Infusing FineTuning with Semantic Dependencies

Zhaofeng Wu, Hao Peng, Noah A. Smith
Presented By Anuja Tayal

Transactions of the Association for Computational Linguistics, 2021.

Paradigm Shift

Before

General purpose linguistic modules (pos taggers) were constructed using supervised learning from linguistics datasets

Often used as preprocessing techniques

Today

General purpose representation learning

Pretraining

Representations are fine-tuned on application specific datasets

Introduction

- A syntactic parser may generate constituency or dependency trees from a sentence, but a semantic parser may be built depending upon the task for which inference is required.
- Requires understanding concepts from different word phrases
- Modern NLP models (BERT, GPT, etc) are typically trained in the end to end manner, without specifying the features explicitly
- Linguistic features (like part-of-speech, co-reference, etc) have played a key role in the classical NLP.
- Hence, it is important to understand how modern NLP models are arriving at decisions by "**probing**" into what all they learn.

Previous works

- Probing focused more on syntactic dependency
- Pretraining does not much useful information for entity labeling or coreference resolution
- Bert does not capture FrameNet
- Authors extend these findings to show for predicate-argument substructure

Contribution

Apply Novel probes to recent language models focusing on predicate-argument structure as operationalised by semantic dependencies

Unlike syntax, semantics is not brought to the surface by today's pre-trained models

Use convolutional graph encoders to explicitly incorporate semantic parses into task-specific finetuning

NLU tasks in GLUE Benchmarks

Predicate Argument Structure

The crocodile devoured a doughnut.

Crocodile - subject
Dougnut - object
Devoured - predicate

Predicate-Argument Semantics

- Experiments focus on DELPH-IN dependency (DM)
- Deep Linguistic Processing with HPSG - INitiative (DELPH-IN)
- DELPH-IN adopt two linguistic formalisms for deep linguistic analysis, viz. head-driven phrase structure grammar (HPSG) and minimal recursion semantics (MRS).
- Vertices are words
- 59 labels to characterize argument and adjunct relationships

Syntactic and Semantic Dependency

- Both highlight bilexical relationships
- Semantically empty words are excluded from semantic graphs (is, to)
- Technique->apply, impossible->apply (connection between semantic related words)

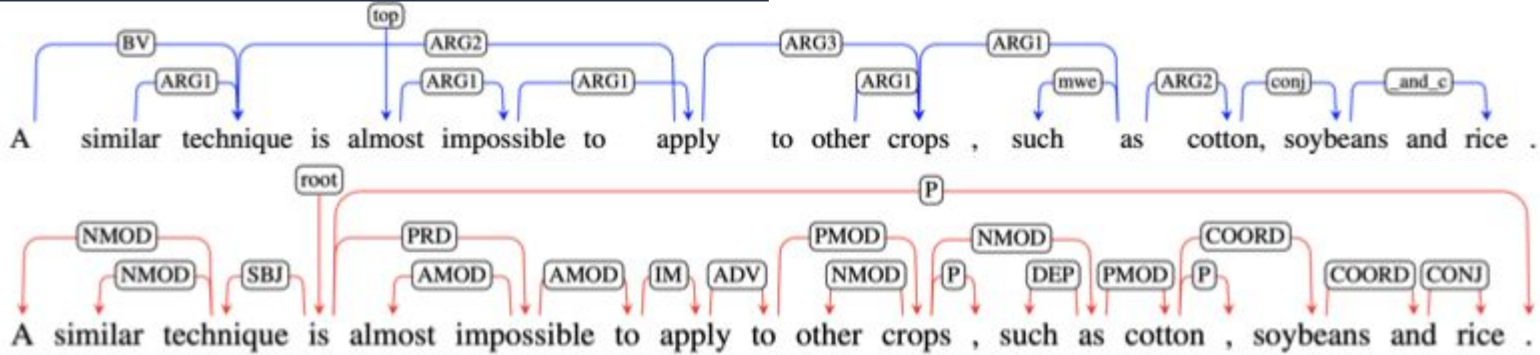

top
BV
ARG2
ARG1
ARG1
ARG1
ARG3
ARG1
ARG1
mwe
ARG2
conj
_and_c
A similar technique is almost impossible to apply to other crops , such as cotton, soybeans and rice .
root
P
NMOD
NMOD
SBJ
PRD
AMOD
AMOD
IM
ADV
PMOD
NMOD
P
NMOD
DEP
PMOD
COORD
P
COORD
CONJ
A similar technique is almost impossible to apply to other crops , such as cotton , soybeans and rice .

Figure 1: An example sentence in the DM (top, blue) and Stanford Dependencies (bottom, red) format, taken from Oepen et al. (2015) and Ivanova et al. (2012).

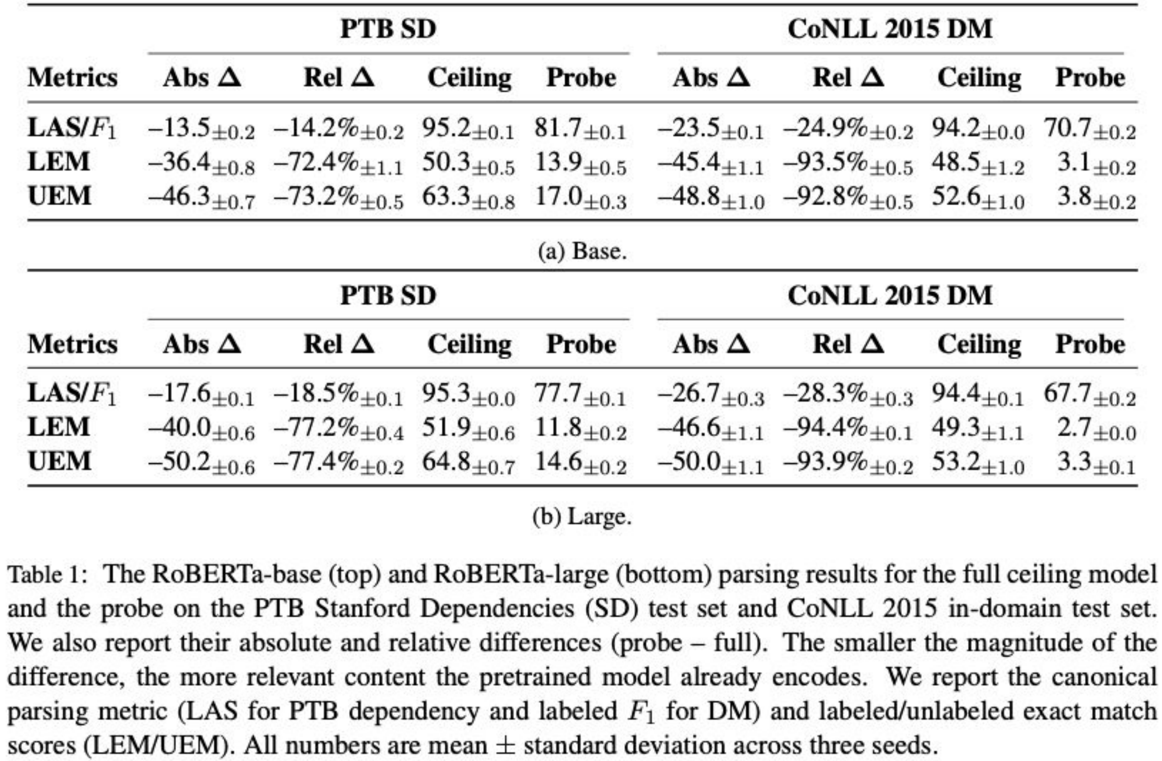
Probing RoBERTa

- Linguistic probing
- Determine the level a pretrained model incorporates linguistic abstraction
- Select annotated dataset
- Pretain- Considered RoBERTa and BERT
- Train end-to-end model- ceiling model- finetuning
- Train a supervised “probe model” with pretrained representations(frozen). - Use linear probe classifier
- Compare on held out data the probe model and ceiling model
- Learn the extent to which pertained model learns?

Results

- Based on Dozat n Manning Parser
- Labeled attachment score for dependency parsing
- Labeled F1 score for semantic parsing
- Use Chu-Liu-Edmonds Algorithm to decode syntactic dependency trees

- Difference between probe and ceiling
- Previous works- semantic probe will exhibit larger difference than syntactic one
- Pretraining surfaces syntactic abstractions than semantic
- Larger pretrained models do not necessarily come with better structural information



	PTB SD				CoNLL 2015 DM			
Metrics	**Abs Δ**	**Rel Δ**	**Ceiling**	**Probe**	**Abs Δ**	**Rel Δ**	**Ceiling**	**Probe**
LAS/F_1	–13.5$_{\pm 0.2}$	–14.2%$_{\pm 0.2}$	95.2$_{\pm 0.1}$	81.7$_{\pm 0.1}$	–23.5$_{\pm 0.1}$	–24.9%$_{\pm 0.2}$	94.2$_{\pm 0.0}$	70.7$_{\pm 0.2}$
LEM	–36.4$_{\pm 0.8}$	–72.4%$_{\pm 1.1}$	50.3$_{\pm 0.5}$	13.9$_{\pm 0.5}$	–45.4$_{\pm 1.1}$	–93.5%$_{\pm 0.5}$	48.5$_{\pm 1.2}$	3.1$_{\pm 0.2}$
UEM	–46.3$_{\pm 0.7}$	–73.2%$_{\pm 0.5}$	63.3$_{\pm 0.8}$	17.0$_{\pm 0.3}$	–48.8$_{\pm 1.0}$	–92.8%$_{\pm 0.5}$	52.6$_{\pm 1.0}$	3.8$_{\pm 0.2}$

(a) Base.

	PTB SD				CoNLL 2015 DM			
Metrics	**Abs Δ**	**Rel Δ**	**Ceiling**	**Probe**	**Abs Δ**	**Rel Δ**	**Ceiling**	**Probe**
LAS/F_1	–17.6$_{\pm 0.1}$	–18.5%$_{\pm 0.1}$	95.3$_{\pm 0.0}$	77.7$_{\pm 0.1}$	–26.7$_{\pm 0.3}$	–28.3%$_{\pm 0.3}$	94.4$_{\pm 0.1}$	67.7$_{\pm 0.2}$
LEM	–40.0$_{\pm 0.6}$	–77.2%$_{\pm 0.4}$	51.9$_{\pm 0.6}$	11.8$_{\pm 0.2}$	–46.6$_{\pm 1.1}$	–94.4%$_{\pm 0.1}$	49.3$_{\pm 1.1}$	2.7$_{\pm 0.0}$
UEM	–50.2$_{\pm 0.6}$	–77.4%$_{\pm 0.2}$	64.8$_{\pm 0.7}$	14.6$_{\pm 0.2}$	–50.0$_{\pm 1.1}$	–93.9%$_{\pm 0.2}$	53.2$_{\pm 1.0}$	3.3$_{\pm 0.1}$

(b) Large.

Table 1: The RoBERTa-base (top) and RoBERTa-large (bottom) parsing results for the full ceiling model and the probe on the PTB Stanford Dependencies (SD) test set and CoNLL 2015 in-domain test set. We also report their absolute and relative differences (probe – full). The smaller the magnitude of the difference, the more relevant content the pretrained model already encodes. We report the canonical parsing metric (LAS for PTB dependency and labeled F_1 for DM) and labeled/unlabeled exact match scores (LEM/UEM). All numbers are mean ± standard deviation across three seeds.

Semantics Infused Finetuning (SIFT)

- Proposed Semantics Infused FineTuning
- Input sentences are passed through semantic dependency parser
- Used DELPH-IN MRS derived dependency (DM)
- Architecture learned is combined with pretrained model (RoBERTa) with relational graph convolutional network (RGCN)

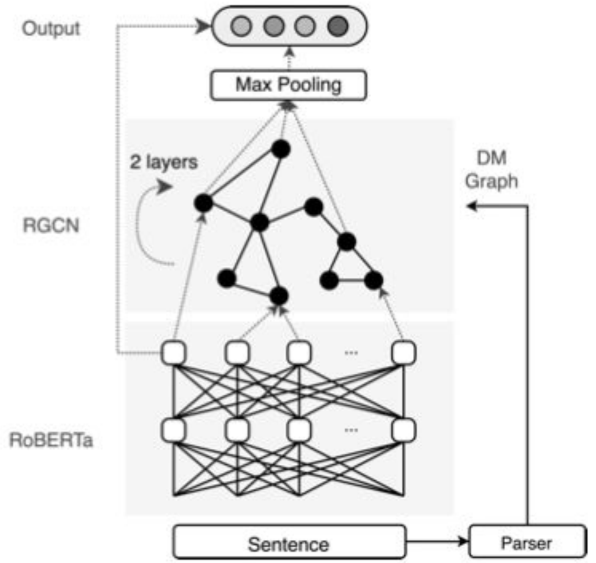


Figure 2: SIFT architecture. The sentence is first contextualized using RoBERTa, and then parsed. RGCN encodes the graph structures on top of RoBERTa. We max-pool over the RGCN's outputs for onward computation.

SIFT Light

- Lightweight variant of SIFT that aim to reduce time and memory overhead at test time
- Learns 2 classifier
  - Main linear classifier on top of RoBERTa $f_{RoBERTa}$
  - An auxiliary classifier f_{RGCN} based on SIFT

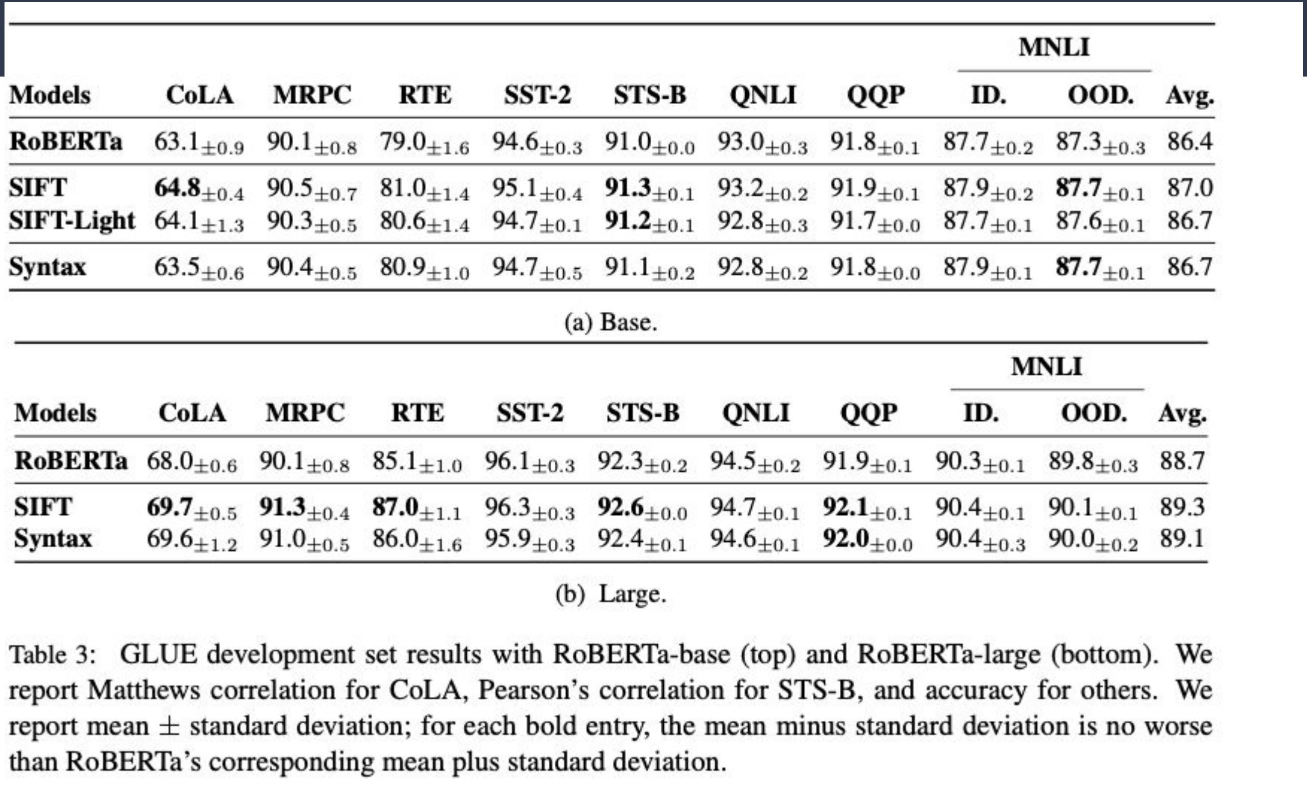
Discussion

- Previous works used GCN to encode unlabeled syntactic structures
- Used RGCN to encode labeled semantic graphs
- Outperforms GCN
- Number of parameters linearly increases with the number of relation types

- Incorporates semantic information during finetuning
- Previous work incorporated in pretraining

Experiments

- Used GLUE dataset
- Ran models across 3 seeds for large datasets (QNLI, MNLI, QQP)
- 4 seeds among others
- SIFT used 2 RGCN layers
- Used semantic parsing
  - Held 1st place in CoNLL 2019 shared task
  - 92.5 labeled F1 for DM



								MNLI		
Models	**CoLA**	**MRPC**	**RTE**	**SST-2**	**STS-B**	**QNLI**	**QQP**	**ID.**	**OOD.**	**Avg.**
RoBERTa	63.1$_{\pm 0.9}$	90.1$_{\pm 0.8}$	79.0$_{\pm 1.6}$	94.6$_{\pm 0.3}$	91.0$_{\pm 0.0}$	93.0$_{\pm 0.3}$	91.8$_{\pm 0.1}$	87.7$_{\pm 0.2}$	87.3$_{\pm 0.3}$	86.4
SIFT	**64.8**$_{\pm 0.4}$	90.5$_{\pm 0.7}$	81.0$_{\pm 1.4}$	95.1$_{\pm 0.4}$	**91.3**$_{\pm 0.1}$	93.2$_{\pm 0.2}$	91.9$_{\pm 0.1}$	87.9$_{\pm 0.2}$	**87.7**$_{\pm 0.1}$	87.0
SIFT-Light	64.1$_{\pm 1.3}$	90.3$_{\pm 0.5}$	80.6$_{\pm 1.4}$	94.7$_{\pm 0.1}$	**91.2**$_{\pm 0.1}$	92.8$_{\pm 0.3}$	91.7$_{\pm 0.0}$	87.7$_{\pm 0.1}$	87.6$_{\pm 0.1}$	86.7
Syntax	63.5$_{\pm 0.6}$	90.4$_{\pm 0.5}$	80.9$_{\pm 1.0}$	94.7$_{\pm 0.5}$	91.1$_{\pm 0.2}$	92.8$_{\pm 0.2}$	91.8$_{\pm 0.0}$	87.9$_{\pm 0.1}$	**87.7**$_{\pm 0.1}$	86.7

(a) Base.

								MNLI		
Models	**CoLA**	**MRPC**	**RTE**	**SST-2**	**STS-B**	**QNLI**	**QQP**	**ID.**	**OOD.**	**Avg.**
RoBERTa	68.0$_{\pm 0.6}$	90.1$_{\pm 0.8}$	85.1$_{\pm 1.0}$	96.1$_{\pm 0.3}$	92.3$_{\pm 0.2}$	94.5$_{\pm 0.2}$	91.9$_{\pm 0.1}$	90.3$_{\pm 0.1}$	89.8$_{\pm 0.3}$	88.7
SIFT	**69.7**$_{\pm 0.5}$	**91.3**$_{\pm 0.4}$	**87.0**$_{\pm 1.1}$	96.3$_{\pm 0.3}$	**92.6**$_{\pm 0.0}$	94.7$_{\pm 0.1}$	**92.1**$_{\pm 0.1}$	90.4$_{\pm 0.1}$	90.1$_{\pm 0.1}$	89.3
Syntax	69.6$_{\pm 1.2}$	91.0$_{\pm 0.5}$	86.0$_{\pm 1.6}$	95.9$_{\pm 0.3}$	92.4$_{\pm 0.1}$	94.6$_{\pm 0.1}$	**92.0**$_{\pm 0.0}$	90.4$_{\pm 0.3}$	90.0$_{\pm 0.2}$	89.1

(b) Large.

Table 3: GLUE development set results with RoBERTa-base (top) and RoBERTa-large (bottom). We report Matthews correlation for CoLA, Pearson's correlation for STS-B, and accuracy for others. We report mean ± standard deviation; for each bold entry, the mean minus standard deviation is no worse than RoBERTa's corresponding mean plus standard deviation.

When do Semantic Structures help?

- SIFT helps guards the model against frequent but invalid heuristics in the data
- Better captures nuanced sentence level linguistic phenomenon than RoBERTa

Results

- Performing experiment on HANS Dataset
- Performing Ablation studies
-

Conclusion

- Presented strong evidence that RoBERTa and BERT do not bring predicate argument substructure as compared to syntactic dependency
- Propose SIFT to incorporate semantic structure
- Presented experiments and results to support their hypothesis

Thank You

Questions?