In-depth exploration of the syntactic capabilities of autoencoding language models

PhD Student: Laura Pérez-Mayos
TALN Research Group, Pompeu Fabra University, Barcelona, Spain

Supervisors: Dr. Leo Wanner, Dr. Miguel Ballesteros

upf. Universitat Pompeu Fabra Barcelona
Department of Information and Communication Technologies

About me

I ❤️ Python

- @lpmayos
- Computer Scientist.
- Former web app developer.
- President of PyBCN and co-organiser of PyLadiesBCN.

www.pybcn.org

Knowledge-based Natural Language Interaction Lab

Dr. Leo Wanner

- Natural Language Text Generation
- Deep language analysis
- Computational Lexicography and Lexicology
- Text Meta-Analysis
- Communicative Structure and Thematic Progression

Large Scale Text Understanding Systems Lab

Dr. Horacio Saggion

- Automatic text simplification
- Sentiment analysis and opinion mining
- Sign Language processing

TALN Research group
www.taln.upf.edu

How much pretraining data do language models need to learn syntax?

EMNLP 2021

While pretraining methods are very convenient, they are expensive in terms of time and resources. This calls for a study of the impact of pretraining data size on the syntactic knowledge of the models.

On the Evolution of Syntactic Information Encoded by BERT's Contextualized Representations

EACL 2021

Pretrained models are often fine-tuned on downstream tasks, and therefore it becomes increasingly important to understand how the encoded knowledge evolves along the fine-tuning process.

BERT & friends: the beginning of a new era in NLP

Syntactic assessment of language models

Supervised probing

Hewitt & Manning structural probe (2019)

Entire syntax trees are embedded implicitly in the vector geometry of the deep models in both ELMo and BERT

Tree distance evaluation

Evaluates how well the predicted distances between all pairs of words in a model reconstruct gold parse trees.

- UUAS (Undirected Unlabeled Attachment Score). Percentage of undirected edges placed correctly.
- Dspr. Spearman correlation between true and predicted distances for each word in each sentence.

Tree depth evaluation

Evaluates the ability to recreate the order of words specified by their depth in the parse tree.

- Root %. Ability of the models to identify the root of the sentence as the least deep word.
- Nspr. Spearman correlation between the predicted and the true depth ordering.

Targeted syntactic evaluation

Hu et al. (2020) syntactic tests

The targeted syntactic evaluation incorporates methods from psycholinguistic experiments, allowing to distinguish models with human-like representations of syntactic structure
(Linzen et al., 2016; Lau et al., 2017; Gulordava et al., 2018; Marvin and Linzen, 2018; Futrell et al., 2019) .

Hu et al. (2020) test 20 model type combinations and data sizes on 34 English syntactic test suites, finding substantial differences in syntactic generalization performance by model architecture.

2. Targeted syntactic evaluation

How well do models generalise over syntactic phenomena?

Agreement:

The author that the senators hurt [is] good.

* The author that the senators hurt [are] good.

Center embedding:

The painting that the artist [painted] deteriorated.

* The painting that the artist [deteriorated] painted.

Garden-path effects:

* As the ship crossed the waters [remained] blue and calm.

As the ship crossed, the waters [remained] blue and calm.

Gross Syntactic Expectation:

? **As** the doctor studied the book[.]

The doctor studied the book[.]

Licensing:

No teacher that the ministers hated has failed [any] student.

The teacher that the ministers hated has failed [any] student.

Long-Distance Dependencies:

* My neighbor told me **what** the dog caught [the mouse] in full view of the neighbors yesterday.

My neighbor told me **that** the dog caught [the mouse] in full view of the neighbors yesterday.

How much pretraining data do language models need to learn syntax?

Laura Pérez-Mayos[1], Miguel Ballesteros[2], Leo Wanner[3,1]
[1] TALN Research Group, Pompeu Fabra University, Barcelona, Spain
[2] Amazon AI
[3] Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
{laura.perezm|leo.wanner}@upf.edu
ballemig@amazon.com

Impact of the pretraining data size on the syntactic abilities of the models

While pretraining methods are very convenient, they are expensive in terms of time and resources. What is the impact of pretraining data size on the knowledge of the models?

(i) As the models are pretrained on more data, and their **perplexity** improves, do they encode more **syntactic information** and offer a better **syntactic generalization**?

(ii) Do the models with more pretraining perform better when applied in **downstream tasks**?

(iii) Do we always need models pretrained on internet-scale corpora?

The MiniBERTa models

The MiniBERTas are a set of 12 RoBERTa models pretrained from scratch by Warstadt et al. (2020b) on 4 datasets containing 1B, 100M, 10M and 1M tokens, available through HuggingFace Transformers.

We refer to models trained on the same amount of data as a family of models, and models inside a family as intra-family members:

e.g. the *roberta-base-100M-1* model is a member of the *roberta-base-100M* family.

For each dataset size, pretraining is run 25 times (10 times for 1B) with varying hyperparameter values, keeping the three models with the lowest perplexity.

For the smaller dataset, a smaller model size is used to prevent over-fitting.

Experiments

1. Syntactic structural probing
2. Targeted syntactic evaluation
3. Targeted downstream tasks evaluation

1. Syntactic structural probing

Do models pretrained on more data encode a higher amount of syntactic information?

Hewitt and Manning structural probe (2019) measures how well syntax trees are embedded in a linear transformation of the network representation space applying two different evaluations:

Tree distance evaluation:

- **UUAS**. Percentage of undirected edges placed correctly.
- **Dspr**. Spearman correlation between true and predicted distances.

Tree depth evaluation:

- **Root %.** Ability to identify the root of the sentence as the least deep word.
- **Nspr**. Spearman correlation between true and predicted depth ordering.

	Tree distance evaluation		Tree depth evaluation	
Model	UUAS	Dspr.	Root %	Nspr.
roberta-1B-1	70.75	78.82	83.92	85.38
roberta-1B-2	72.93	79.86	83.53	85.92
roberta-1B-3	77.23	82.66	85.13	86.87
roberta-100M-1	68.46	76.95	81.21	84.06
roberta-100M-2	70.02	78.11	81.25	84.53
roberta-100M-3	69.35	78.73	79.88	84.59
roberta-10M-1	61.48	73.19	70.88	81.65
roberta-10M-2	62.01	73.78	70.07	81.89
roberta-10M-3	60.12	72.58	67.14	80.62
roberta-1M-1	56.96	71.70	57.12	74.16
roberta-1M-2	55.78	71.33	56.56	74.74
roberta-1M-3	55.84	71.33	57.41	74.46

2. Targeted syntactic evaluation

How well do models generalise over syntactic phenomena?

We assess the syntactic generalization performance of the different MiniBERTas models using Hu et al. (2020)'s test suites to answer the following questions:

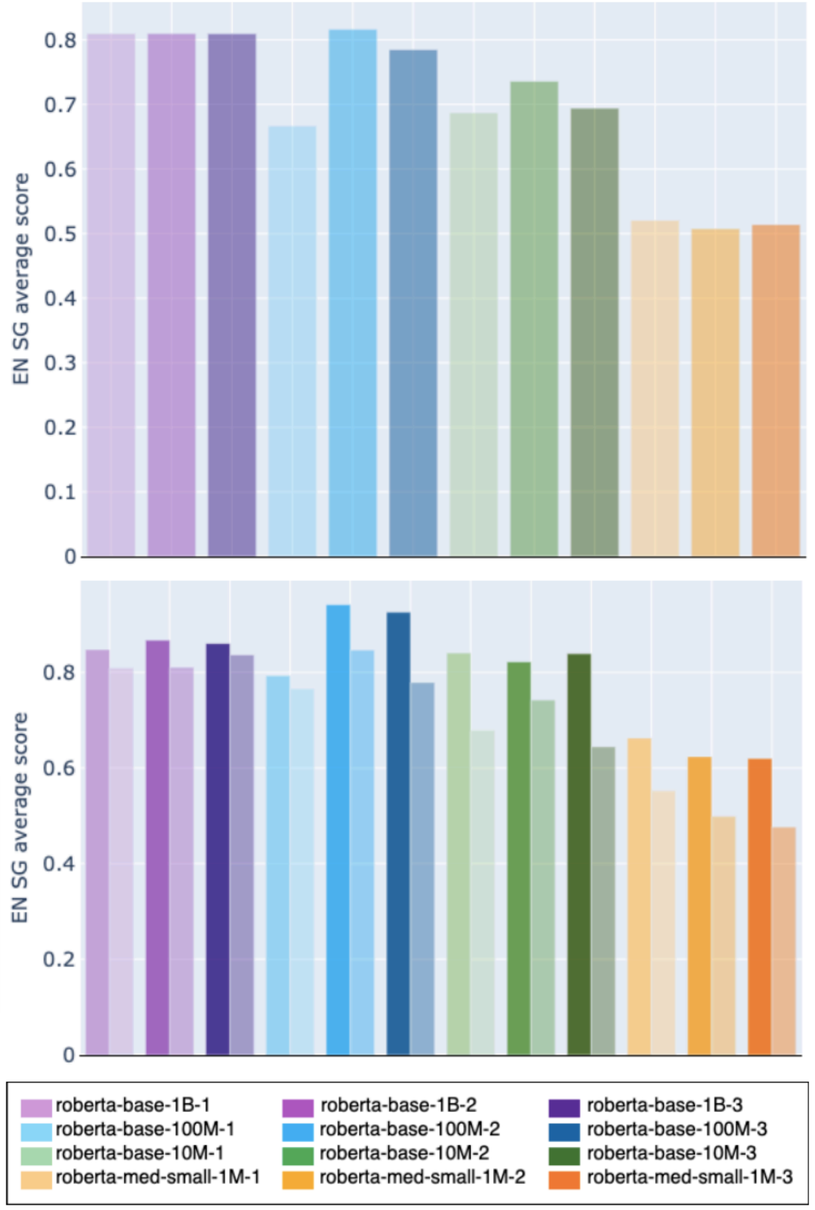
- Do models pretrained on more data generalise better over syntactic phenomena?
- Are models pretrained on more data more robust to the presence of modifiers?
- Do models with lower perplexity perform better?

2. Targeted syntactic evaluation

Do models pretrained on more data generalize better?

- High variability between family members, especially for roberta-base-100M, with a difference of 15 points between models 1 and 2.
- The smallest family offers the lowest performance.
- roberta-base-100M-1 performs worse than the whole roberta-base-10M family, and roberta-base-100M-2 performs better than the whole roberta-base-1B family.
- All models are affected by the presence of modifiers, but narrower difference for the biggest model.

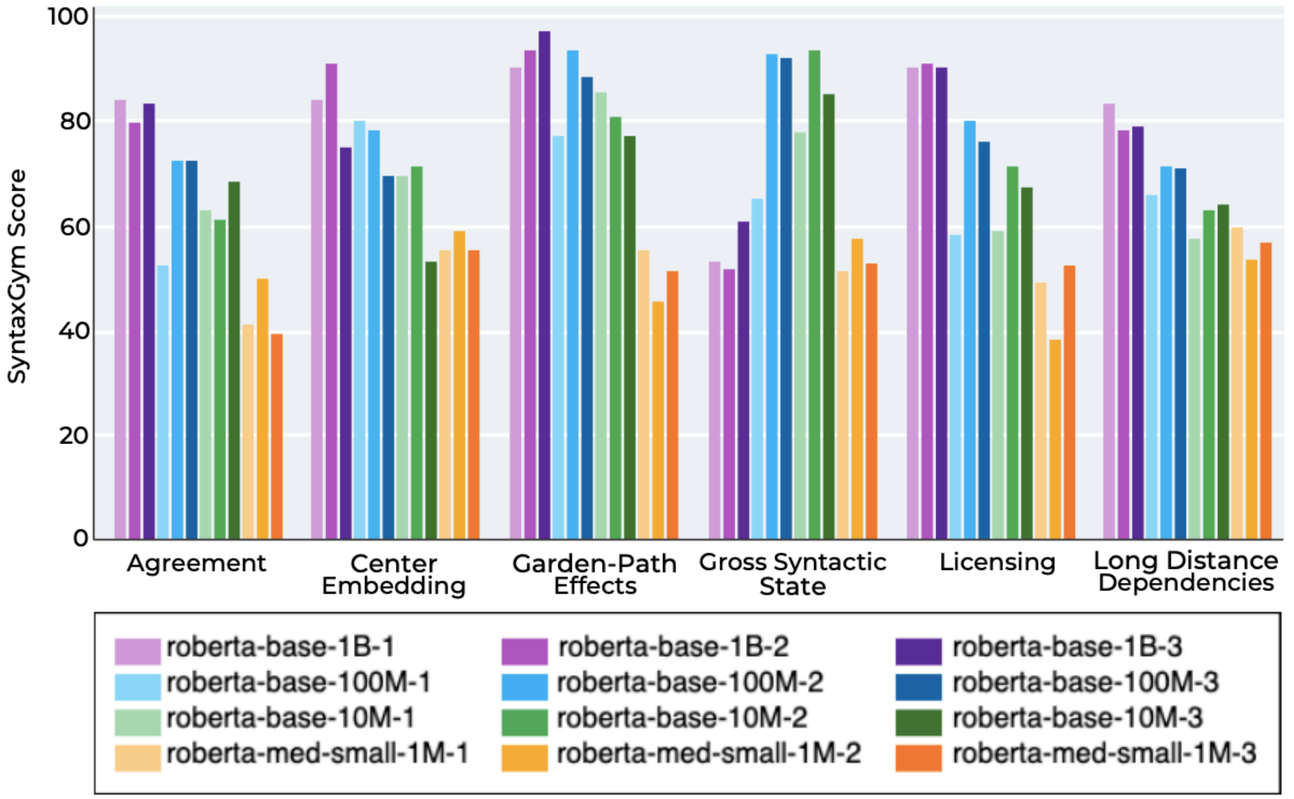
Tests with modifiers (**dark bars**) vs tests without modifiers (light bars)



2. Targeted syntactic evaluation

Do models pretrained on more data generalize better?

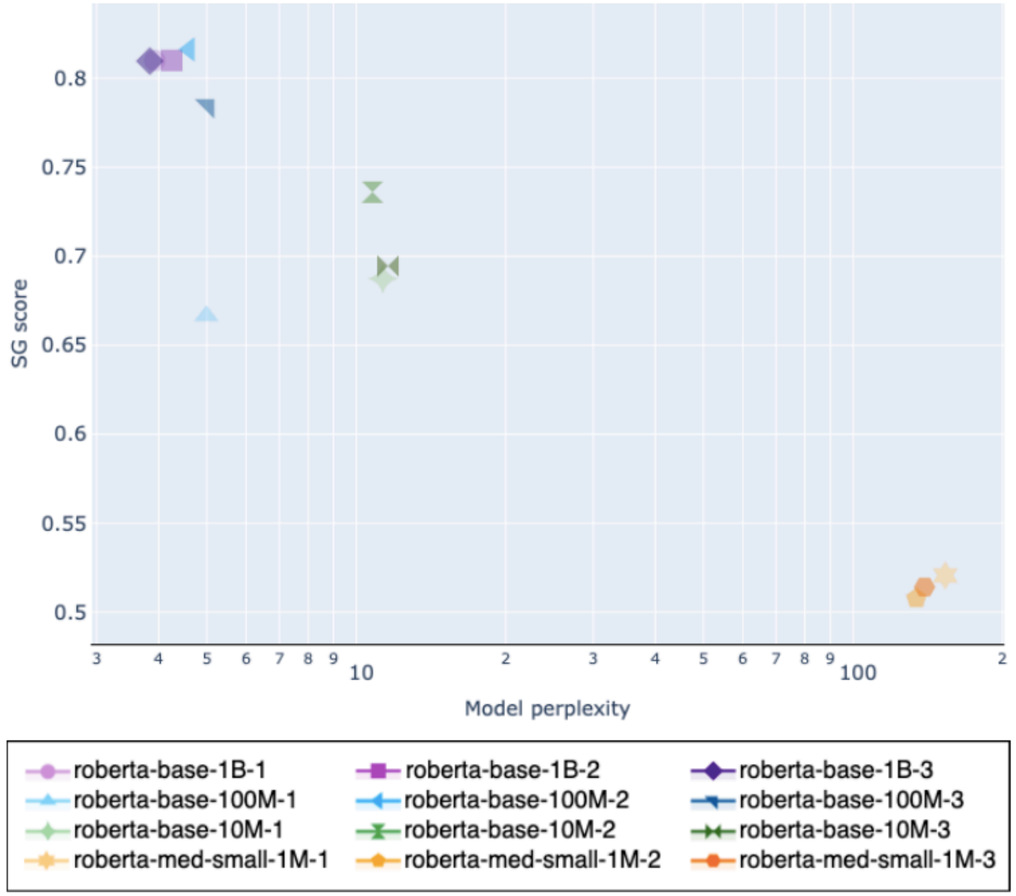
- The biggest family outperforms all other families in 3 out of 6 circuits, but offers a surprisingly low performance in Gross Syntactic State.
- The smallest family offers the lowest performance across all circuits.
- There is a high variability between the scores achieved by the models of the same family in the same circuit.
- There is not a single model for any family that performs best (nor worst) across all tests.



2. Targeted syntactic evaluation

Do models with lower perplexity perform better?

- (not perfect) negative correlation between perplexity and SyntaxGym score when comparing different families.
- No clear correlation when comparing points corresponding to the same family of models.
- Both metrics capture different aspects of the knowledge of the models.



3. Targeted downstream tasks evaluation

Do models pretrained on more data perform better on downstream tasks?

- Fine-tune the models in three different tasks, 1) PoS tagging, 2) dependency parsing and 3) paraphrase identification.
- Cost–benefit analysis of the performance gains compared with an estimate of the financial and environmental cost of developing the models.
  - We rely on the data provided in (Strubell et al., 2019) to **estimate** the cost of developing each individual model based on the costs of RoBERTa, trained on 30B words, in proportion to the amount of words used to train each family of models.

3. Targeted downstream tasks evaluation

Do models pretrained on more data perform better on downstream tasks?

- The bigger models offer a better performance, and the smallest model performs remarkably worse.
- While the increase of training data between families is exponential (1M, 10M, 100M, 1B), the performance grows at a slower rate.
- Small performance gains come at high financial and environmental costs.

Model	Cost	CO_2e (lbs)	POS	Dep. parsing	Paraphrase id.
roberta-1B	$20k (x4)	2330 (x4)	96.03 (+0.5)	85.73 (+1.76)	89.59 (+2.02)
roberta-100M	$5k (x10)	582.5 (x10)	95.53 (+1.11)	83.97 (+4.04)	87.57 (+2.79)
roberta-10M	$500 (x10)	58.25 (x10)	94.42 (+2.73)	79.93 (+4.48)	84.78 (+5.34)
roberta-1M	$50	5.825	91.69	65.45	79.44

Conclusions

Conclusions

Models pretrained with more data …

… encode more syntactic information as measured by the Hewitt and Manning probe.

Structural probing

… do not always generalise better over syntactic phenomena, but are more robust to the presence of modifiers.

Targeted Syntactic eval.

… perform better on downstream taks, at a higher financial and environmental cost (not proportional to the performance gains).

Downstream tasks eval.

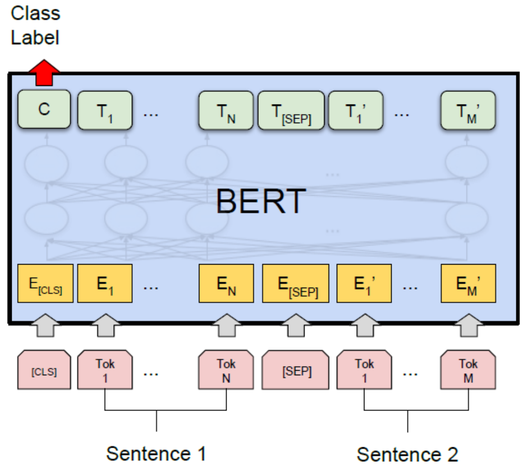
We should weight between the benefit of making models bigger and the costs this implies, prioritizing computationally efficient hardware and algorithms.

On the Evolution of Syntactic Information Encoded by BERT's Contextualized Representations

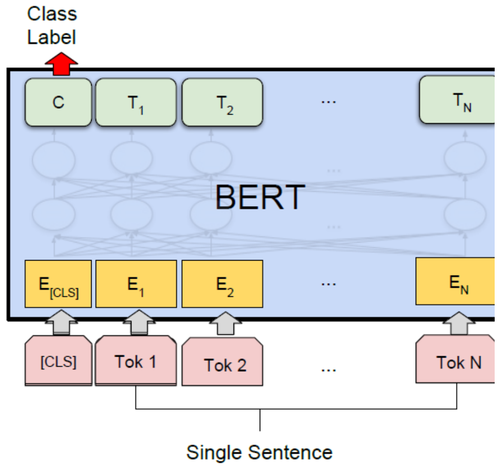
Laura Pérez-Mayos[1], Roberto Carlini[1], Miguel Ballesteros[2], Leo Wanner[3,1]
[1] TALN Research Group, Pompeu Fabra University, Barcelona, Spain
[2] Amazon AI
[3] Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain
{laura.perezm|roberto.carlini|leo.wanner}@upf.edu
ballemig@amazon.com

Fine-tuning BERT

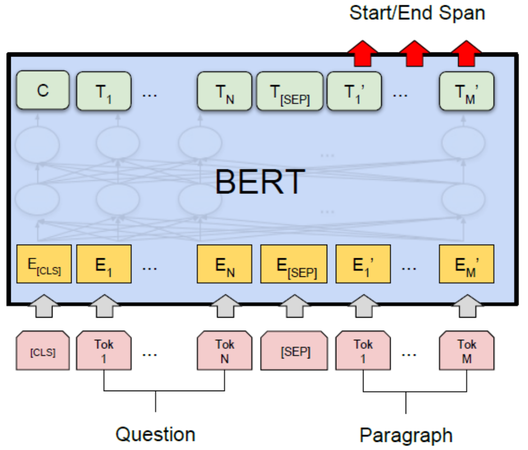
Typically, using BERT involves two stages: unsupervised pretraining followed by supervised task-specific fine-tuning.



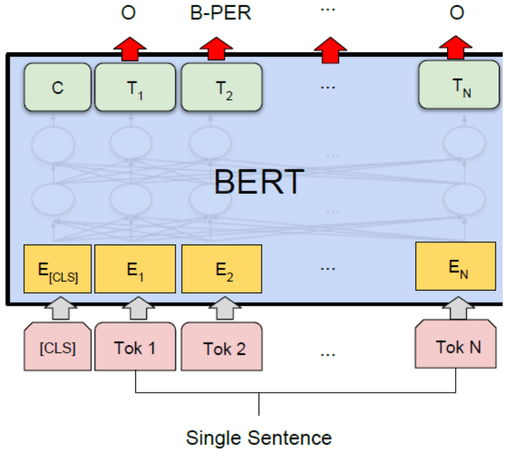
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



(b) Single Sentence Classification Tasks: SST-2, CoLA



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

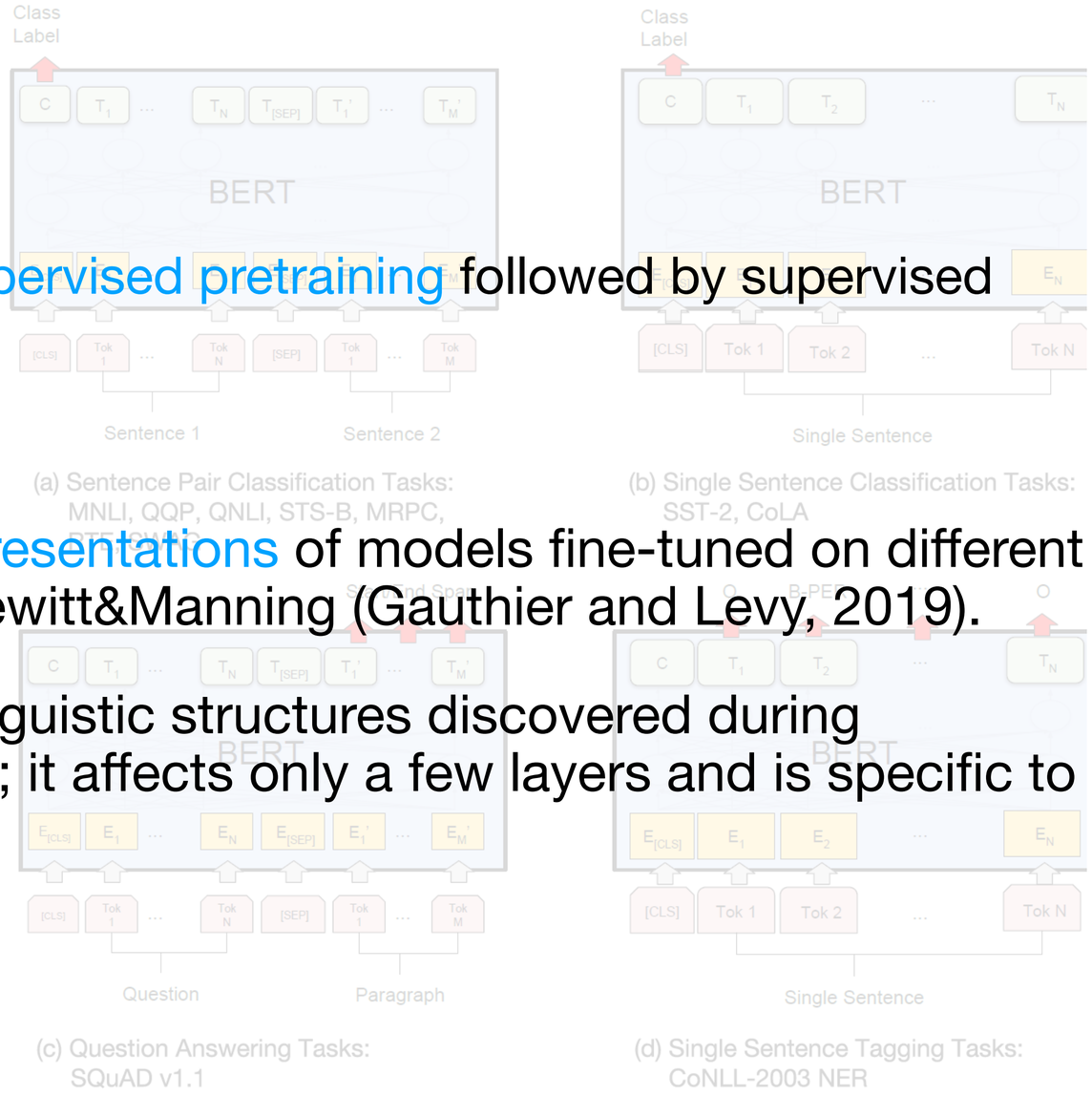
Source: Devlin et al. 2019

Fine-tuning BERT

Typically, using BERT involves two stages: unsupervised pretraining followed by supervised task-specific fine-tuning.

Evolution of internal representations:

- significant divergence between the final representations of models fine-tuned on different tasks when using the structural probe of Hewitt&Manning (Gauthier and Levy, 2019).
- fine-tuning is a conservative process: the linguistic structures discovered during pretraining remain available after fine-tuning; it affects only a few layers and is specific to in-domain examples (Merchant et al., 2020).



Experimental setup

Probing Methods

Hewitt & Manning structural probe (2019)

- Entire syntax trees are embedded implicitly in the vector geometry of the deep models in both ELMo and BERT

Tree distance evaluation

Evaluates how well the predicted distances between all pairs of words in a model reconstruct gold parse trees.

- UUAS (Undirected Unlabeled Attachment Score). Percentage of undirected edges placed correctly.
- Dspr. Spearman correlation between true and predicted distances for each word in each sentence.

Tree depth evaluation

Evaluates the ability to recreate the order of words specified by their depth in the parse tree.

- Root %. Ability of the models to identify the root of the sentence as the least deep word.
- Nspr. Spearman correlation between the predicted and the true depth ordering.

Experimental setup

- HuggingFace Transformers library 
- We fine-tune the 12-layer cased variant of BERT on 6 different tasks for 3 epochs, with a learning rate of $5e^{-5}$.
- We save 10 evenly-spaced checkpoints per epoch, and we probe layer 7 of each checkpoint using H&M probe.

Morpho-syntactic tasks:

- **PoS tagging** (English, UD 2.5 English EWT).
- **Constituency parsing** (English, PTB SD 3.3.0).
- **Dependency parsing**:
  - English: UD 2.5 English EWT
  - Multilingual UD: DE, EN, ES, FR, PT, SV UD 2.5.
  - English: PTB SD 3.3.0.

Semantics-related tasks:

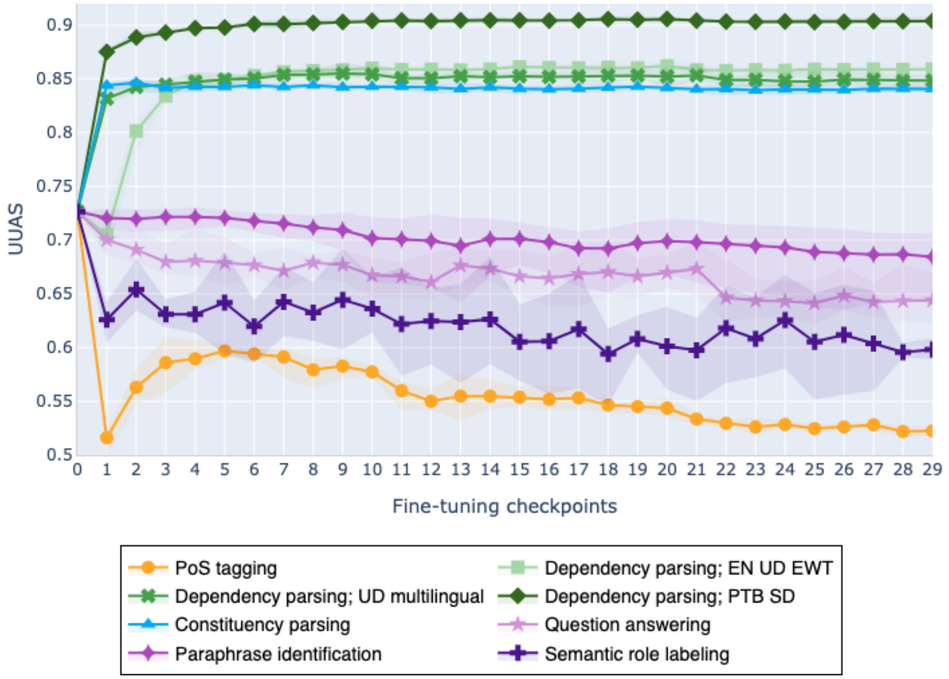
- **Semantic role labelling** (English, OntoNotes)
- **Question answering** (English, Stanford Question Answering Dataset)
- **Paraphrase identification** (English, Microsoft Research Paraphrase Corpus)

Evolution of syntax trees

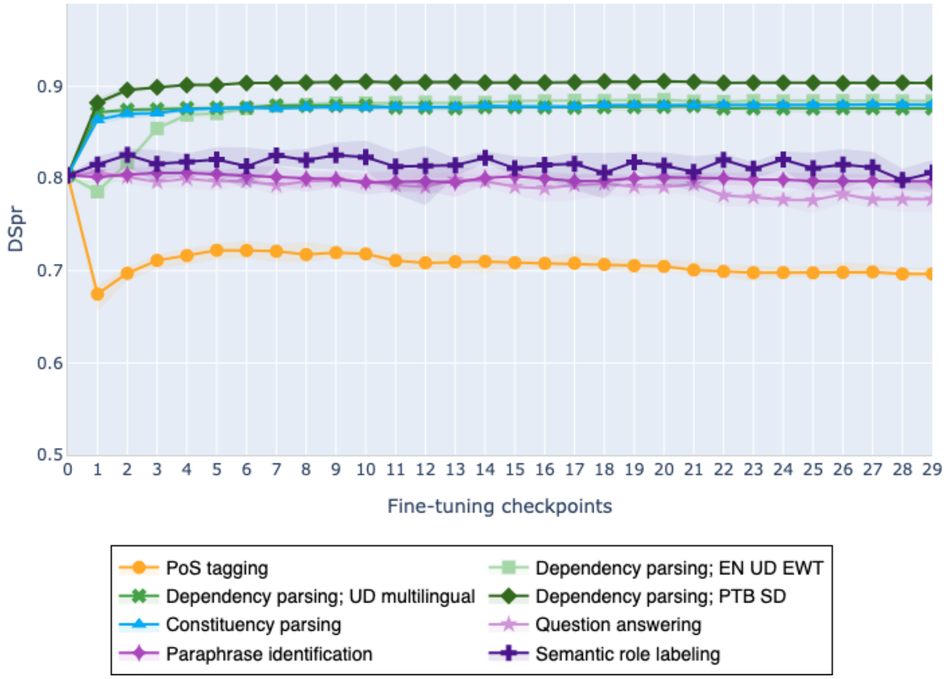
Tree distance evaluation

Evaluates how well the predicted distances between all pairs of words in a model reconstruct gold parse trees.

UUAS (Undirected Unlabeled Attachment Score). Percentage of undirected edges placed correctly.



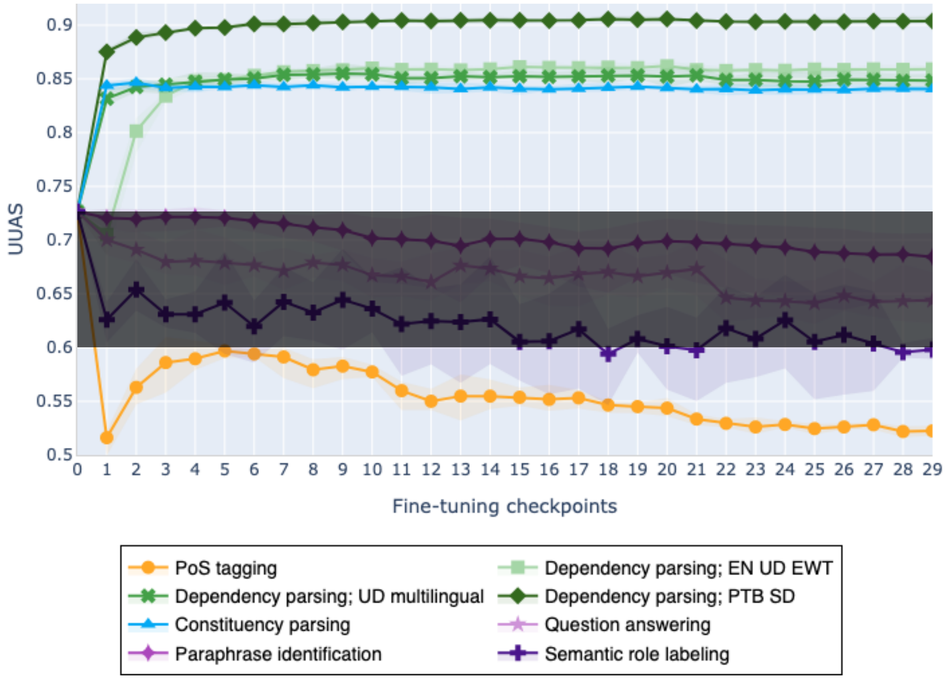
Dspr. Spearman correlation between true and predicted distances for each word in each sentence, averaging across all sentences with lengths between 5 and 50.



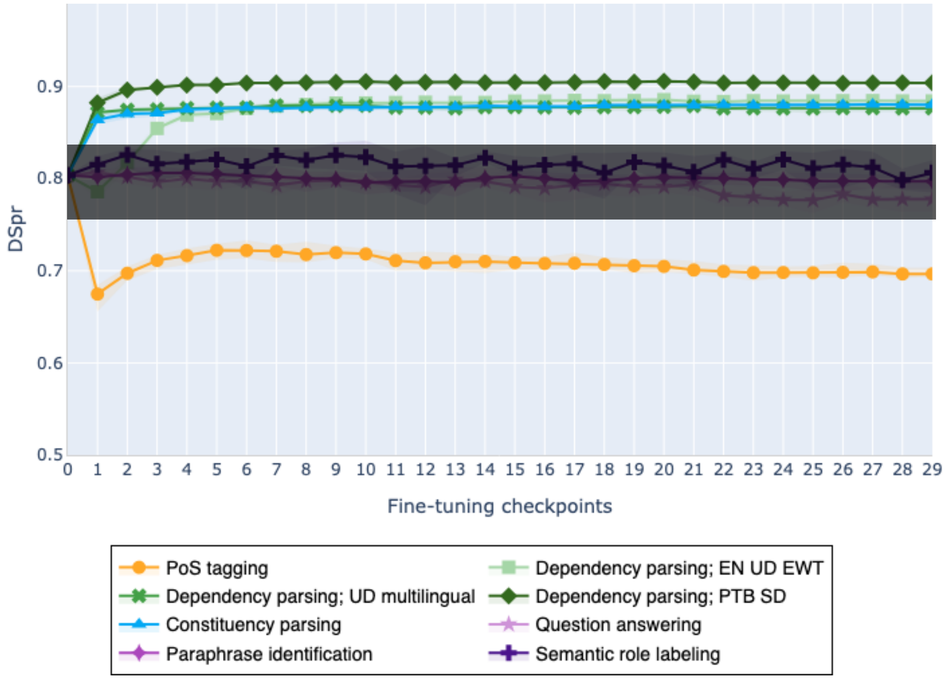
Tree distance evaluation

Evaluates how well the predicted distances between all pairs of words in a model reconstruct gold parse trees.

UUAS (Undirected Unlabeled Attachment Score). Percentage of undirected edges placed correctly.



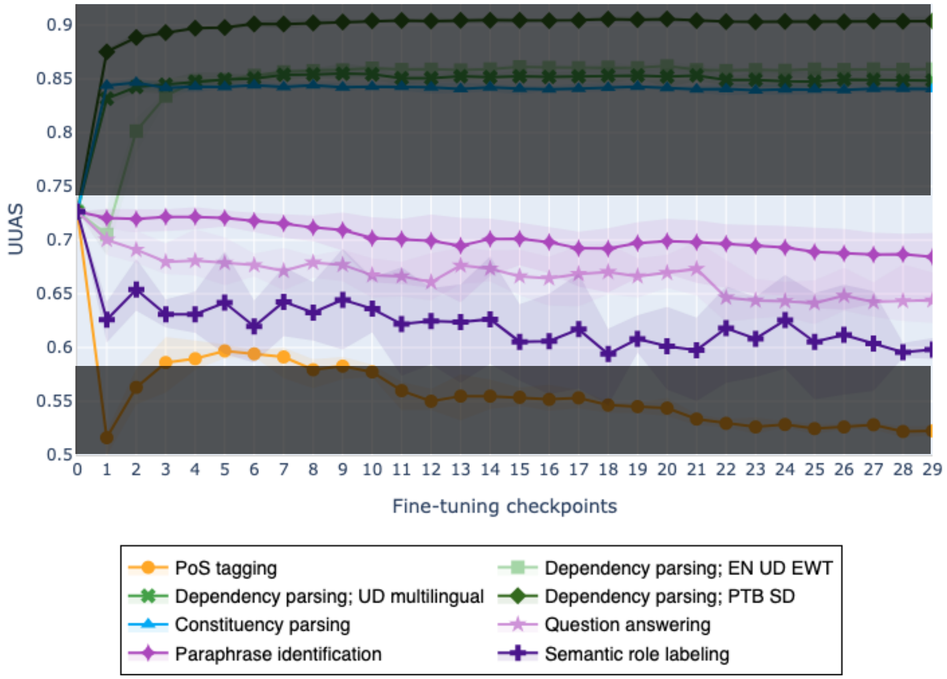
Dspr. Spearman correlation between true and predicted distances for each word in each sentence, averaging across all sentences with lengths between 5 and 50.



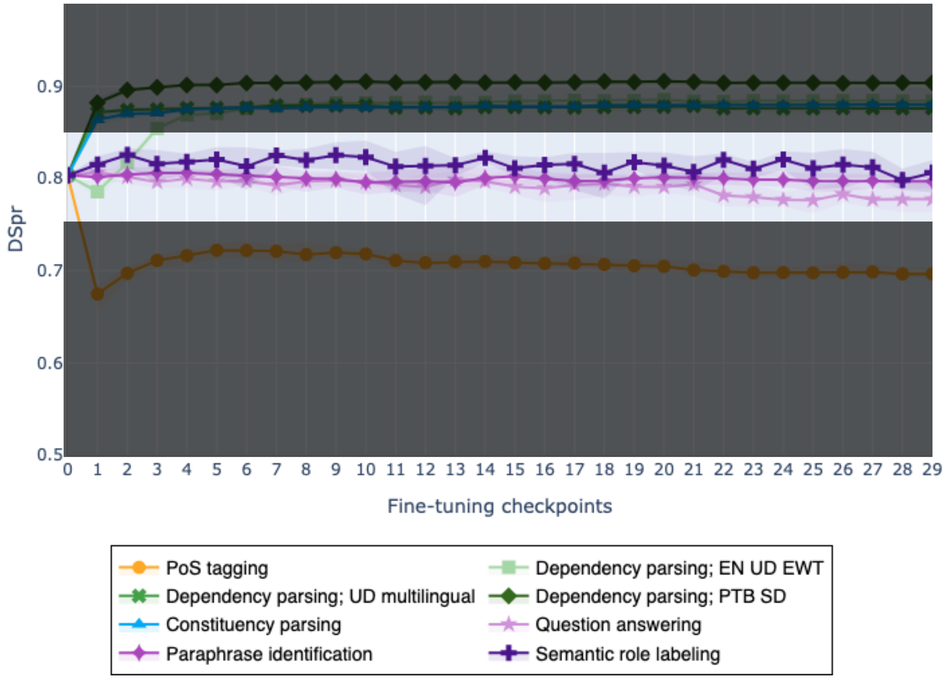
Tree distance evaluation

Evaluates how well the predicted distances between all pairs of words in a model reconstruct gold parse trees.

UUAS (Undirected Unlabeled Attachment Score). Percentage of undirected edges placed correctly.



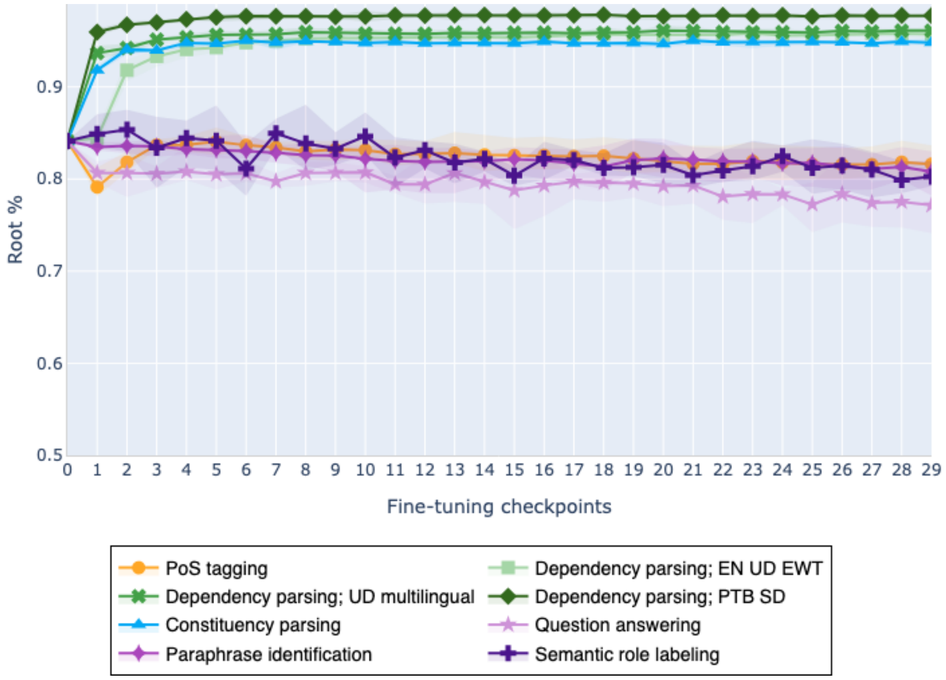
Dspr. Spearman correlation between true and predicted distances for each word in each sentence, averaging across all sentences with lengths between 5 and 50.



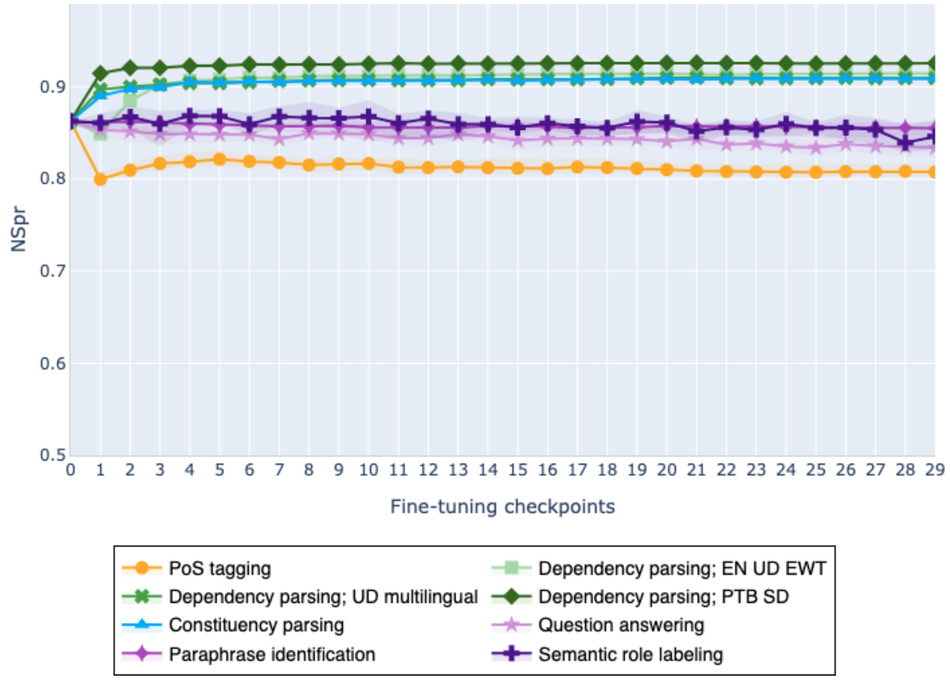
Tree depth evaluation

Evaluates models with respect to their ability to recreate the order of words specified by their depth in the parse tree.

Root %. Ability of the models to identify the root of the sentence as the least deep word.



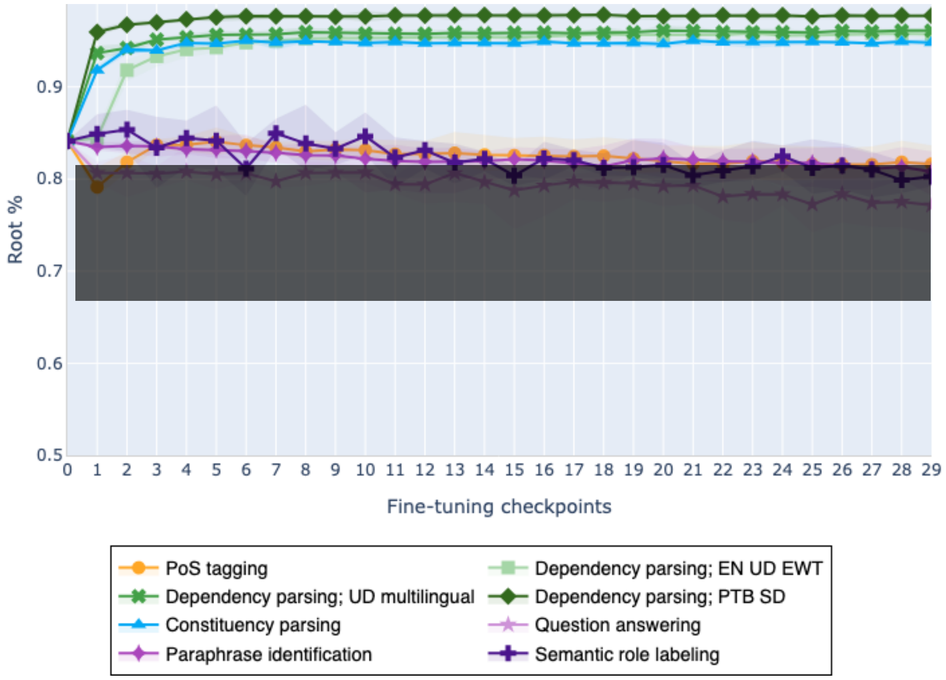
Nspr. Spearman correlation between the predicted and the true depth ordering, averaging across all sentences with lengths between 5 and 50.



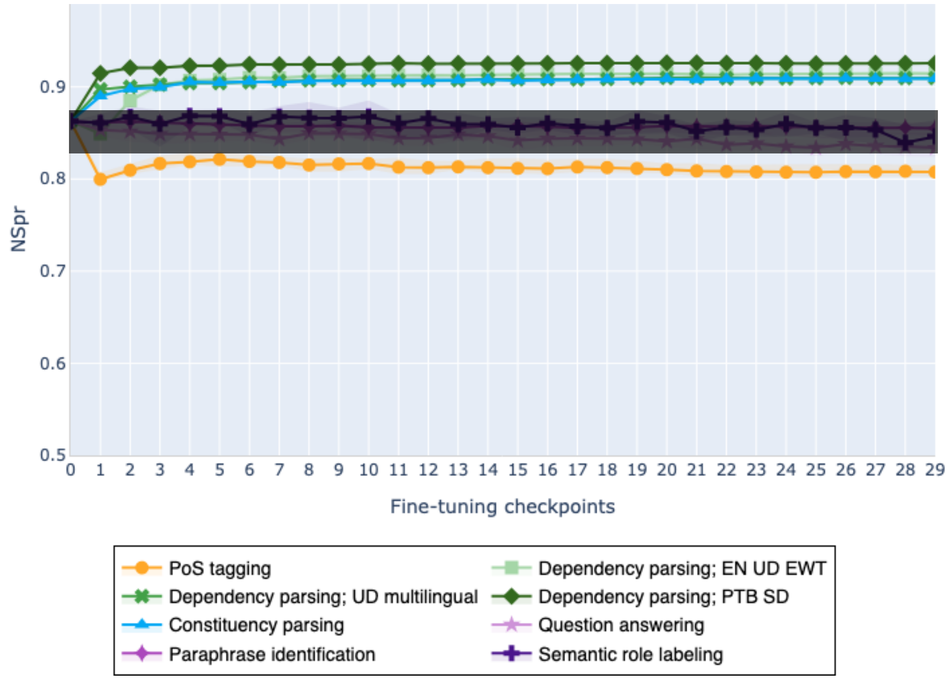
Tree depth evaluation

Evaluates models with respect to their ability to recreate the order of words specified by their depth in the parse tree.

Root %. Ability of the models to identify the root of the sentence as the least deep word.



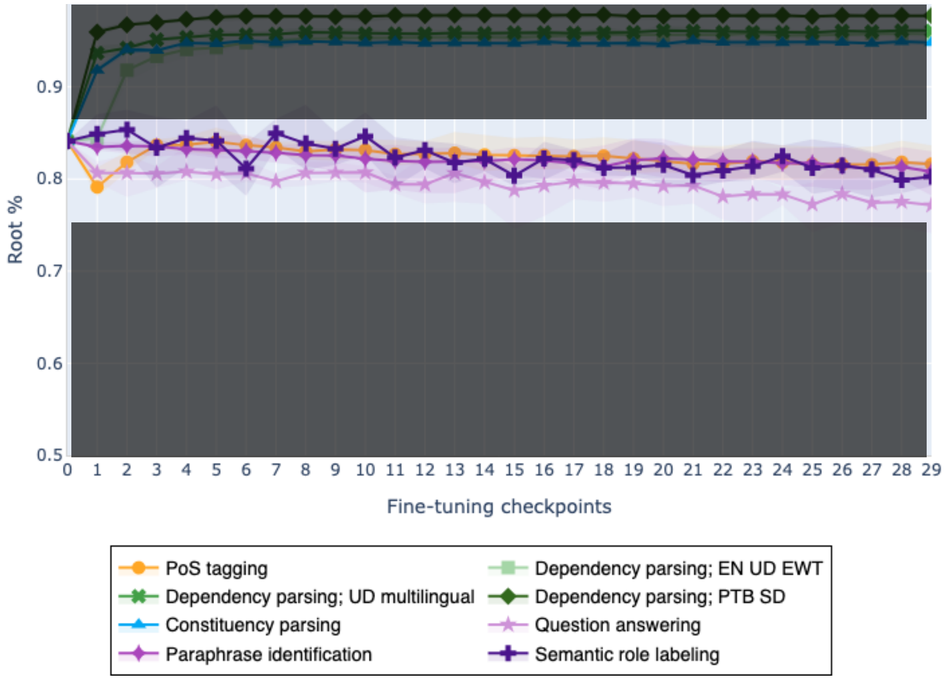
Nspr. Spearman correlation between the predicted and the true depth ordering, averaging across all sentences with lengths between 5 and 50.



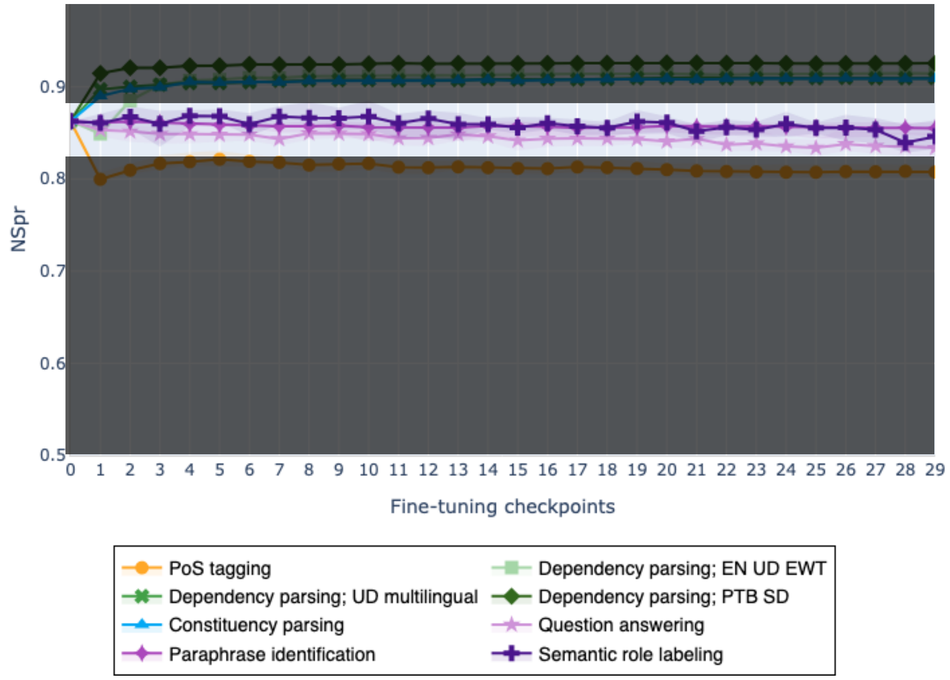
Tree depth evaluation

Evaluates models with respect to their ability to recreate the order of words specified by their depth in the parse tree.

Root %. Ability of the models to identify the root of the sentence as the least deep word.



Nspr. Spearman correlation between the predicted and the true depth ordering, averaging across all sentences with lengths between 5 and 50.



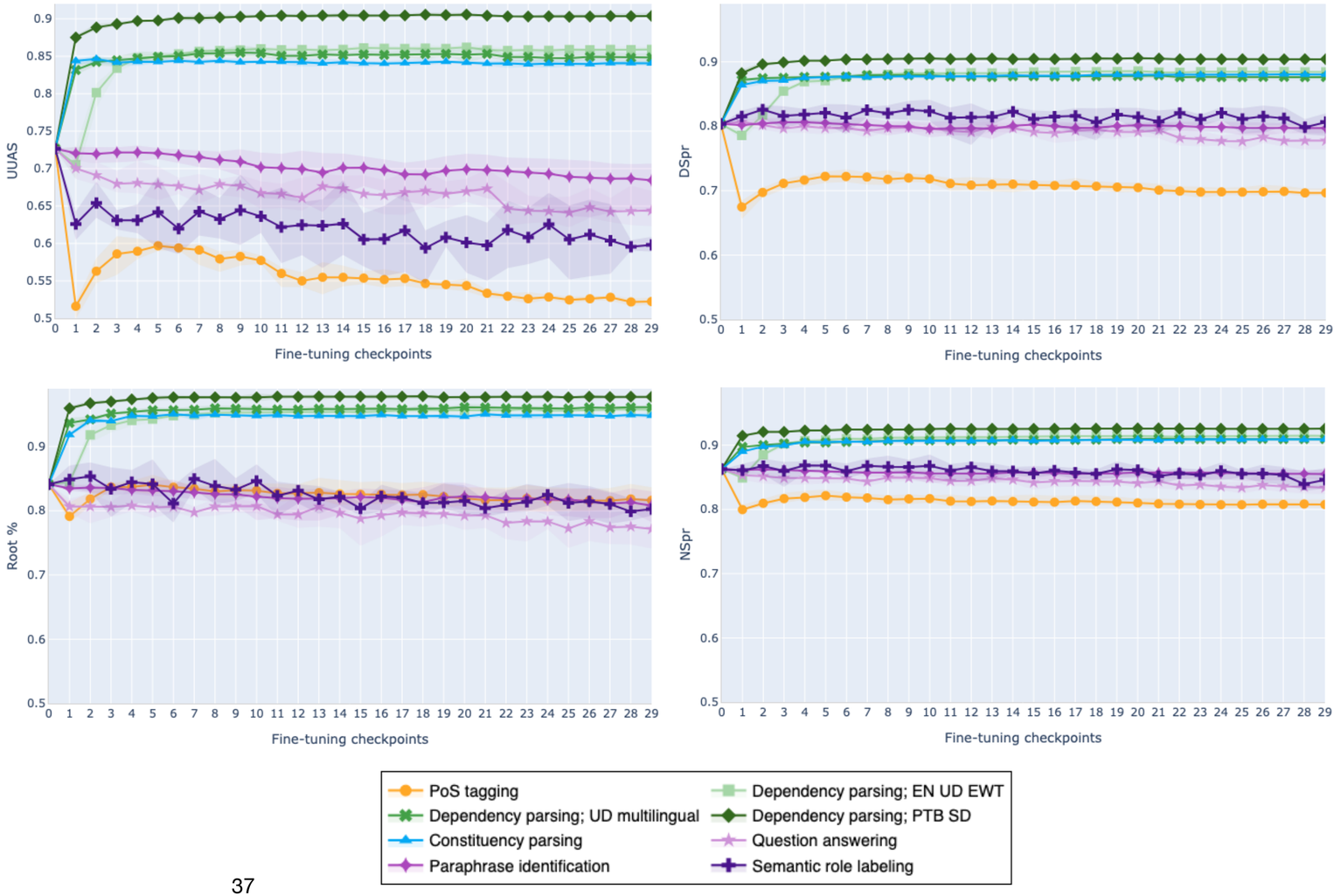
Conclusions

Conclusions

Fine-tuning is not always a conservative process.

The syntactic information initially encoded in the models is forgotten (PoS tagging), reinforced (parsing) or preserved (semantics-related tasks) in different ways along the fine-tuning.

- **Morpho-syntactic tasks** experiment substantial information gains in the initial phases.
- **Semantic-related tasks** maintain a more stable trend, mostly preserving the syntactic knowledge initially encoded.

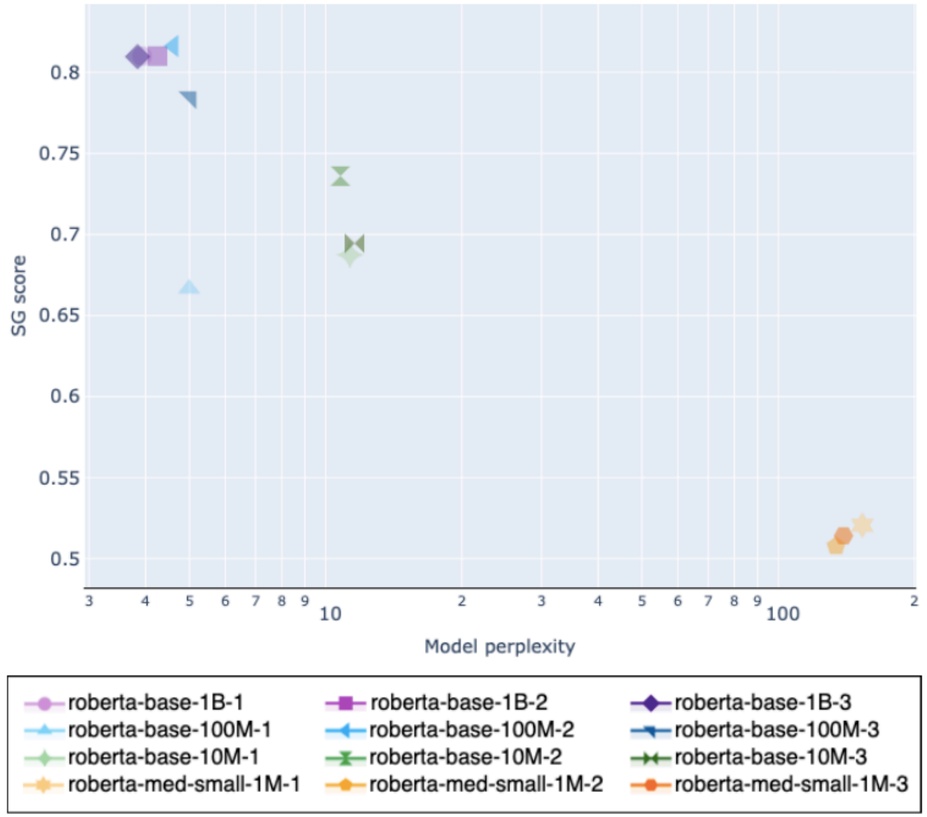


Further questions

Towards robust and efficient models

Perplexity and SyntaxGym Score seem to be measuring different aspects of the knowledge of the models.

Would it be possible to complement information-theoretical metrics such as perplexity with metrics measuring specific types of knowledge, e.g., syntax, in order to develop and select more robust and efficient models to address NLU tasks?



Thank you!

Comments, questions and feedback

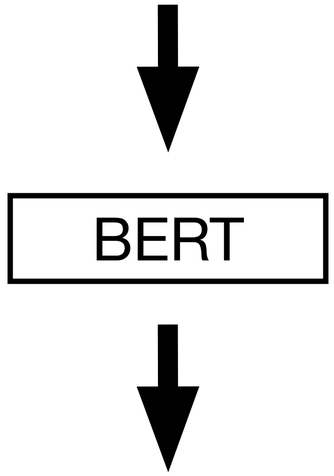
Laura Pérez-Mayos
TALN Research Group, Pompeu Fabra University, Barcelona, Spain
lpmayos@gmail.com

Hidden slides

Encoding unidirectional context with bidirectional models

Agreement test: The girls [run/runs] fast.

[[bos] [The] [girls] [mask] [mask] [mask] [mask]]



$p(run_{pos4}) > p(runs_{pos4})$

Constituency parsing

- Constituency Parsing is the process of analyzing the sentences by breaking down it into **sub-phrases** also known as **constituents**.
- These sub-phrases belong to a specific **category of grammar** like NP (noun phrase) and VP(verb phrase).