Probe-less probing of Bert's layer

Rithvik Sama – M15458313

Akash Mehta - M15863466

Anthology ID: 2022.naacl-srw.25

Volume: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop

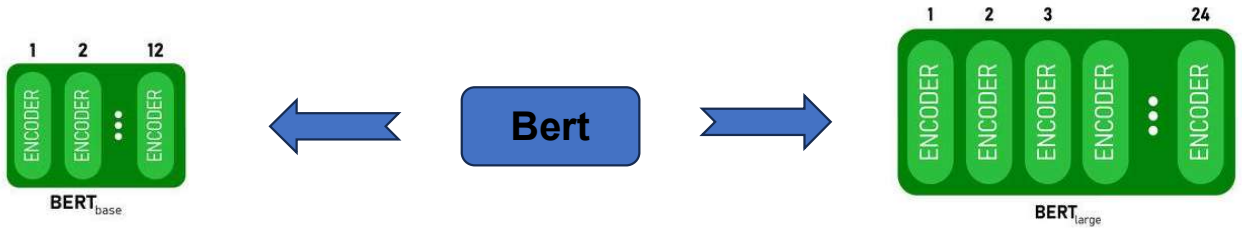
Project Description:

Probe-less probing of BERT's Layer-Wise Linguistic Knowledge with Masked Word Prediction

What is Bert(Bidirectional Encoder Representations from Transformers):

- Model introduced by GOOGLE in 2018.
- Transformer-based architecture that uses attention and self-attention to weight the significance of different words in a sentence.
- Designed to train from both left and right contexts at every layer, by utilizing unlabeled text to condition.
- Pre-trained and finely-tuned for specific tasks like question answering, sentimental analysis, and language inferences.

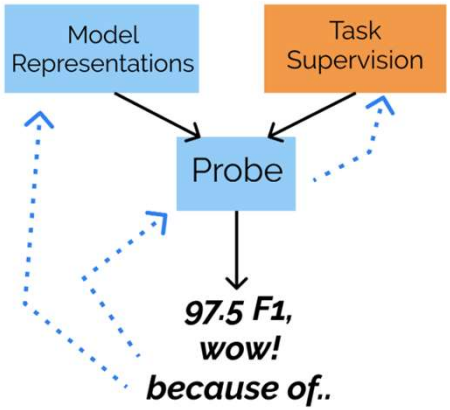
Types of BERT



Probing-

In simple terms: A technique used to analyze and understand how well different layers of neural network models understand or represent various aspects of language.

Involves a pre-training the model which can be called as the probe classifier. These are specially designed to predict specific linguistic properties.



Probe-less Probing-

Analyzing linguistic knowledge captured by the models without relying on an external probing classifier.

Probing

Probe classifier ✗

Dataset Used:

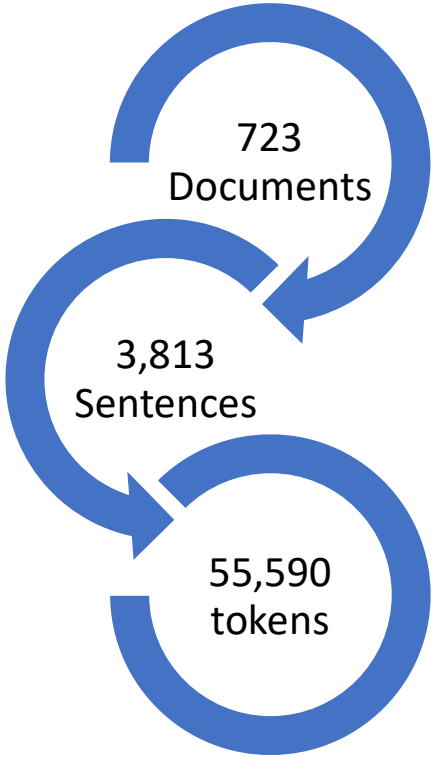
STREUSLE 4.4 –

A corpus of web reviews written in English.

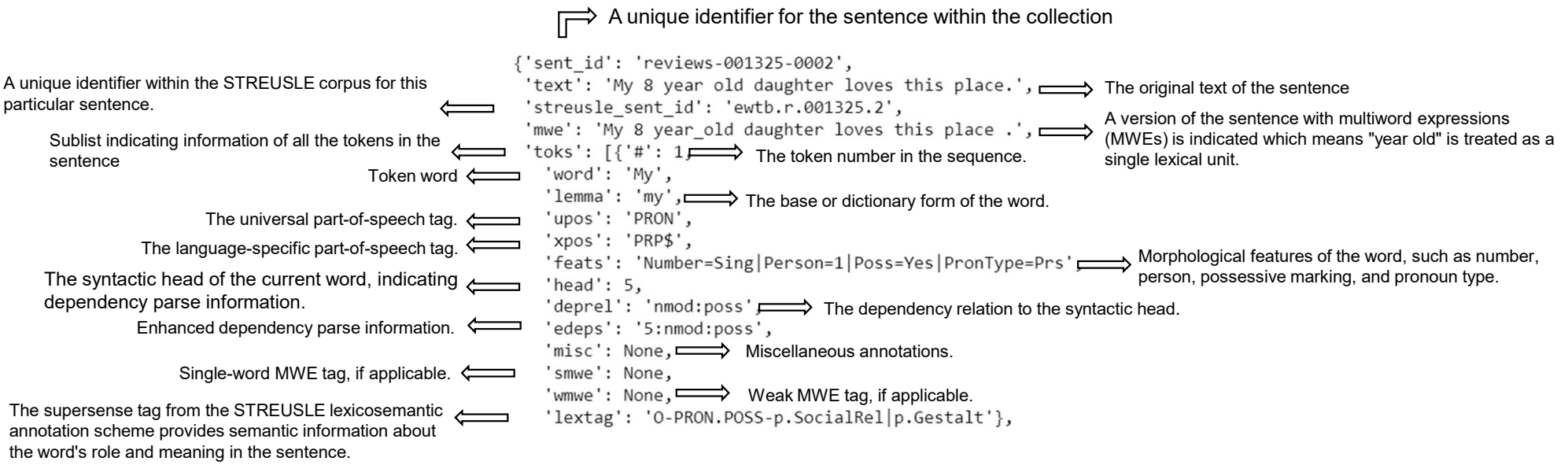
With annotations of 3,013 strong multiword expressions.

↓

Phrases with multiple words that collectively have a meaning that is not deducible from the meanings of individual words.



STREUSLE Breakdown

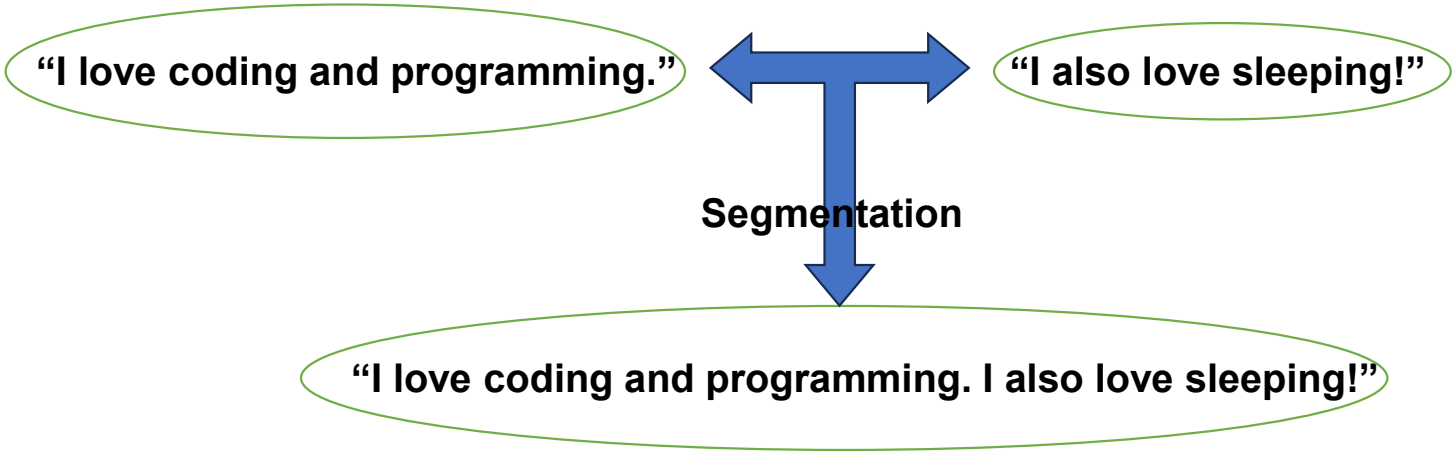


Method

Step 1.

Segmentation/ Tokenization

Dividing the large group of sentences collectively called passages into individual sentences.

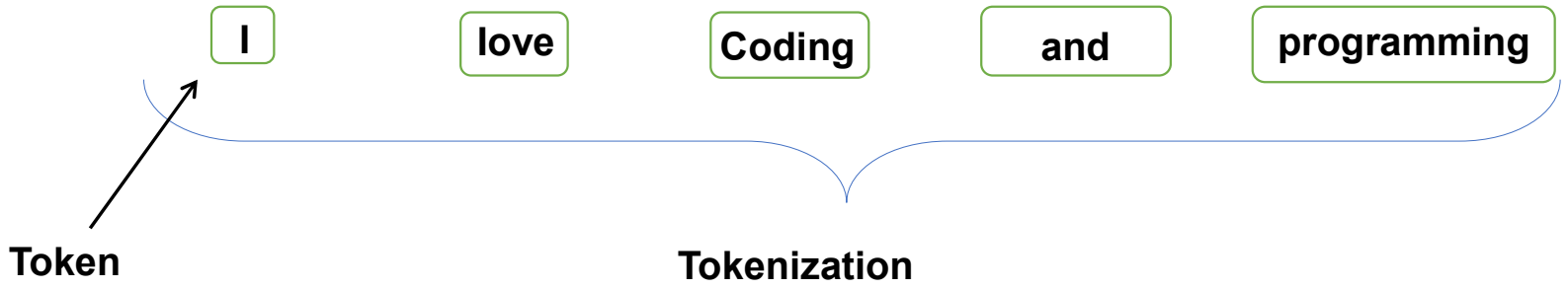


Method

Step 1.

Tokenization

Individual Sentences are divided into further small units which are termed as tokens.



“I love coding and programming.”

Method

Step 2.

For each sentence, create n variants, where n is the number of tokens in the sentence by replacing one token with [MASK] token.

'It is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.'

[MASK] is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It [MASK] next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It is [MASK] to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It is next [MASK] Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It is next to [MASK] du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare [MASK] Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du [MASK] and a five minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord [MASK] a five minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and [MASK] five minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and a [MASK] minute walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and a five [MASK] walk to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and a five minute [MASK] to Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and a five minute walk [MASK] Sacre Coeur which is excellent for shopping.
It is next to Gare du Nord and a five minute walk to [MASK] Coeur which is excellent for shopping.
It is next to Gare du Nord and a five minute walk to Sacre [MASK] which is excellent for shopping.
It is next to Gare du Nord and a five minute walk to Sacre Coeur [MASK] is excellent for shopping.
It is next to Gare du Nord and a five minute walk to Sacre Coeur which [MASK] excellent for shopping.
It is next to Gare du Nord and a five minute walk to Sacre Coeur which is [MASK] for shopping.
It is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent [MASK] shopping.
It is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for [MASK]

STEP 3.

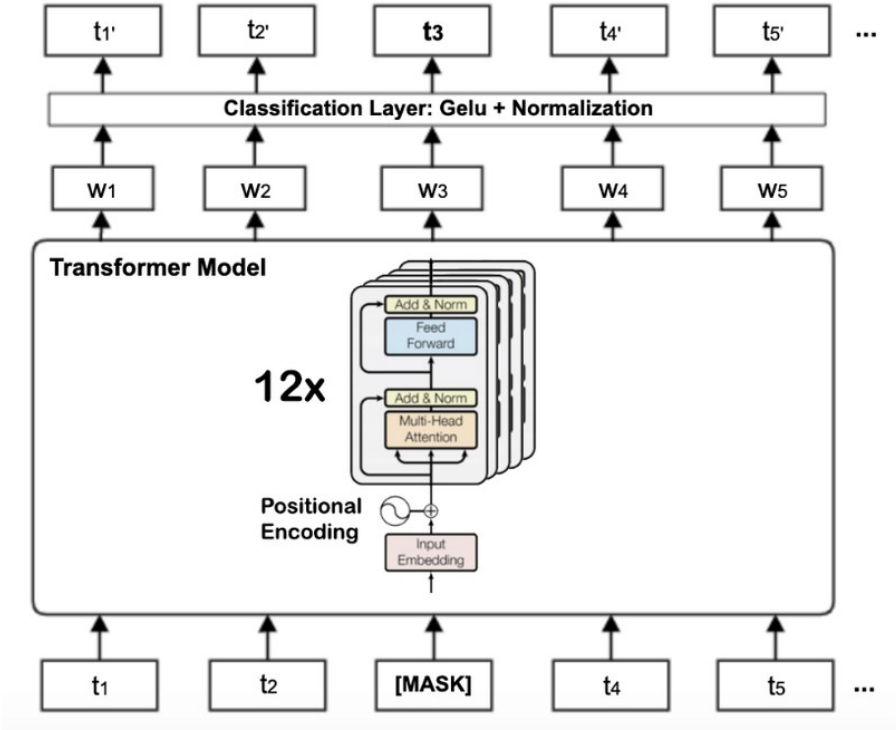
Each variant is fed through the neural network.

Which neural network?

STEP 3.

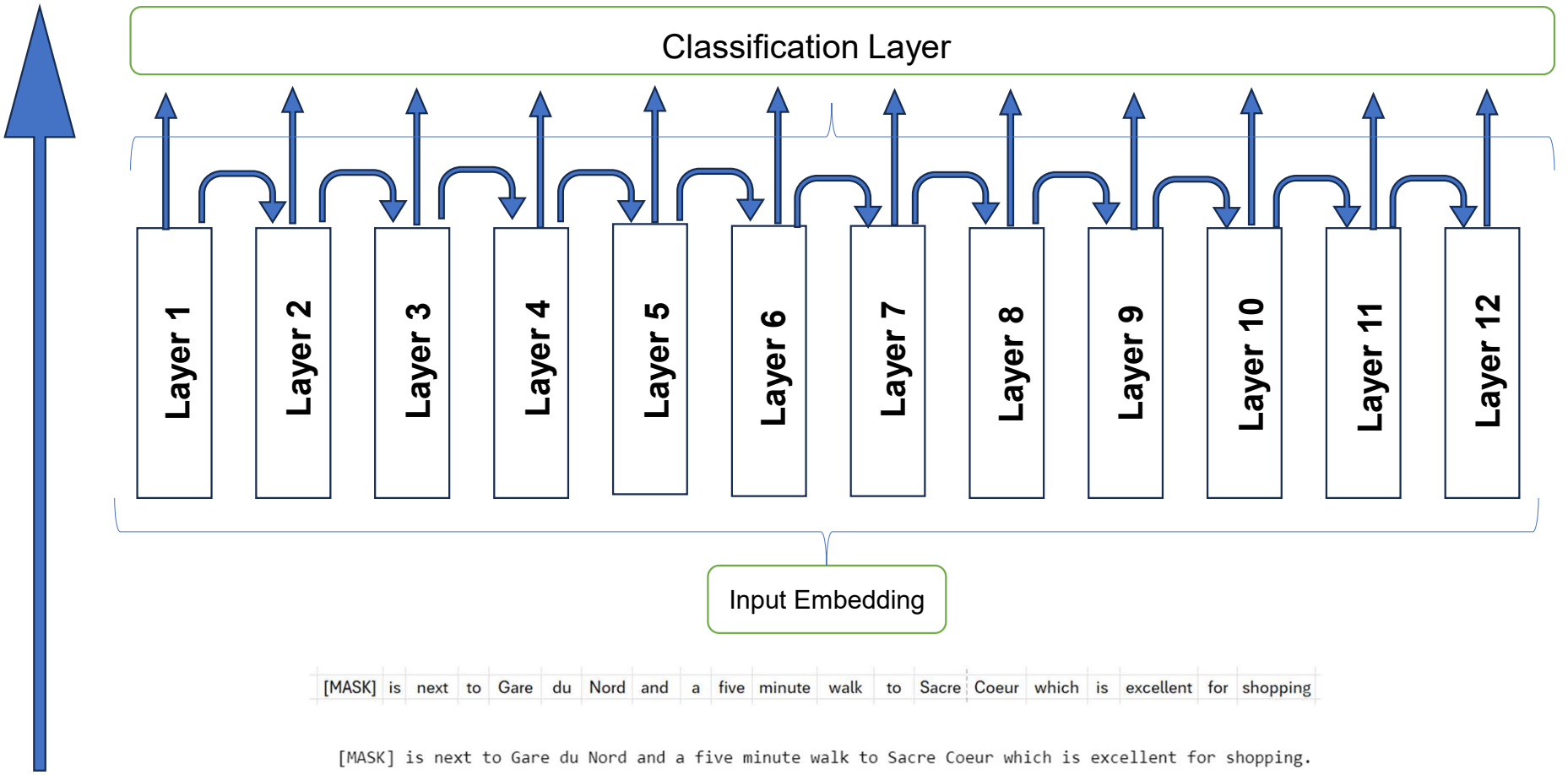
Bert-base

- Bert base comprises 12 layers (transformer blocks), 768 hidden units, and 12 self-attention heads.
- BERT is pre-trained on a large corpus of text using two unsupervised tasks:
    1. Masked Language Modeling (In MLM, some percentage of the input tokens are masked, and the model needs to predict the masked word based on its context.
    2. Next Sentence Prediction, the model learns to predict whether two sentences are consecutive or not.
- After pre-training, the model is trained on a smaller dataset specific to the task, allowing it to adjust its parameters to perform well on that particular task like question answering, sentiment analysis, and named entity recognition.



```
predicted_token: 'it'
actual_token: 'It'
predicted_pos: 'PUNCT'
actual_pos: 'PRON'
```

```
It is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
```



[MASK]	is	next	to	Gare	du	Nord	and	a	five	minute	walk	to	Sacre	Coeur	which	is	excellent	for	shopping

```
[MASK] is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.
```

Results

Layer wise prediction:

{'id': {'layer0': {'predicted_token': '.',
'actual_token': 'It',
'predicted_pos': 'PUNCT',
'actual_pos': 'PRON',
'predicted_sentence': '. is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.'},
'layer1': {'predicted_token': '.',
'actual_token': 'It',
'predicted_pos': 'PUNCT',
'actual_pos': 'PRON',
'predicted_sentence': '. is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.'},
'layer2': {'predicted_token': 'there',
'actual_token': 'It',
'predicted_pos': 'ADV',
'actual_pos': 'PRON',
'predicted_sentence': 'there is next to Gare du Nord and a five minute walk to Sacre Coeur which is excellent for shopping.'},

Results

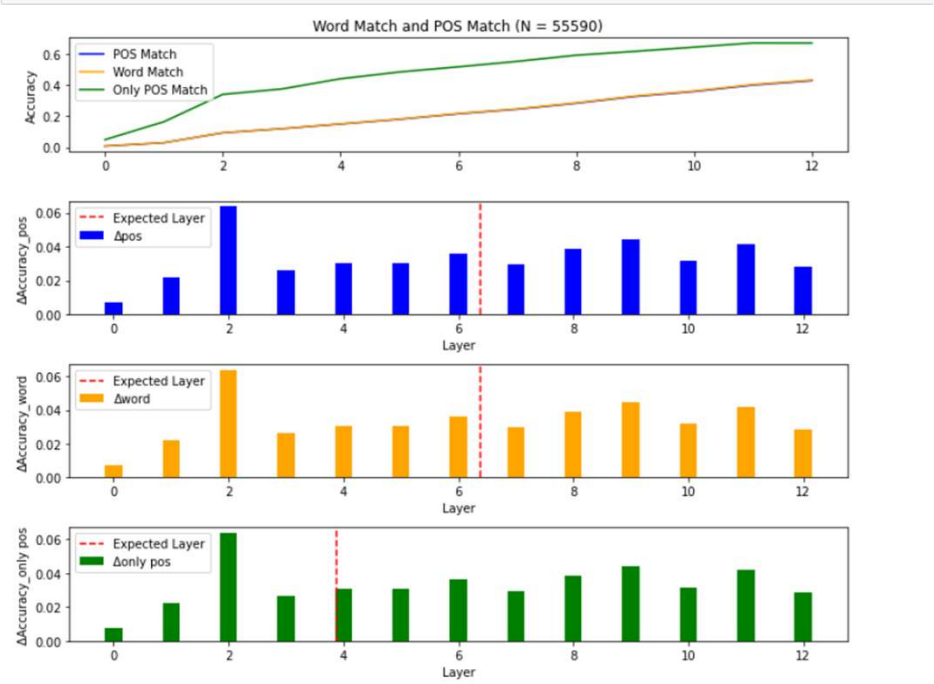
Accuracy is calculated as the total number of correct predictions over the total number of predictions.

Incremental gain at a layer l is defined as:

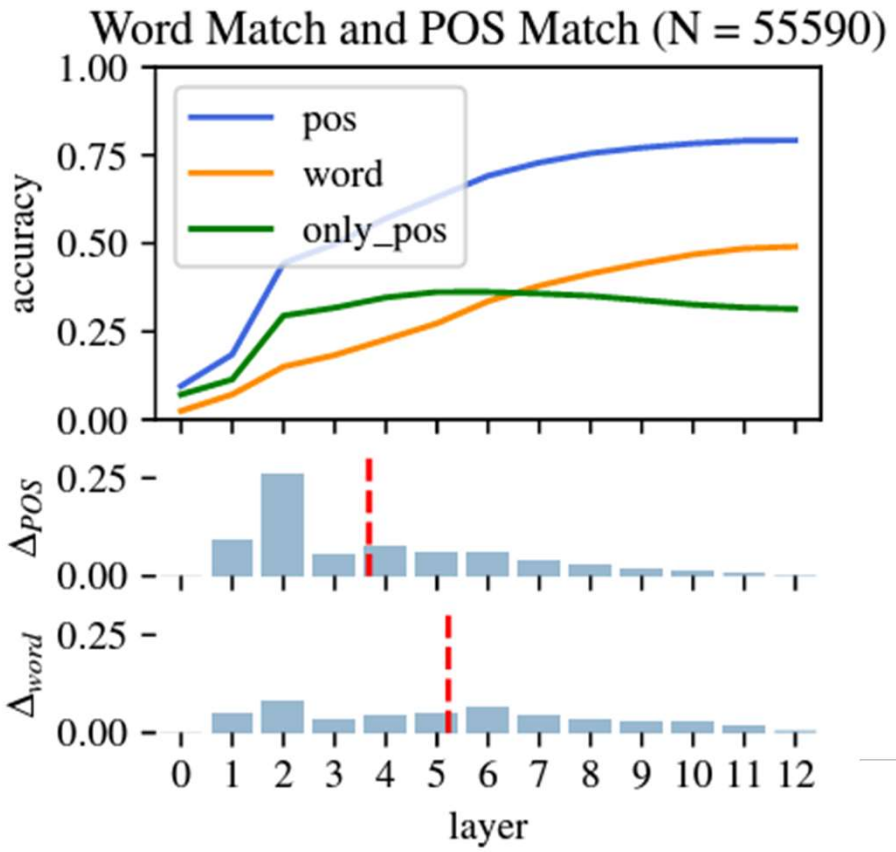
$$\Delta_T^{(\ell)} = Score_T^{(\ell)} - Score_T^{(\ell-1)}$$

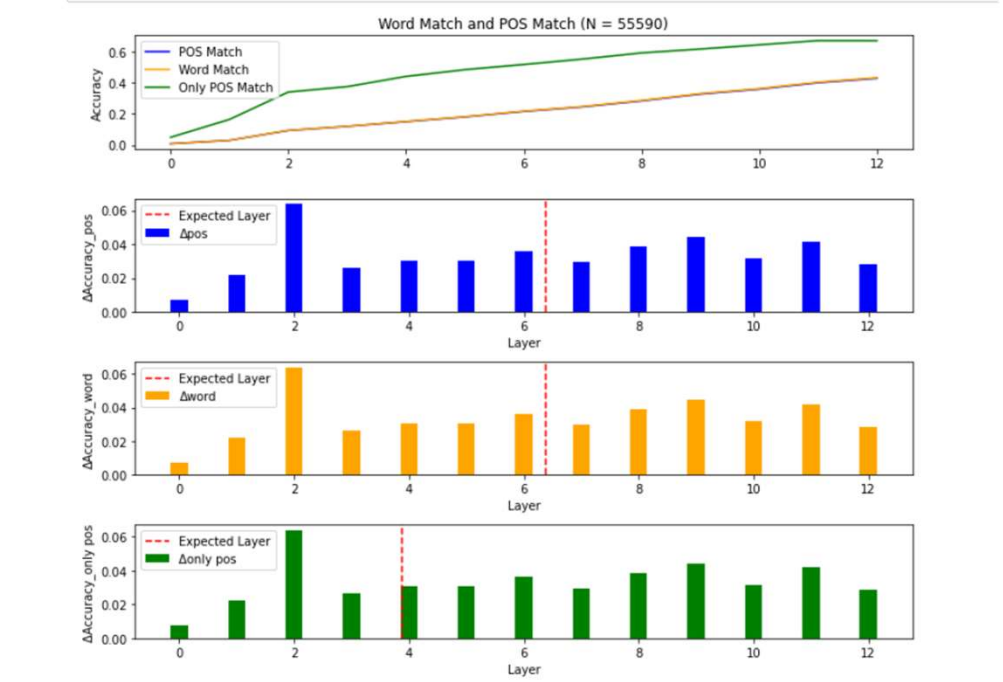
The expected layer at which gain scores are centered:

$$\bar{E}_\Delta[\ell] = \frac{\sum_{\ell=1}^{L} l \cdot \Delta_T^{(\ell)}}{\sum_{\ell=1}^{L} \Delta_T^{(\ell)}}$$



Results


Word Match and POS Match (N = 55590)
accuracy
pos
word
only_pos
Δ_{POS}
Δ_{word}
layer


Word Match and POS Match (N = 55590)
Accuracy
POS Match
Word Match
Only POS Match
ΔAccuracy_pos
Expected Layer
Δpos
ΔAccuracy_word
Δword
ΔAccuracy_only pos
Δonly pos
Layer

Results

UPOS	N	POS	POSm	WORD	WORDm
ADJ	3169	7.80	7.89	8.70	8.94
ADV	3080	5.98	5.85	7.29	8.31
INTJ	108	7.60	0.00	8.18	10.25
NOUN	7265	7.25	7.21	9.11	9.01
PROPN	1406	6.73	6.85	8.63	8.59s
VERB	5328	6.27	6.36	8.42	8.65
ADP	3368	6.07	6.01	7.38	7.84
AUX	2950	6.79	6.76	7.62	7.60
CCONJ	1803	5.56	5.49	5.51	5.55
DET	3525	9.25	0.00	6.59	6.60
NUM	555	7.21	8.67	7.66	9.10
PART	1314	6.09	5.77	6.15	5.70
PRON	5264	6.49	6.35	7.82	8.41
SCONJ	808	7.05	7.09	7.79	5.83
X	69	7.18	5.67	8.69	0.00
PUNCT	5958	5.56	3.79	5.90	7.57
SYM	159	7.78	7.75	7.86	8.12

Expected layer by UPOS, POSm, and wordm.

		N	POS	POS_M	word	$word_M$
open	ADJ	3169	4.04	4.24	6.45	6.35
	ADV	3080	3.42	3.76	5.74	5.30
	INTJ	108	3.48	**9.33**	7.13	**8.75**
	NOUN	7265	3.98	4.48	7.53	6.99
	PROPN	1406	**6.68**	6.11	**8.05**	7.88
	VERB	5328	3.96	3.68	6.73	6.38
closed	ADP	3368	3.16	3.52	5.01	5.18
	AUX	2950	3.10	4.43	5.14	5.08
	CCONJ	1803	5.48	5.32	5.88	4.74
	DET	3525	2.16	2.54	3.11	3.43
	NUM	555	5.70	6.81	6.73	7.23
	PART	1314	1.80	1.31	2.08	1.40
	PRON	5264	3.91	4.61	6.76	5.96
	SCONJ	808	5.05	4.45	5.71	5.45

Expected layer by UPOS, POSm, and wordm according to the paper.

- POS accuracy achieved was generally higher across most UPOS categories as compared to the paper. For example, in ADJ, your POS accuracy is 7.80, while the paper reports 4.04. This could indicate that your model is better at predicting the correct part of speech, or it might be a sign of overfitting or a difference in the evaluation method.
- POS accuracy for multiword expressions. In some cases, the POSm values we achieved are significantly higher, such as in NUM (your value is 8.67 compared to 2.16 in the paper) or DET (your value is 9.25 compared to 2.54 in the paper). These results suggest a very different performance on multiword expressions, which is quite striking.
- Word Match Accuracy: Generally higher.
- The discrepancies in wordM, the word accuracy for multiword expressions, are again very notable. Your NOUN wordM accuracy is 9.01 compared to 6.99 in the paper.

Results

UPOS	N	POS	POSm	WORD	WORDm
ADJ	3169	7.80	7.89	8.70	8.94
ADV	3080	5.98	5.85	7.29	8.31
INTJ	108	7.60	0.00	8.18	10.25
NOUN	7265	7.25	7.21	9.11	9.01
PROPN	1406	6.73	6.85	8.63	8.59s
VERB	5328	6.27	6.36	8.42	8.65
ADP	3368	6.07	6.01	7.38	7.84
AUX	2950	6.79	6.76	7.62	7.60
CCONJ	1803	5.56	5.49	5.51	5.55
DET	3525	9.25	0.00	6.59	6.60
NUM	555	7.21	8.67	7.66	9.10
PART	1314	6.09	5.77	6.15	5.70
PRON	5264	6.49	6.35	7.82	8.41
SCONJ	808	7.05	7.09	7.79	5.83
X	69	7.18	5.67	8.69	0.00
PUNCT	5958	5.56	3.79	5.90	7.57
SYM	159	7.78	7.75	7.86	8.12

Expected layer by UPOS, POSm, and wordm.

		N	POS	POS_M	word	$word_M$
open	ADJ	3169	4.04	4.24	6.45	6.35
	ADV	3080	3.42	3.76	5.74	5.30
	INTJ	108	3.48	**9.33**	7.13	**8.75**
	NOUN	7265	3.98	4.48	7.53	6.99
	PROPN	1406	**6.68**	6.11	**8.05**	7.88
	VERB	5328	3.96	3.68	6.73	6.38
closed	ADP	3368	3.16	3.52	5.01	5.18
	AUX	2950	3.10	4.43	5.14	5.08
	CCONJ	1803	5.48	5.32	5.88	4.74
	DET	3525	2.16	2.54	3.11	3.43
	NUM	555	5.70	6.81	6.73	7.23
	PART	1314	1.80	1.31	2.08	1.40
	PRON	5264	3.91	4.61	6.76	5.96
	SCONJ	808	5.05	4.45	5.71	5.45

Expected layer by UPOS, POSm, and wordm according to the paper.

Furthermore, we have 3 extra UPOS categories, specifically X, PUNCT and SYM. The tag X often represents a token that does not fit into any other category, and if such tokens were never part of MWEs in the enteire 69 instances, then that could explain the expected layer being 0,00 for wordm. The same goes for INT where posm has 0 instances out of the whole population of 108, Perhaps there were no instances of interjections in MWEs, or they were not captured due to some reason.

Challenges Faced

- Reimplementation of this paper requires a significant computational source. A total of 55590 masked tokens ran across the multiple layers for numerous tokens which has been computationally very expensive.
- STREUSLE 4.4, a corpus of web reviews with rich syntactic and lexical-semantic annotations is used Accessing this particular dataset and understanding its structure format for the reimplementation has been very challenging.
- The paper vaguely explores the complex interaction between syntactic categories, multiword expressions, and the performance of different BERT layers which has become challenging for understanding them particularly what they have used for representing the linguistic information.

Future Aspects:

Extend the probing to other linguistic phenomena such as anaphora resolution, negation handling, or idiomatic expressions to see how well these aspects are captured across different layers. This could involve developing new datasets or leveraging existing ones tailored to these phenomena.

Questions?

Thank You