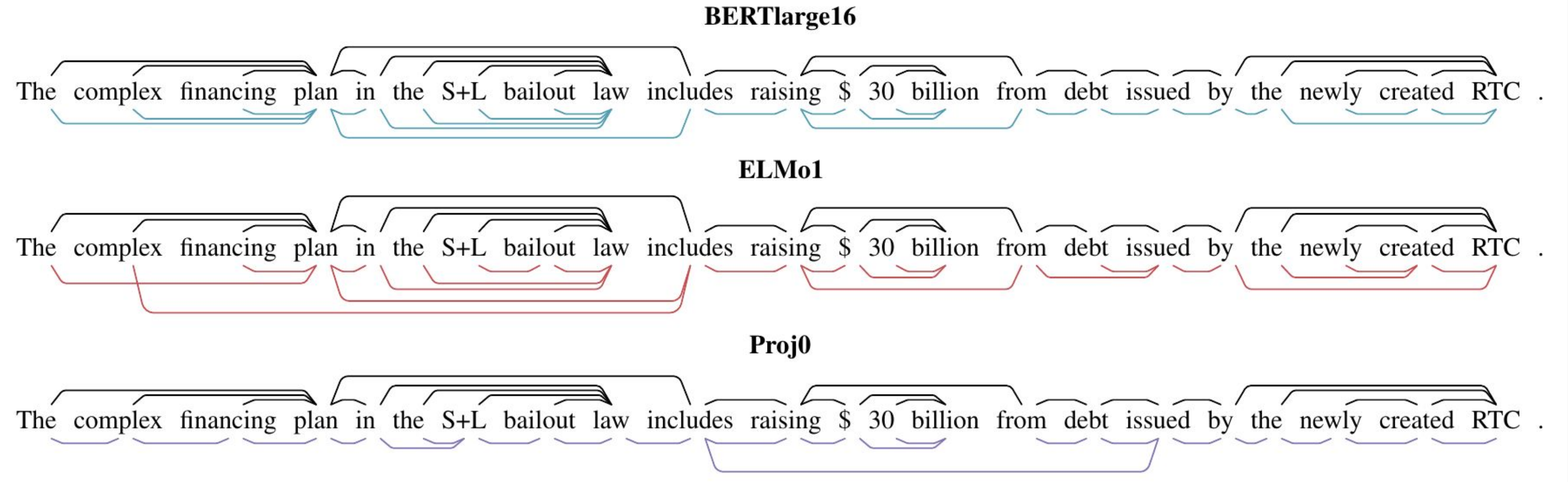
Constructive Evidence: The model may encoder the phenomenon of interest, and we train a probe supervisely to recover it

Experiments

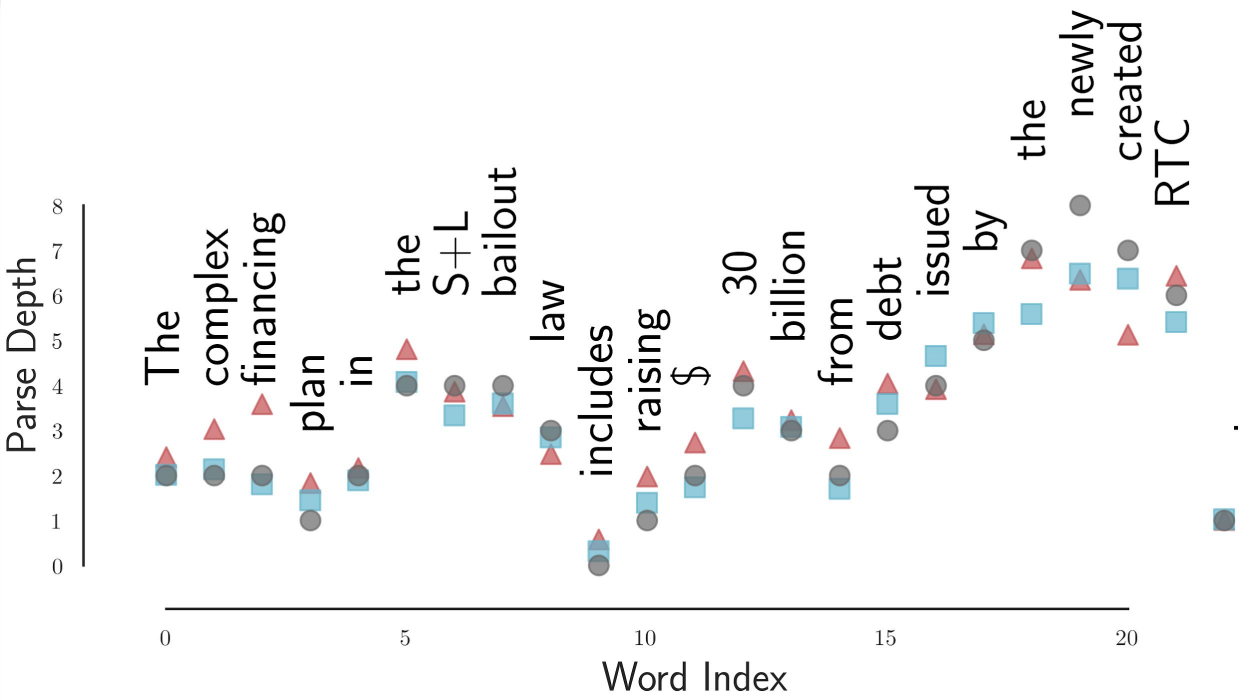
Method	Distance UUAS	Distance DSpr.	Depth Root%	Depth NSpr.
LINEAR	48.9	0.58	2.9	0.27
ELMO0	26.8	0.44	54.3	0.56
DECAY0	51.7	0.61	54.3	0.56
PROJ0	59.8	0.73	64.4	0.75
ELMO1	77.0	0.83	86.5	0.87
BERTBASE7	79.8	0.85	88.0	0.87
BERTLARGE15	**82.5**	0.86	89.4	0.88
BERTLARGE16	81.7	**0.87**	**90.1**	**0.89**

Table 1: Results of structural probes on the PTB WSJ test set; baselines in the top half, models hypothesized to encode syntax in the bottom half. For the distance probes, we show the Undirected Unlabeled Attachment Score (UUAS) as well as the average Spearman correlation of true to predicted distances, DSpr. For the norm probes, we show the root prediction accuracy and the average Spearman correlation of true to predicted norms, NSpr.

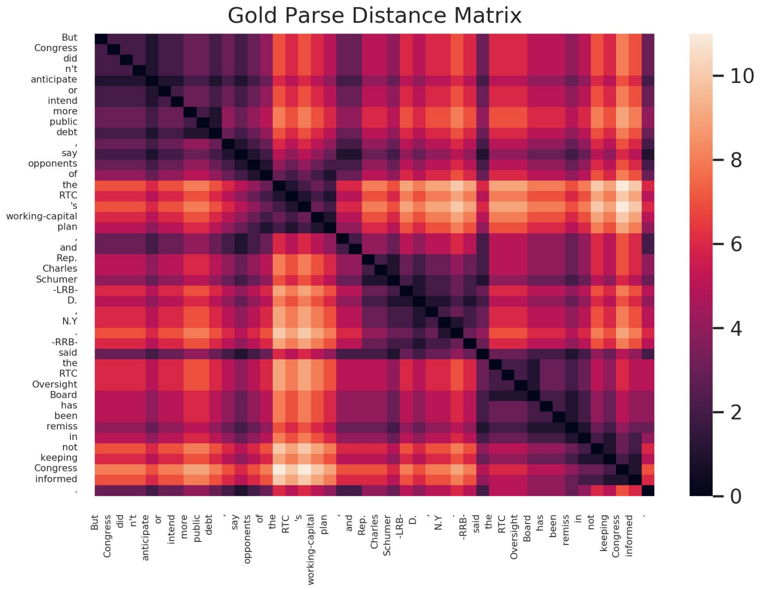
Reconstructed parse trees

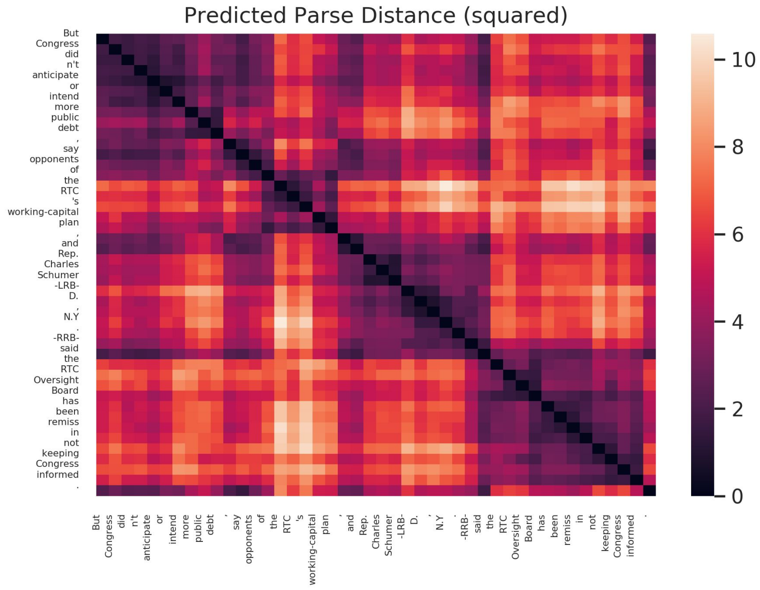


Parse diff



Parse Depth
Word Index
0
1
2
3
4
5
6
7
8
0
5
10
15
20
The
complex
financing
plan
in
the
S+L
bailout
law
includes
raising
$
30
billion
from
debt
issued
by
the
newly
created
RTC
.


Gold Parse Distance Matrix
But
Congress
did
n't
anticipate
or
intend
more
public
debt
,
say
opponents
of
the
RTC
's
working-capital
plan
,
and
Rep.
Charles
Schumer
-LRB-
D.
,
N.Y
.
-RRB-
said
the
RTC
Oversight
Board
has
been
remiss
in
not
keeping
Congress
informed
.
0
2
4
6
8
10


Predicted Parse Distance (squared)
But
Congress
did
n't
anticipate
or
intend
more
public
debt
,
say
opponents
of
the
RTC
's
working-capital
plan
,
and
Rep.
Charles
Schumer
-LRB-
D.
,
N.Y
.
-RRB-
said
the
RTC
Oversight
Board
has
been
remiss
in
not
keeping
Congress
informed
.
0
2
4
6
8
10

Different hidden layers

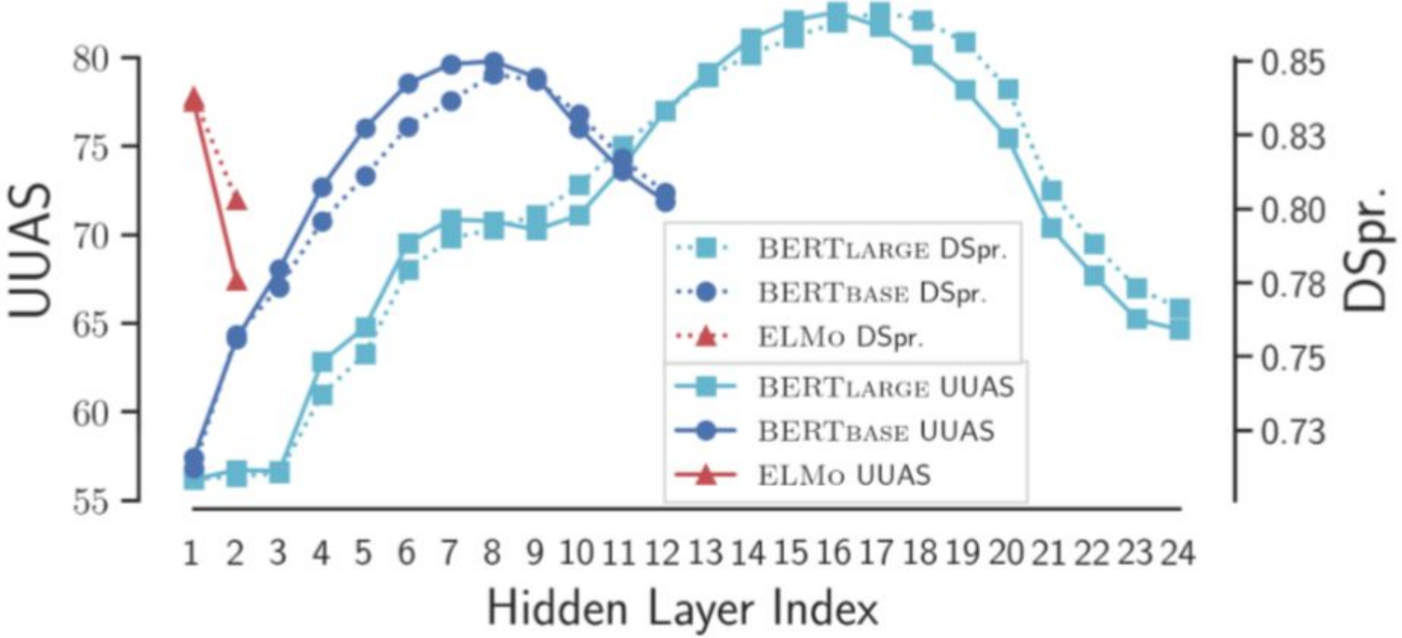
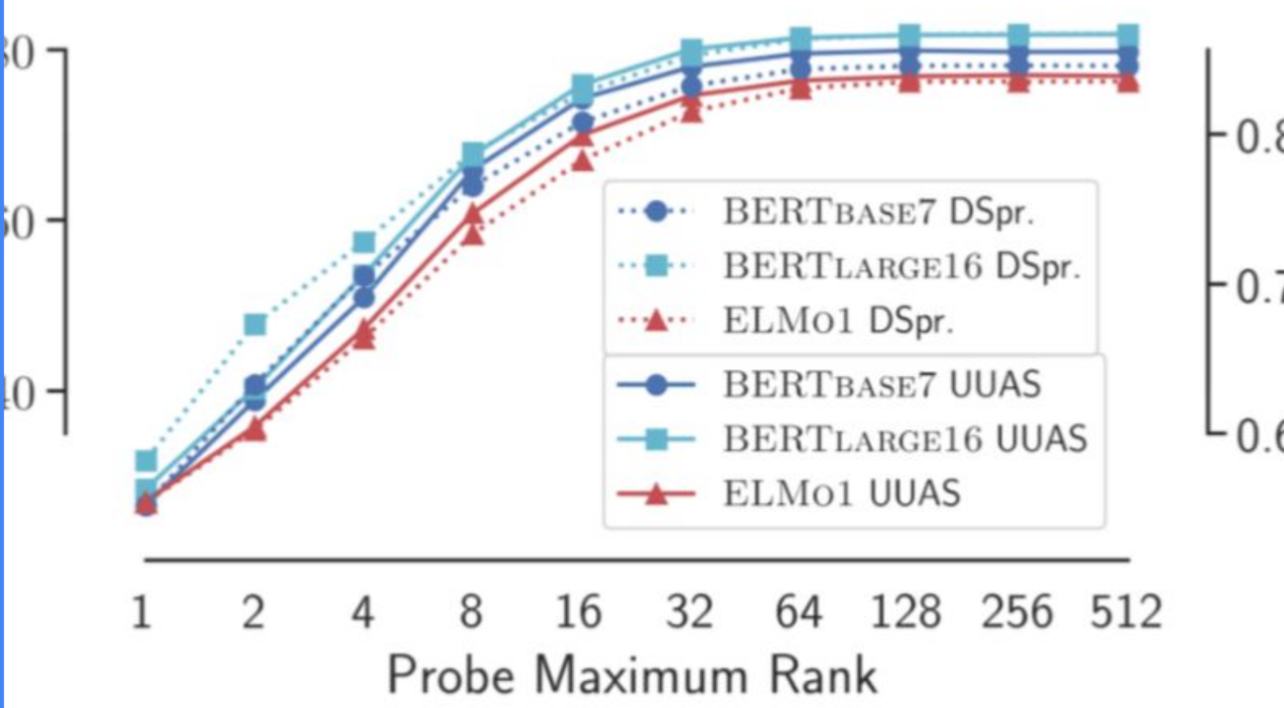


Figure 1: Parse distance UUAS and distance Spearman correlation across the BERT and ELMo model layers.

Rank of the linear transformation

Intuitively, larger k means a more expressive probing model, and a larger fraction of the representational capacity of the model being devoted to syntax.



ure 5: Parse distance tree reconstruction accuracy
linear transformation is constrained to varying maxi
ensionality.

Thanks