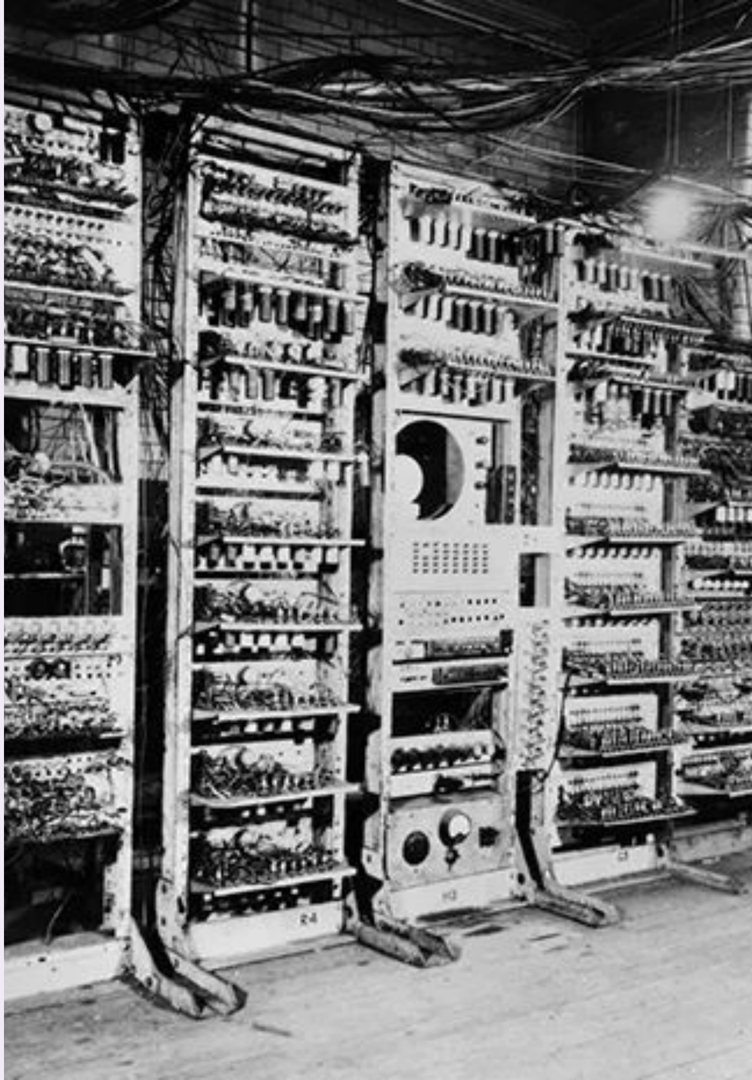
Systematic Language Understanding

Exploring the limits of systematicity of natural language models

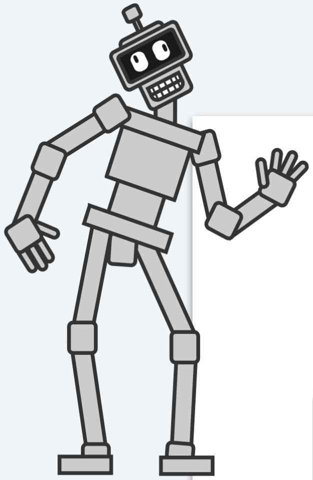
Koustuv Sinha

PhD Thesis Defense, McGill University



Natural Language Understanding (NLU)

- Goal: understand ambiguous and contextual natural language
- Several tasks in NLU to measure progress (NLI, Q&A, RC, WSD etc)
- State-of-the-art Pre-trained **Transformer**-based architectures outperform most baselines



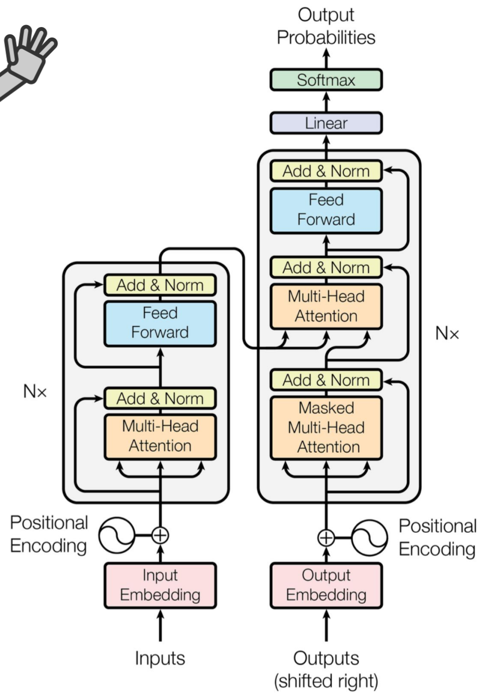
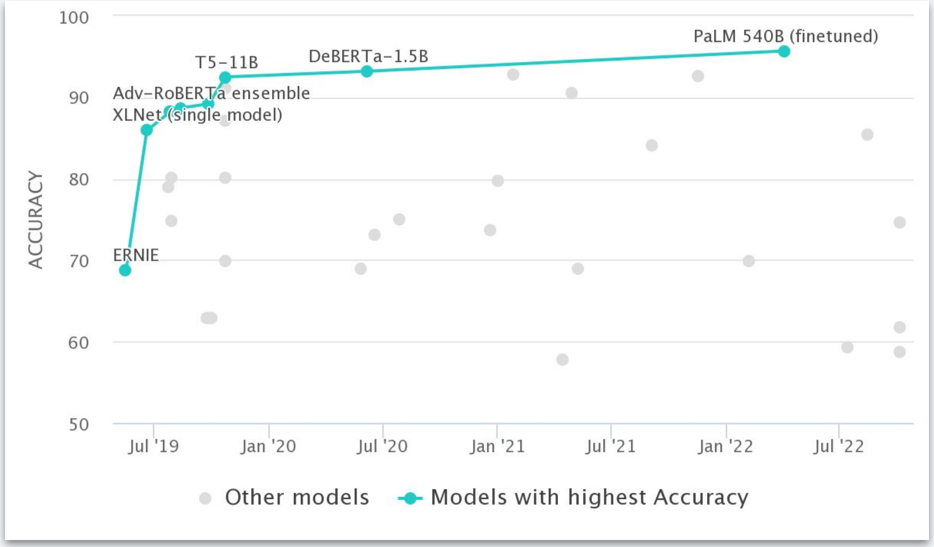


Figure 1: The Transformer - model architecture.

Impressive success of Pre-trained Transformers

- Success recipe: Pre-training on large data
- Impressive performance by Transformers!
- More parameters and more data -> Scaling is all you need? [1]



[1] Scaling laws for Neural Language Models, Kaplan et al 2021

How are these models so successful?

- **Probing** - proxy to evaluate latent knowledge by learning a function

- Large pre-trained models have been shown to contain [1]:
    - **Semantic** knowledge
    - **Syntax** knowledge
    - **World** knowledge

Emergent linguistic structure in artificial neural networks trained by self-supervision

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy

[a]*Computer Science Department, Stanford University, Stanford, CA 94305;*

[b]*Facebook Artificial Intelligence Research, Facebook Inc., Seattle, WA 98109*

– Hide authors and affiliations

PNAS December 1, 2020 117 (48) 30046-30054; first published June 3, 2020; https://doi.org/10.1073/pnas.1907367117

Edited by Matan Gavish, Hebrew University of Jerusalem, Jerusalem, Israel, and accepted by Editorial Board Member David L. Donoho April 13, 2020 (received for review June 3, 2019)

[1] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842-866.

However, models are not robust

Issues of Large Language Models:

- Brittle to adversarial input
- Exploit statistical artefacts
- Leverage spurious correlations
- Employ simple heuristics

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Are we really testing generalization?

- Benchmarks contain exploitable, unwanted statistical and social biases
- Increase in model parameters -> reduction of “distributional gap” -> dataset saturation
- Need dynamic, updated datasets to test for generalization

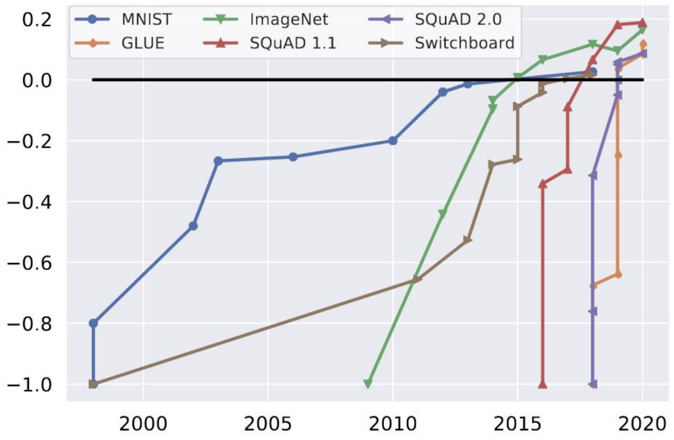


Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., & Williams, A. (2021). Dynabench: Rethinking Benchmarking in NLP. ArXiv, abs/2104.14337.. NAACL 2021

Thesis overview: measuring NLU progress through *systematicity*

The ability to produce / understand some sentences is intrinsically connected to the ability to produce / understand certain others

Fodor & Pylyshyn, 1988

Thesis overview: measuring NLU progress through *systematicity*

We humans are **consistent** in our language understanding in different contexts.

- ✅ We can reason consistently once we learn the rules
- ✅ We fail to understand consistently on inputs which doesn't agree with our learned rules

Investigations in Systematicity

Measuring consistency in reasoning

- CLUTRR: A diagnostic benchmark for inductive reasoning from text

  K Sinha, S Sodhani, J Dong, J Pineau, W Hamilton; EMNLP 2019 (Oral)

- Probing Linguistic Systematicity

  E Goodwin, K Sinha, T J O'Donnell; ACL 2020

- Measuring Systematic Generalization in Neural Proof Generation with Transformers

  N Gontier, K Sinha, S Reddy, C Pal; NeurIPS 2020

Measuring consistency in understanding

- UnNatural Language Inference

  K Sinha, P Parthasarathi, J Pineau, A Williams; ACL 2021 (Oral, Outstanding Paper Award)

- Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little

  K Sinha, R Jia, D Hupkes, J Pineau, A Williams, D Kiela; EMNLP 2021

- Sometimes we want ungrammatical translations

  P Parthasarathi, K Sinha, J Pineau, A Williams; EMNLP Findings 2021

- The Curious Case of Absolute Position Embeddings

  K Sinha, A Kazemnejad, S Reddy, J Pineau, D Hupkes, A Williams; EMNLP Findings 2022

Investigations in Systematicity

Measuring consistency in reasoning

- CLUTRR: A diagnostic benchmark for inductive reasoning from text
  K Sinha, S Sodhani, J Dong, J Pineau, W Hamilton; EMNLP 2019 (Oral)
- Probing Linguistic Systematicity
  E Goodwin, K Sinha, T J O'Donnell; ACL 2020
- Measuring Systematic Generalization in Neural Proof Generation with Transformers
  N Gontier, K Sinha, S Reddy, C Pal; NeurIPS 2020

Measuring consistency in understanding

- UnNatural Language Inference
  K Sinha, P Parthasarathi, J Pineau, A Williams; ACL 2021 (Oral, Outstanding Paper Award)
- Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little
  K Sinha, R Jia, D Hupkes, J Pineau, A Williams, D Kiela; EMNLP 2021
- Sometimes we want ungrammatical translations
  P Parthasarathi, K Sinha, J Pineau, A Williams; EMNLP Findings 2021
- The Curious Case of Absolute Position Embeddings
  K Sinha, A Kazemnejad, S Reddy, J Pineau, D Hupkes, A Williams; EMNLP Findings 2022

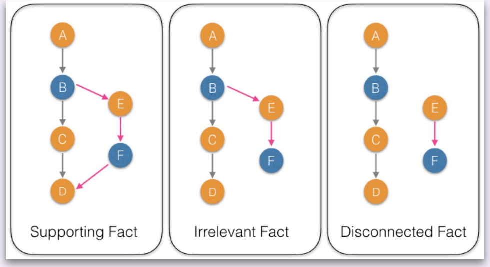
CLUTRR

A Diagnostic Benchmark for Inductive Reasoning from Text

K Sinha, S Sodhani, J Dong, J Pineau, WL. Hamilton

EMNLP 2019

Oral



Measuring reasoning through Question Answering

SQuAD2.0
The Stanford Question Answering Dataset

- Several datasets available, such as SQuAD, COQA, etc.
- Explicit reasoning
- Models surpass human accuracy

The English name "Normans" comes from the French words Normans/Normanz, plural of Normant, modern French normand, which is itself borrowed from Old Low Franconian Nortmann "Northman" or directly from Old Norse Norðmaðr, Latinized variously as Nortmannus, Normannus, or Nordmannus (recorded in Medieval Latin, 9th century) to mean "Norseman, Viking".

What is the original meaning of the word Norman?
Ground Truth Answers: Viking | Norseman, Viking | Norseman, Viking
Prediction: Norseman, Viking

When was the Latin version of the word Norman first recorded?
Ground Truth Answers: 9th century | 9th century | 9th century
Prediction: 9th century

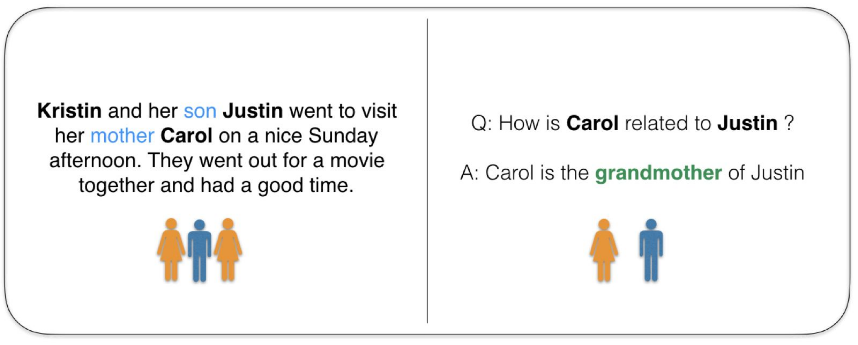
Measuring consistency in reasoning

- Implicit reasoning
- Finite set of rules

Son(Kristin, Justin) + Mother(Kristin, Carol) = grandmother(Justin, Carol)

CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text

Koustuv Sinha [1,3,4], **Shagun Sodhani** [2,3], **Jin Dong** [1,3],
Joelle Pineau [1,3,4] and **William L. Hamilton** [1,3,4]
[1] School of Computer Science, McGill University, Canada
[2] Université de Montréal, Canada
[3] Montreal Institute of Learning Algorithms (Mila), Canada
[4] Facebook AI Research (FAIR), Montreal, Canada



Measuring consistency in reasoning

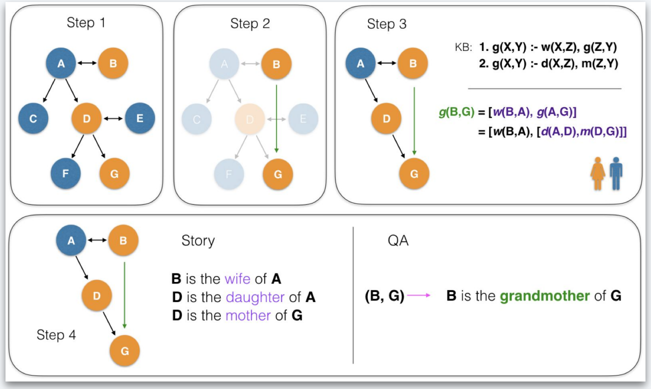
- **Length Generalization**
- Reasoning gets more complex
- Data is *procedurally generated*

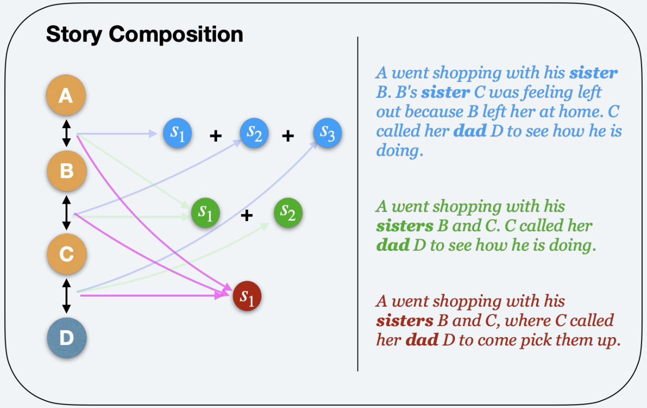
Sister(Mario, Marianne) + Mother(Jean, Marianne) + Sister(Jean, Darlene) + Brother(Darlene, Roy) + Father(Teri, Mario) + Daughter(Agnes, Teri) = Nephew(Agnes, Roy)

Puzzle	Question	Gender	Answer
Mario wanted to get a good gift for his sister, *Marianne Jean* and her sister *Darlene* were going to a party held by *Jean*'s mom, *Marianne*. *Darlene* invited her brother *Roy* to come, too, but he was too busy. *Teri* and her father, *Mario*, had an argument over the weekend. However, they made up by Monday. *Agnes* wants to make a special meal for her daughter *Teri*'s birthday.	Roy is the _____ of Agnes	Agnes:female, Teri:female, Mario:male, Marianne:female, Jean:female, Darlene:female, Roy:male	nephew

Procedural Data Generation

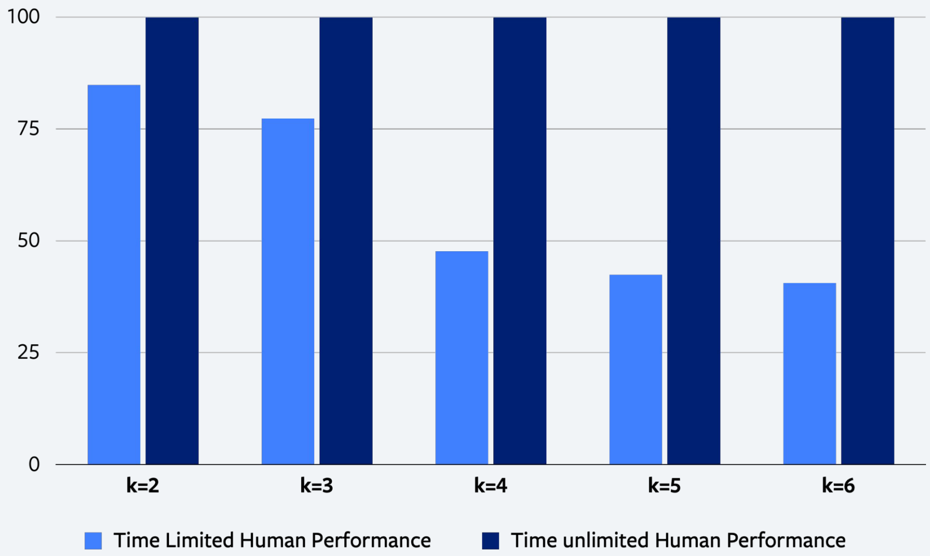
- Start from a predefined "Rule Base"
- Generate graphs.
- Sample an edge
- Sample a path enclosing the edge
- Stitch to a story!





How do we (humans) do?

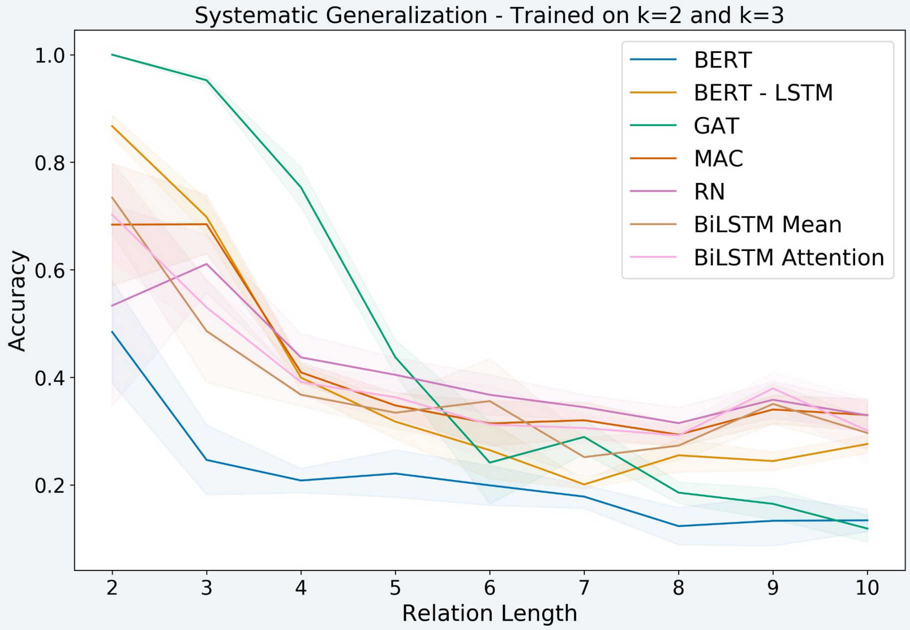
- Humans find the task difficult in a time limited setting
- Given **unlimited time,** human workers were able to solve the task with perfect accuracy



Relation Length	Human Performance Time Limited	Human Performance Unlimited Time	Reported Difficulty
2	0.848	1	1.488 +- 1.25
3	0.773	1	2.41 +- 1.33
4	0.477	1	3.81 +- 1.46
5	0.424	1	3.78 +- 0.96
6	0.406	1	4.46 +- 0.87

Q1. Are models able to generalize systematically?

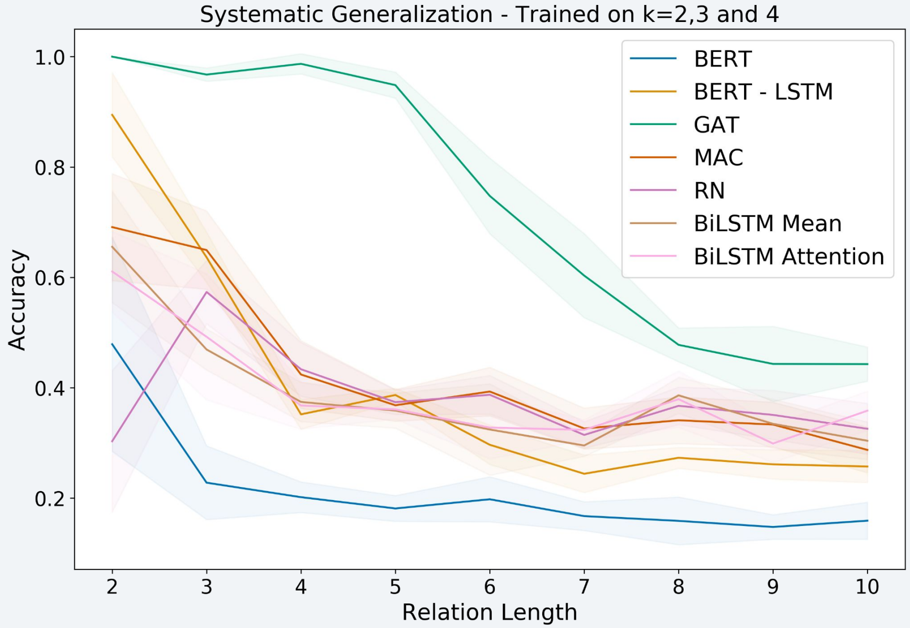
- Train on stories less combinations and test on longer combinations
- Ensure model sees all logical kinship rules during training, but not all combinations of those rules
- Split the AMT templates into train and test



Graph Attention Networks
BiLSTM, Relation Network, MAC, BERT, BERT-LSTM

Q1. Are models able to generalize systematically?

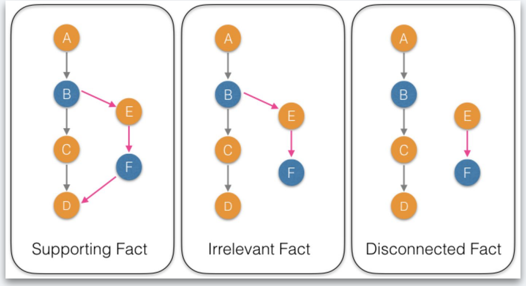
- Train on stories less combinations and test on longer combinations
- Ensure model sees all logical kinship rules during training, but not all combinations of those rules
- Split the AMT templates into train and test



Graph Attention Networks
BiLSTM, Relation Network, MAC, BERT, BERT-LSTM

Q2. Do models reason robustly?

- Supporting fact
- Irrelevant fact
- Disconnected fact

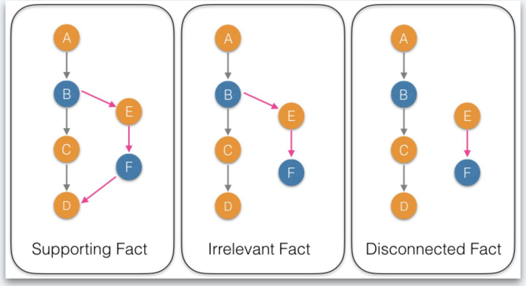


Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Clean	Clean	0.58 ±0.05	0.53 ±0.05	0.49 ±0.06	0.63 ±0.08	0.37 ±0.06	0.67 ±0.03	**1.0** ±0.0
	Supporting	**0.76** ±0.02	0.64 ±0.22	0.58 ±0.06	0.71 ±0.07	0.28 ±0.1	0.66 ±0.06	0.24 ±0.2
	Irrelevant	0.7 ±0.15	**0.76** ±0.02	0.59 ±0.06	0.69 ±0.05	0.24 ±0.08	0.55 ±0.03	0.51 ±0.15
	Disconnected	0.49 ±0.05	0.45 ±0.05	0.5 ±0.06	0.59 ±0.05	0.24 ±0.08	0.5 ±0.06	**0.8** ±0.17
Supporting	Supporting	0.67 ±0.06	0.66 ±0.07	0.68 ±0.05	0.65 ±0.04	0.32 ±0.09	0.57 ±0.04	**0.98** ±0.01
Irrelevant	Irrelevant	0.51 ±0.06	0.52 ±0.06	0.5 ±0.04	0.56 ±0.04	0.25 ±0.06	0.53 ±0.06	**0.93** ±0.01
Disconnected	Disconnected	0.57 ±0.07	0.57 ±0.06	0.45 ±0.11	0.4 ±0.1	0.17 ±0.05	0.47 ±0.06	**0.96** ±0.01
Average		**0.61** ±0.08	0.59 ±0.08	0.54 ±0.07	**0.61** ±0.06	0.30 ±0.07	0.56 ±0.05	**0.77** ±0.09

19

Q2. Do models reason robustly?

- Supporting fact
- Irrelevant fact
- Disconnected fact

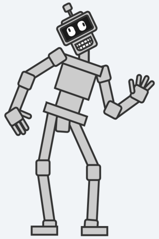


Models		Unstructured models (no graph)						Structured model (with graph)
Training	Testing	BiLSTM - Attention	BiLSTM - Mean	RN	MAC	BERT	BERT-LSTM	GAT
Clean	Clean	0.58 ± 0.05	0.53 ± 0.05	0.49 ± 0.06	0.63 ± 0.08	0.37 ± 0.06	0.67 ± 0.03	**1.0** ± 0.0
	Supporting	**0.76** ± 0.02	0.64 ± 0.22	0.58 ± 0.06	0.71 ± 0.07	0.28 ± 0.1	0.66 ± 0.06	0.24 ± 0.2
	Irrelevant	0.7 ± 0.15	**0.76** ± 0.02	0.59 ± 0.06	0.69 ± 0.05	0.24 ± 0.08	0.55 ± 0.03	0.51 ± 0.15
	Disconnected	0.49 ± 0.05	0.45 ± 0.05	0.5 ± 0.06	0.59 ± 0.05	0.24 ± 0.08	0.5 ± 0.06	**0.8** ± 0.17
Supporting	Supporting	0.67 ± 0.06	0.66 ± 0.07	0.68 ± 0.05	0.65 ± 0.04	0.32 ± 0.09	0.57 ± 0.04	**0.98** ± 0.01
Irrelevant	Irrelevant	0.51 ± 0.06	0.52 ± 0.06	0.5 ± 0.04	0.56 ± 0.04	0.25 ± 0.06	0.53 ± 0.06	**0.93** ± 0.01
Disconnected	Disconnected	0.57 ± 0.07	0.57 ± 0.06	0.45 ± 0.11	0.4 ± 0.1	0.17 ± 0.05	0.47 ± 0.06	**0.96** ± 0.01
Average		**0.61** ± 0.08	0.59 ± 0.08	0.54 ± 0.07	**0.61** ± 0.06	0.30 ± 0.07	0.56 ± 0.05	**0.77** ± 0.09

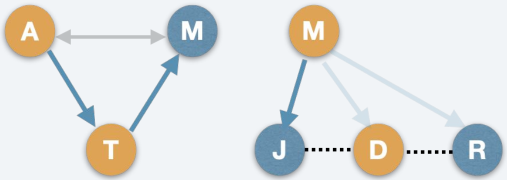
20

Key Takeaways

- **Structure** is required for better generalization and robust reasoning
- **Syntax parsing** could be a bottleneck in understanding structure
- Logic provides a provable way to devise tasks for semantic/syntactic understanding



Puzzle	Question	Gender	Answer
Mario wanted to get a good gift for his sister, *Marianne Jean* and her sister *Darlene* were going to a party held by *Jean*'s mom, *Marianne*. *Darlene* invited her brother *Roy* to come, too, but he was too busy. *Teri* and her father, *Mario*, had an argument over the weekend. However, they made up by Monday. *Agnes* wants to make a special meal for her daughter *Teri*'s birthday.	Roy is the ____ of Agnes	Agnes:female, Teri:female, Mario:male, Marianne:female, Jean:female, Darlene:female, Roy:male	nephew



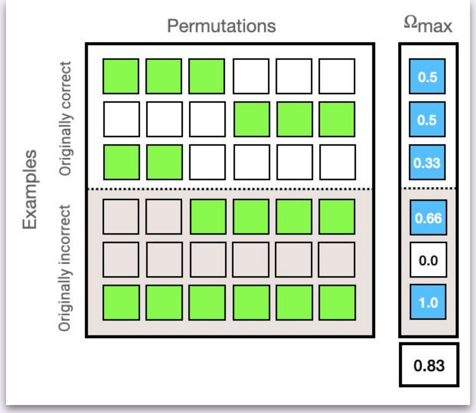
21

UnNatural Language Inference

K Sinha, P Parthasarathi, J Pineau, A Williams

P: Boats in daily use lie within feet of the fashionable bars and restaurants .

H: There are boats close to bars and restaurants .



ACL-IJCNLP 2021

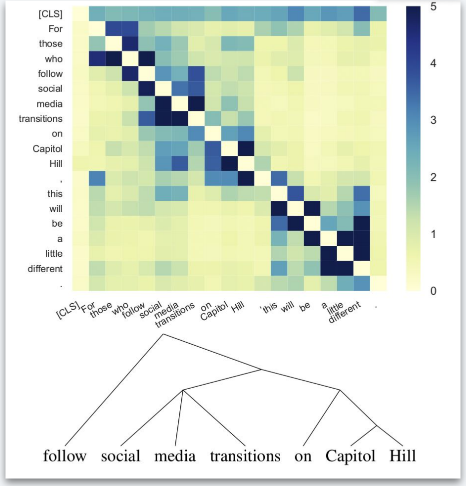
Oral, Outstanding Paper Award

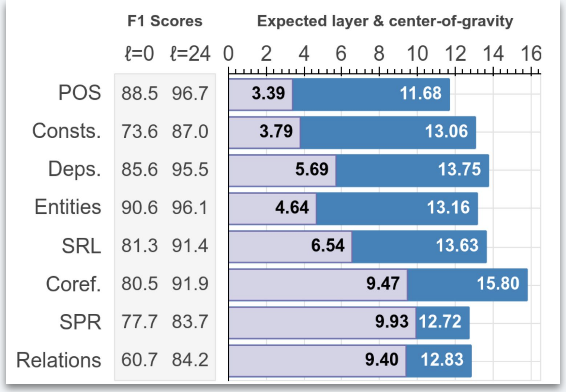
“Pretrained LMs know syntax”

Many papers claim LMs “know syntax” on the basis of probes and diagnostic datasets

- BERT project syntax structure in attention patterns

- BERT ‘*recreates the classical NLP pipeline*’





Goldberg, 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Wu et al., 2020; Tenney et al 2019; Warstadt et al 2019a,b; Warstadt and Bowman 2020; Linzen and Baroni 2021

Test of syntax: **the order of words** conveys important information.

The person* bit *the cat.

The cat* bit *the person.

mean very different things!

Task: Natural Language Inference (NLI)

James Byron Dean refused to move without blue jeans

{entails, contradicts, neither}

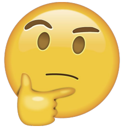
James Dean didn't dance without pants

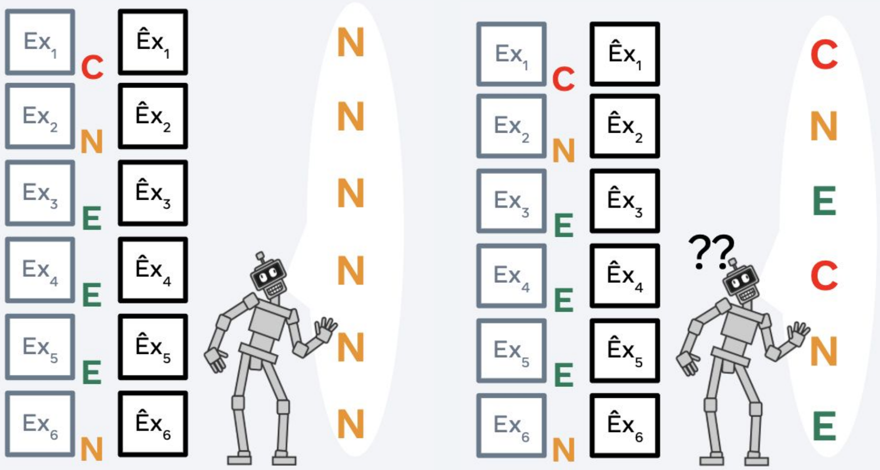
---

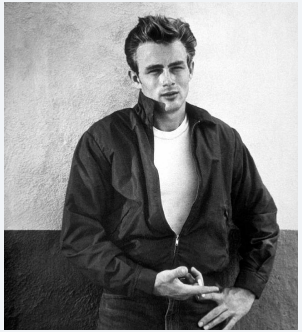
refused James jeans blue without Dean Byron move to

{entails?, contradicts?, neither?}

didn't Dean James pants dance without

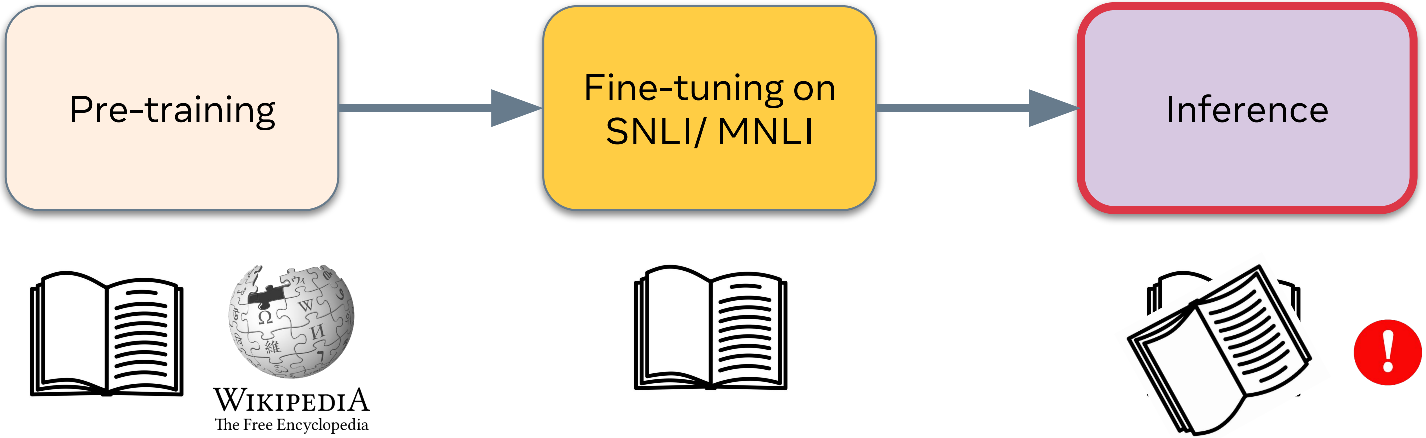


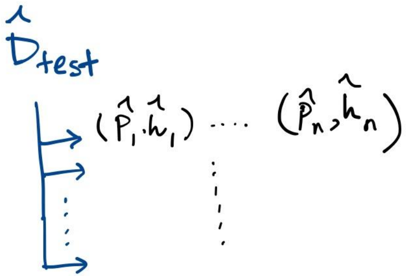




Natural Language Inference, Bill MacCartney PhD Dissertation, 2009; https://nlp.stanford.edu/~wcmac/papers/nli-diss.pdf

Measuring consistency in understanding





- ***No word should appear in its original position***
- A sentence of length *n* has *(n-1)!* possible permutations
- We select only *unique* permutations from this operation

Does word order matter?
Probably Not!

- State-of-the-art NLI **models are largely invariant to word order**!
- **Models often *accept* permuted examples** (i.e. assign the original gold label to them).
- Same for pre-Transformer era neural models, too!

P: Boats in daily use lie within feet of the fashionable bars and restaurants .
H: There are boats close to bars and restaurants .

Concurrently, similar findings on GLUE and QA has been shown by Pham et al 2021, Gupta et al 2021

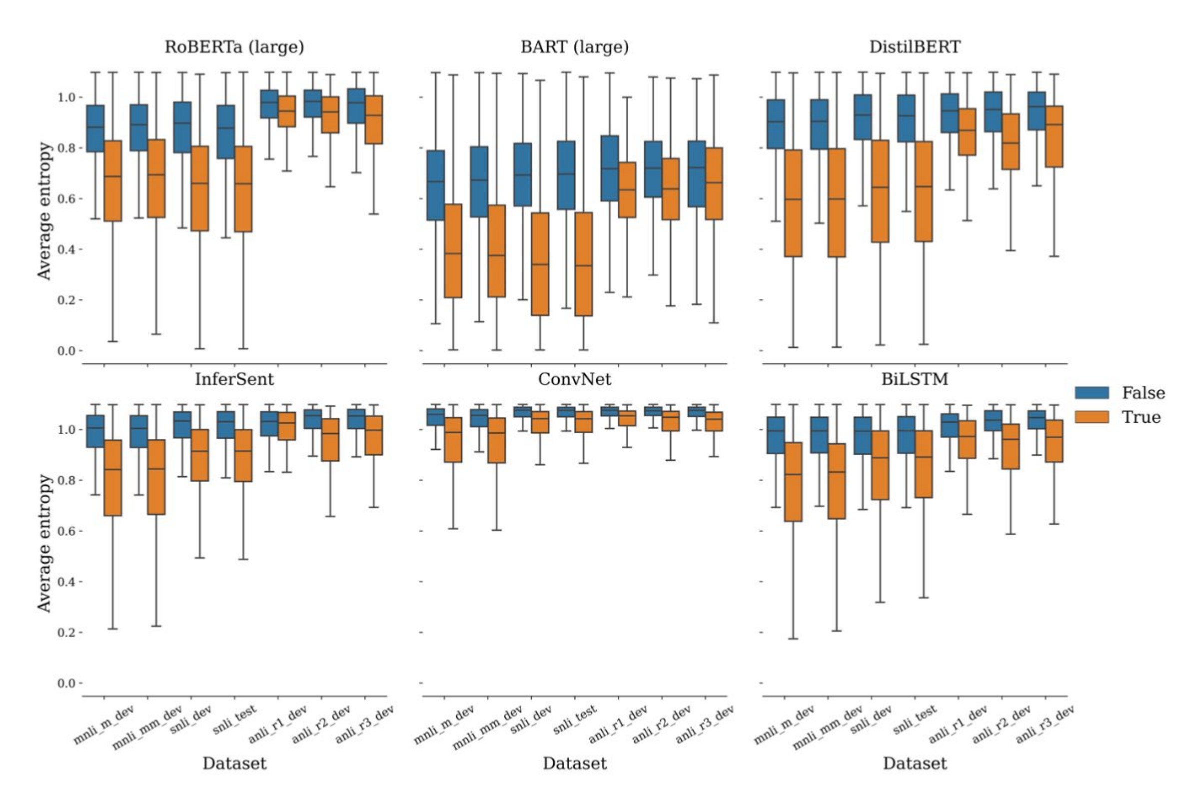
Major findings

Transformer models (RoBERTa, BART, DistilBERT) accepts

- at least one permutation as correct: **98.9%**
- at least 1/3rd (out of 100) permutations as correct for **83.6%**
- all permutations (100/100) correct for **10-20%**
- Humans get only **48%** of permutations correct

35-40% of permutations labeled correct whose original examples were **wrong**!

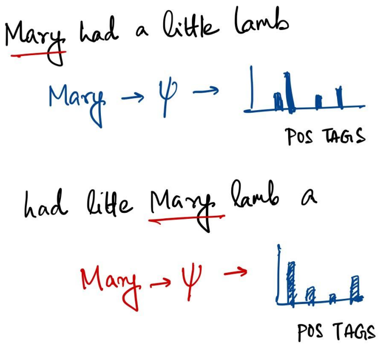
Models display high confidence

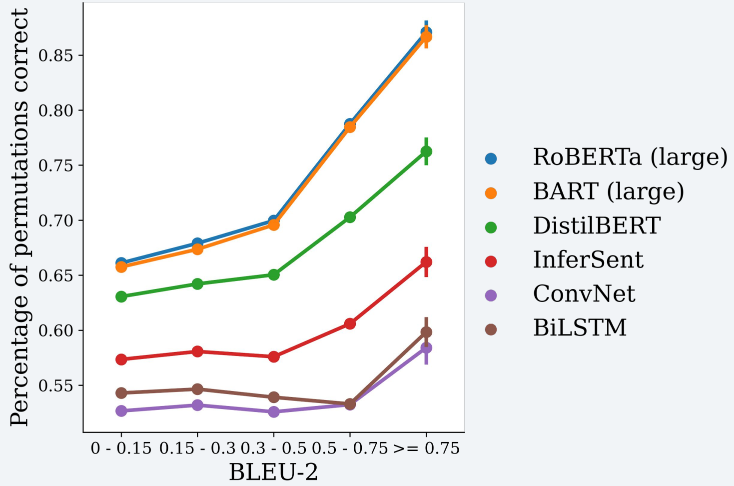


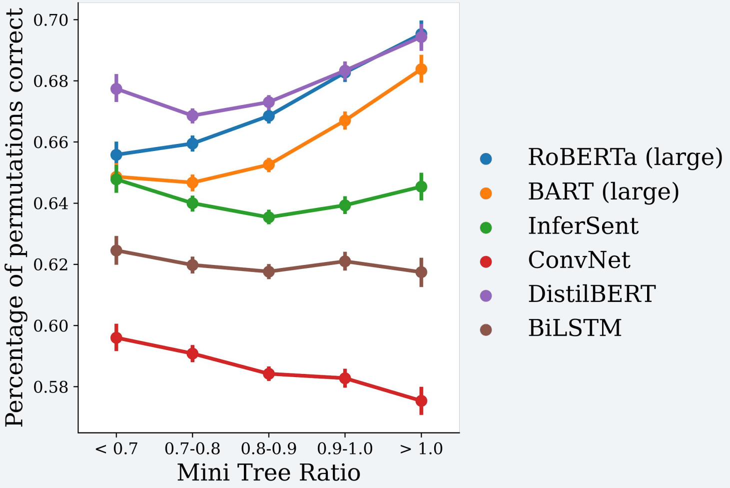
Probable causes of permutation acceptance

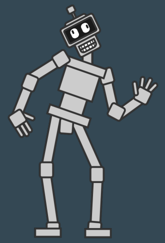
- Preserving local word order leads to accepted permutations

- Transformer LMs aren't entirely BOW, they operate on abstract syntactic information









Takeaways:

1. All tested models are largely insensitive to permutations of word order, though humans are not.

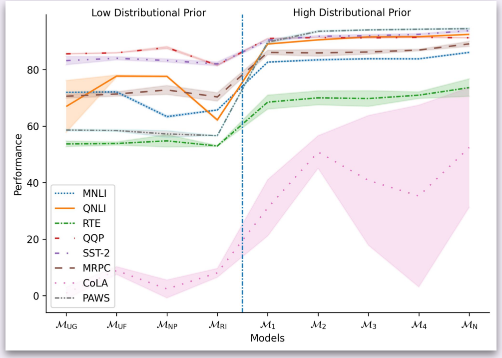
2. Reordering words can cause models to flip classification labels

3. Models have learned some distributional information (*POS neighborhood*) that enable them to perform reasonably well under the permuted set up

Masked Language Modeling and the Distributional Hypothesis

Order Word Matters Pre-training for Little

K Sinha, R Jia, D Hupkes, J Pineau, A Williams, D Kiela

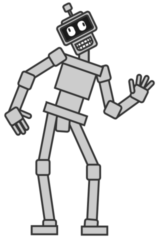


EMNLP 2021

Measuring consistency in syntax representations

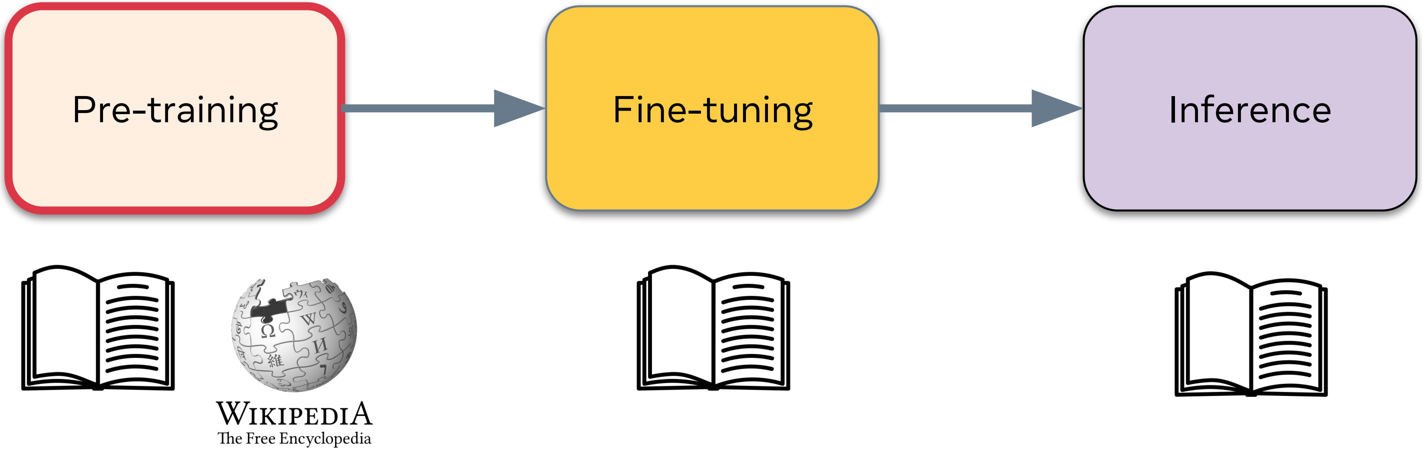
Are Transformer models systematic?

- Should be sensitive to syntactic perturbations
- Should be consistent in learning syntax



Word order as a proxy for syntax

Measuring consistency in syntax representations

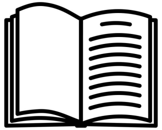


Measuring consistency in syntax representations

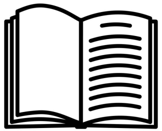
RoBERTa (base) - 125M parameters, 768 hidden size, 12 layers

Pre-training → Fine-tuning → Inference

?



- BookWiki corpus (16GB)
- *"no word should appear in its original position"*
- N-gram shuffles



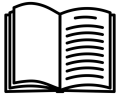
Models and Baselines

Low distributional prior

- No positional embedding
- Random corpus
  - Weighted
  - Uniform
- Random Initialization

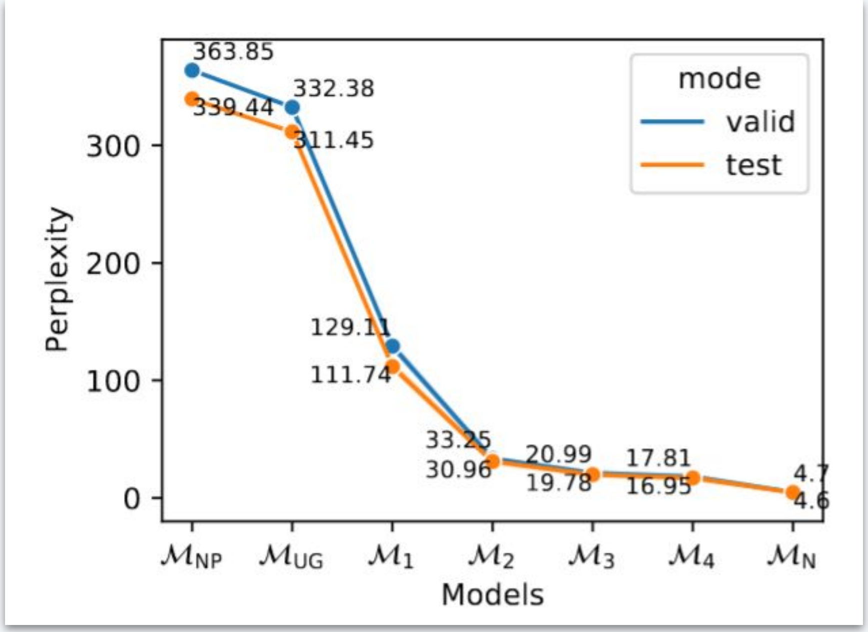
High distributional prior

- Unigram shuffle
- Bigram shuffle
- Trigram shuffle
- Four-gram shuffle



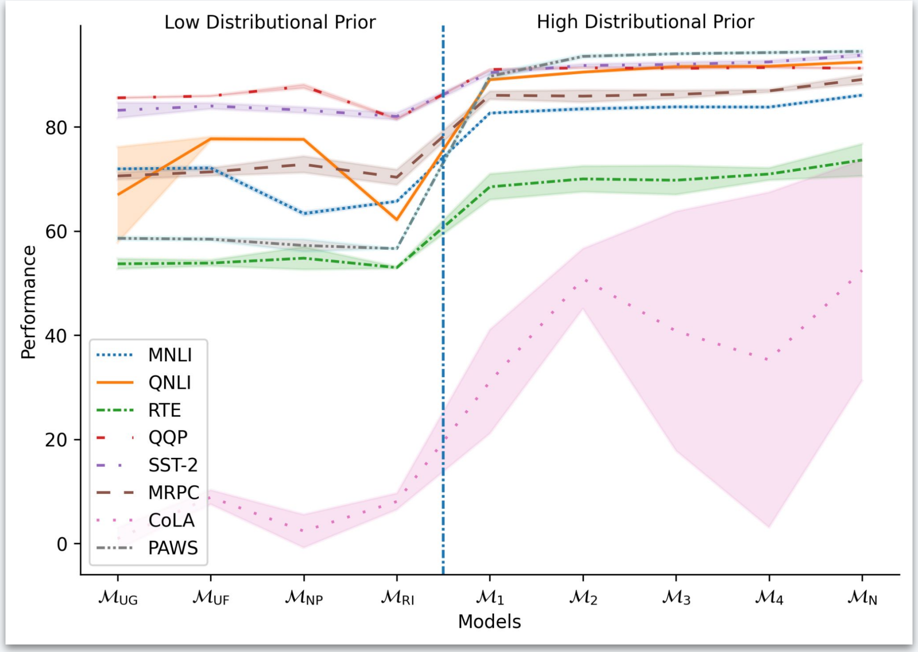
Natural word order model

Roberta (base) trained on BookCorpus + Wikipedia



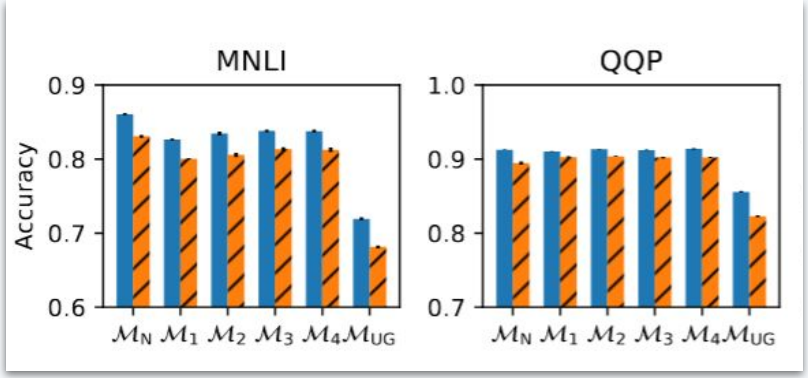
Models pre-trained on shuffled text gets optimal results on downstream tasks!

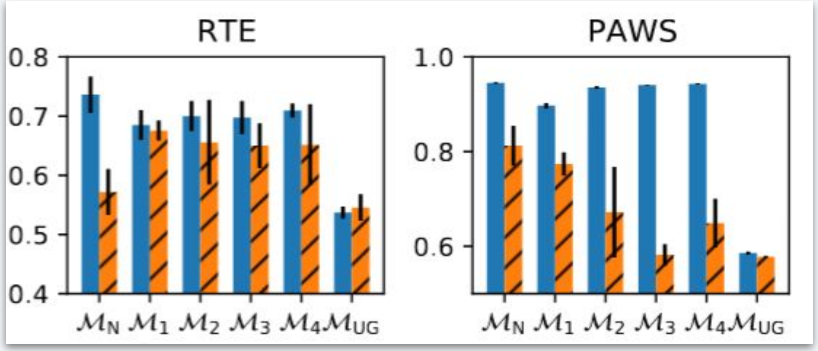
- MNLI (**82%** on n=1 vs **86%** on original)
- QQP **91.01%** vs **91.25%**
- PAWS **89.69%** vs **94.49%**
- CoLA - **31.08** vs **52.48**



Source of word order

- For many tasks, models does equally well when *fine-tuned on shuffled corpus*!
- For word order reliant task, models learn word order *primarily from fine-tuning corpus*





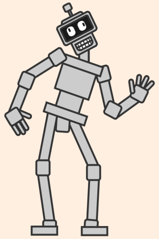
Syntax probes get high accuracy on unnatural models

- *POS tagging*
- *Dependency arc labeling*
- *Dependency parsing*
- *linear and non-linear parametric probes*
- *SentEval task*
- *Subject Verb agreement analysis*

Model	UD EWT		PTB	
	MLP	Linear	MLP	Linear
$\mathcal{M}_{\mathrm{N}}$	80.41 +/- 0.85	66.26 +/- 1.59	86.99 +/- 1.49	66.47 +/- 2.77
$\mathcal{M}_1$	69.26 +/- 6.00	56.24 +/- 5.05	79.43 +/- 0.96	57.20 +/- 2.76
$\mathcal{M}_2$	78.22 +/- 0.88	64.96 +/- 2.32	84.72 +/- 0.55	64.69 +/- 2.50
$\mathcal{M}_3$	77.80 +/- 3.09	64.89 +/- 2.63	85.89 +/- 1.01	66.11 +/- 1.68
$\mathcal{M}_4$	78.04 +/- 2.06	65.61 +/- 1.99	85.62 +/- 1.09	66.49 +/- 2.02
$\mathcal{M}_{\mathrm{UG}}$	74.15 +/- 0.93	65.69 +/- 7.35	80.07 +/- 0.79	57.28 +/- 1.42

Table 2: Unlabeled Attachment Score (UAS) on the dependency parsing task (DEP) on two datasets, UD EWT and PTB, using the Pareto Probing framework (Pimentel et al., 2020a)

Key Takeaways

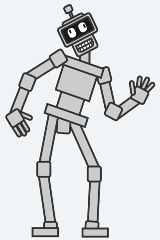


- Word-order doesn’t matter even in pre-training
- Models learn necessary word order from fine-tuning tasks
- Models fail to perform (granular) syntax processing
- Current methods to identify syntax processing are probably not valid
- Distributional statistics is enough
    - Models tend to exploit distributional word co-occurrences to get high scores on downstream tasks

Thesis overview: measuring NLU progress through *systematicity*

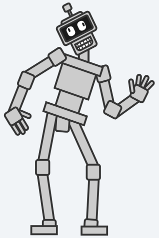
- ⚠ ~~We can~~ Models cannot reason consistently once ~~we~~ they learn the rules
- ⚠ ~~We fail~~ Models does not consistently fail on inputs which doesn't agree with ~~our~~ their learned rules

"It is not enough that models should succeed where humans succeed, they should also fail where humans fail."



Thank you for listening!

Time for your questions!



For a full list of my contributions, check out my website: https://koustuvsinha.com/publication/

@koustuvsinha

koustuv.sinha@mail.mcgill.ca

Thanks to my supervisor and all my collaborators for supporting me throughout my PhD

Mila

Meta AI

McGill School of Computer Science

Extra Slides

Follow Up work: Curious Case of Absolute Position Embeddings

Zero starting position

Who could Thomas observe without distracting Nathan ?

Who	could	Thomas	observe	without	distracting	Nathan	?
0	1	2	3	4	5	6	7

Non-zero starting position

Who could Thomas observe without distracting Nathan ?

Who	could	Thomas	observe	without	distracting	Nathan	?
100	101	102	103	104	105	106	107

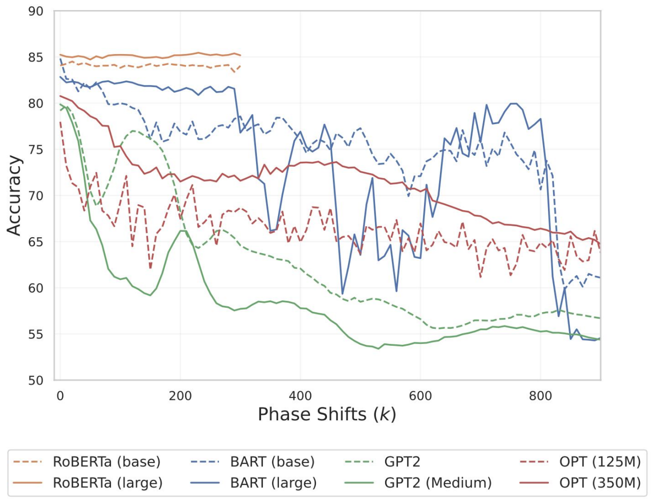


Figure 2: Acceptability Scores in BLiMP (Warstadt et al., 2020) dataset across different phase shifts. RoBERTa only supports context window of size $T = 512$, so we capped the scores to phase shift $k = 300$ to allow for sentences of maximum length in BLiMP to be evaluated.

Thesis overview: measuring NLU progress through *systematicity*

The ability to produce / understand some sentences is intrinsically connected to the ability to produce / understand certain others

Fodor & Pylyshyn, 1988

A human-like, systematic learner must exhibit the following properties:

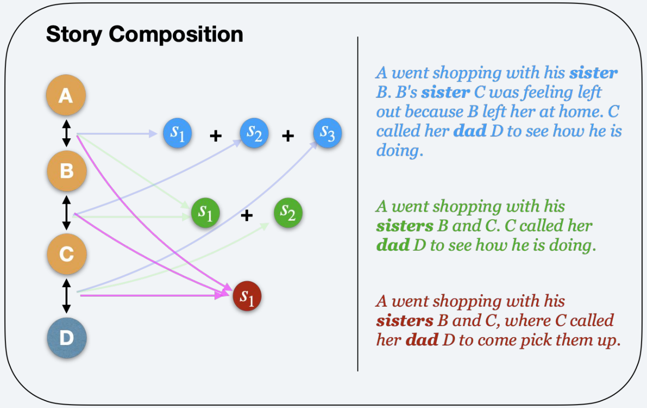
- ✅ Understand the re-combination of known parts and rules
- ✅ Be consistent in understanding in different contexts

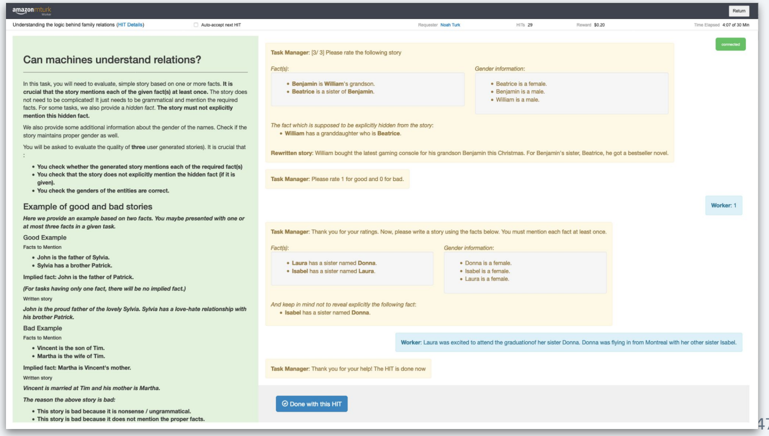
CLUTRR

Extra Slides

Make the data “naturalistic”

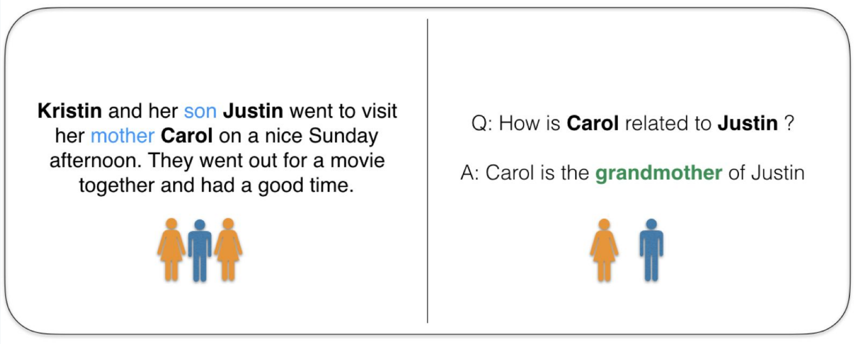
- Collect short stories from Amazon Mechanical Turk
- Build templates based on these short stories
- Apply the templates on the generated graphs





CLUTRR: Compositional Language Understanding with Text-based Relational Reasoning

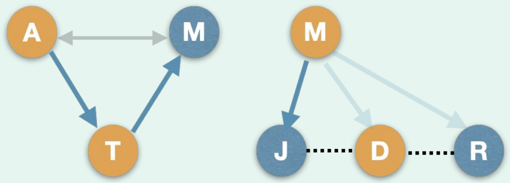
- QA Task of deducing family relations from text
- Inductive reasoning - answer not present explicitly in the text
- Each example has a provable, underlying first-order Horn Clause
- Systematic learner has to learn kinship logical rules and apply to arbitrary stories



48

Dataset snapshot

Puzzle	Question	Gender	Answer
Mario wanted to get a good gift for his sister, *Marianne Jean* and her sister *Darlene* were going to a party held by *Jean*'s mom, *Marianne*. *Darlene* invited her brother *Roy* to come, too, but he was too busy. *Teri* and her father, *Mario*, had an argument over the weekend. However, they made up by Monday. *Agnes* wants to make a special meal for her daughter *Teri*'s birthday.	Roy is the _____ of Agnes	Agnes:female, Teri:female, Mario:male, Marianne:female, Jean:female, Darlene:female, Roy:male	nephew

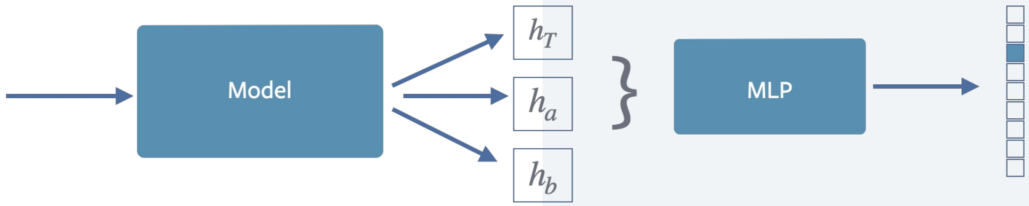


How should the models do?

- Entity extraction and linking
- Coreference resolution
- Rule induction
- Length Generalization

- Models having access to graph underlying the text (Graph Attention Networks)

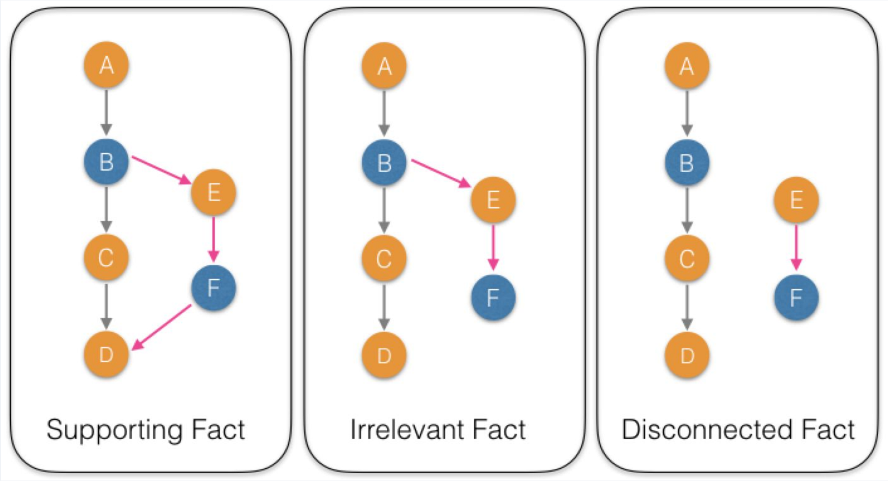
- Models having access to raw text (BiLSTM, Relation Network, MAC, BERT, BERT-LSTM)



Q2. Do models reason robustly?

Alongside consistency, test for robustness

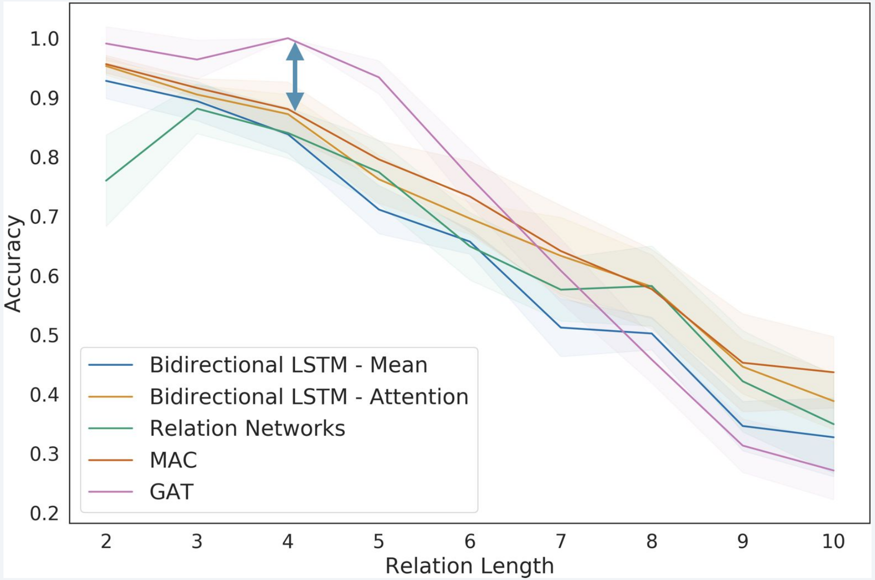
- **Supporting fact:** closed cycle
- **Irrelevant fact**: dangling loops
- **Disconnected fact**: disconnected graph



51

Q1. Are models able to generalize systematically?

- Train on stories less combinations and test on longer combinations
- Ensure model sees all logical kinship rules during training, but not all combinations of those rules
- Split the AMT templates into train and test

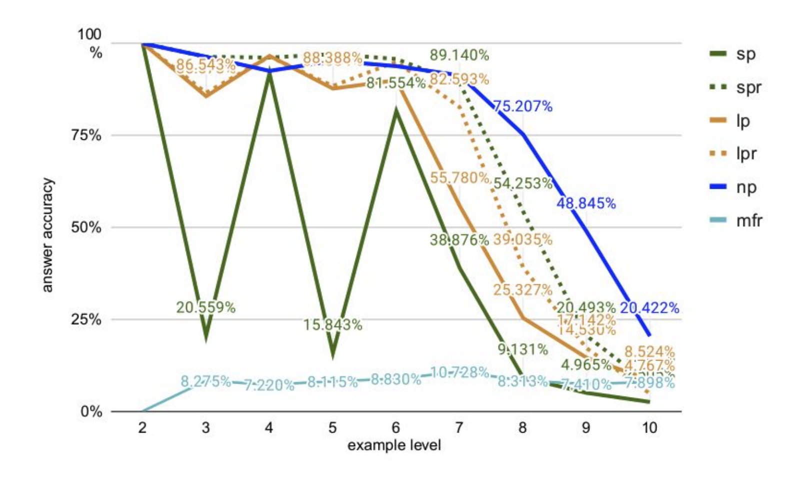


Is syntax understanding the issue of systematic generalization?

How systematic the NLU models are at understanding syntax?

Follow Up Works

- Length Generalization: Interpolation vs Extrapolation
- Models are worse in both scenarios!

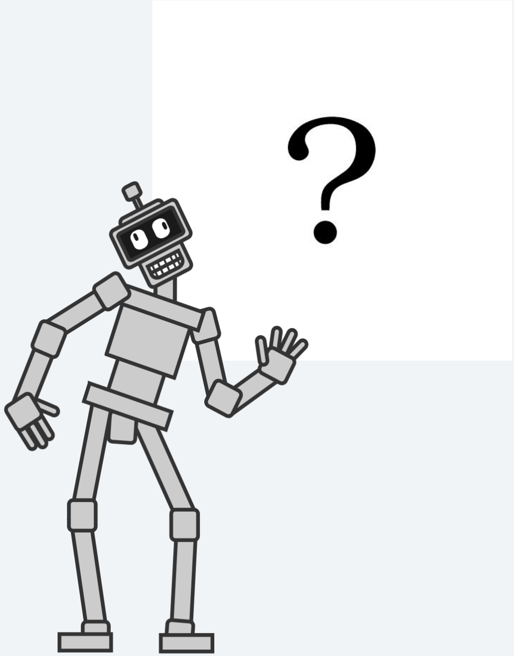


Nicolas Gontier, Koustuv Sinha, Siva Reddy, Chris Pal; *Measuring Systematic Generalization in Neural Proof Generation with Transformers*; NeurIPS 2020

54

Open Questions

- Is **probing** a valid way to extract latent information?
- Do **NLU tasks** require syntax understanding?
  - Or is distributional information is enough?
- Is **distributional overlap** a limiting factor for generalization?
  - Larger datasets, more n-gram statistics in test overlap? [1]



[1] Emami A, Trischler A, Suleman K, Cheung JC. An analysis of dataset overlap on winograd-style tasks. arXiv preprint arXiv:2011.04767. 2020 Nov 9.

UnNatural Language Inference

Probing NLU models using the notion of *systematicity*

The ability to produce / understand some sentences is intrinsically connected to the ability to produce / understand certain others

Fodor & Pylyshyn, 1988

A systematic learner must exhibit the following properties:

- Understand the re-combination of known parts and rules
- **Be consistent in understanding in different contexts**

57

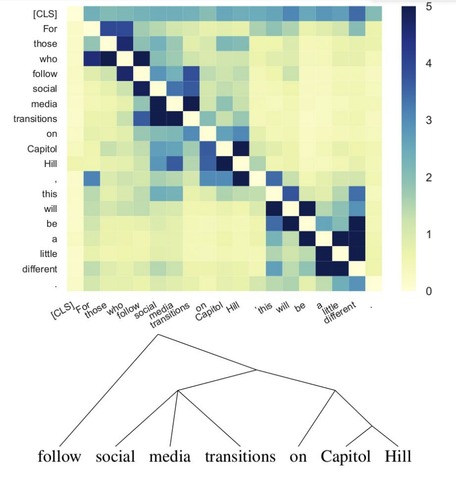
“Pretrained LMs know syntax”

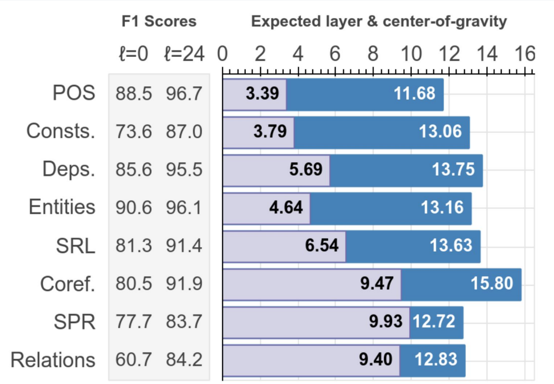
- Wu et al. (2020) recover syntactic trees from BERT considering attention patterns

- Tenney et al. (2019) conclude that BERT ‘recreates the classical NLP pipeline:’ POS tagging, parsing, NER, semantic roles, coreference…

- Many papers claim LMs “know syntax” on the basis of probes and diagnostic datasets

  (Goldberg, 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Wu et al., 2020; Warstadt et al 2019a,b; Warstadt and Bowman 2020; Linzen and Baroni 2021…)





58

If models are genuinely learning syntax, they should know something about word order…

If models are genuinely learning syntax, they should know something about word order… **do they?**

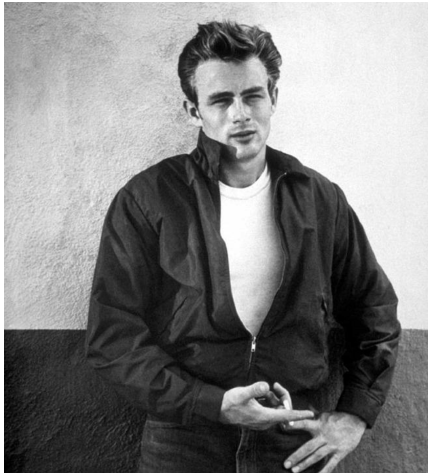
Natural Language Inference (NLI)

also known as recognizing textual entailment (RTE[1])

James Byron Dean refused to move without blue jeans

{entails, contradicts, neither}

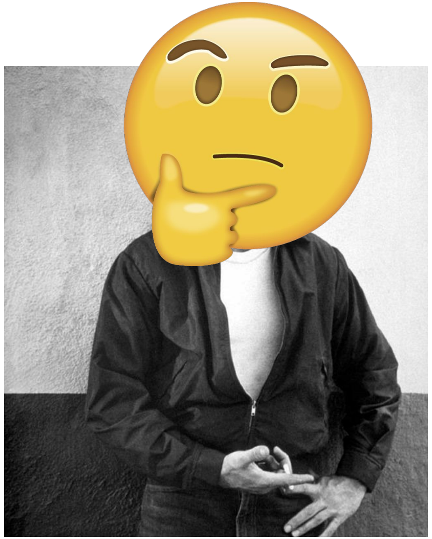
James Dean didn't dance without pants



[1]Fyodorov et al., 2000; Condoravdi et al., 2003; Bos and Markert, 2005; Dagan et al., 2006; MacCartney and Manning, 2009

Example: MacCartney thesis '09

Wait a sec…how *should* a (humanlike) NLI model that's sensitive to word order behave?



refused James jeans blue without Dean Byron move to

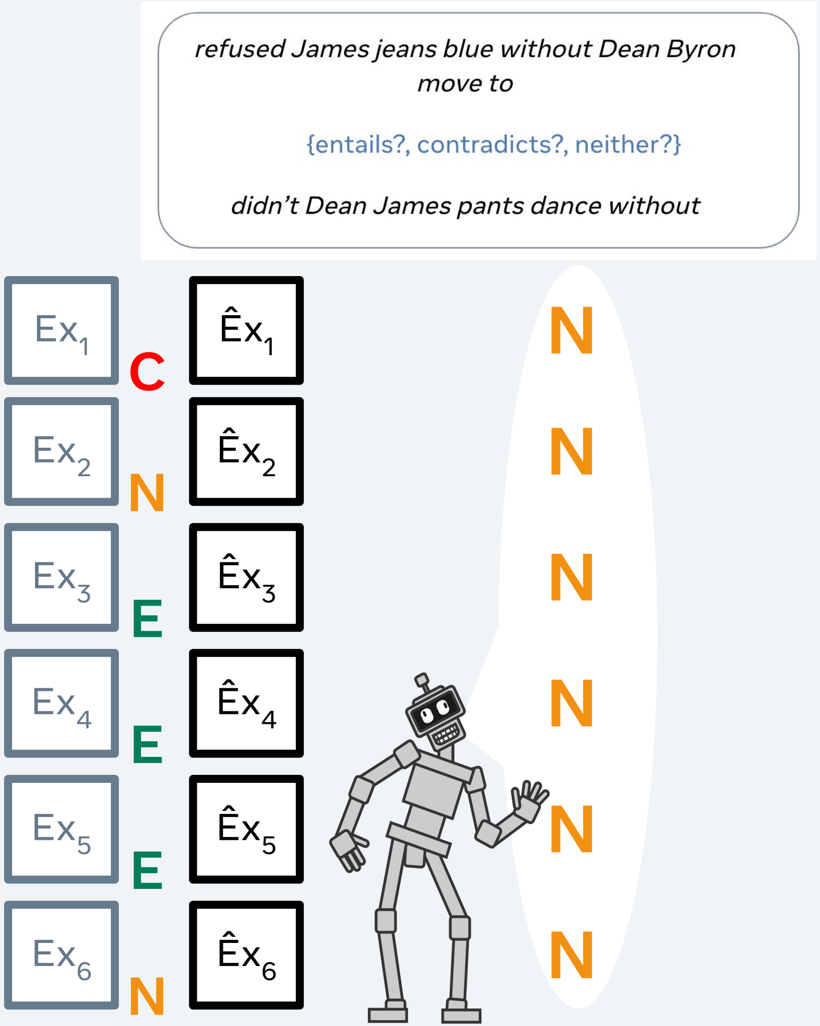
{entails?, contradicts?, neither?}

didn't Dean James pants dance without

(1) Maybe it just performs NLI…

For 3-way NLI, any pair that isn't clearly contradiction or entailment should be **neutral**.

A model that learned this might just assign **neutral** always.

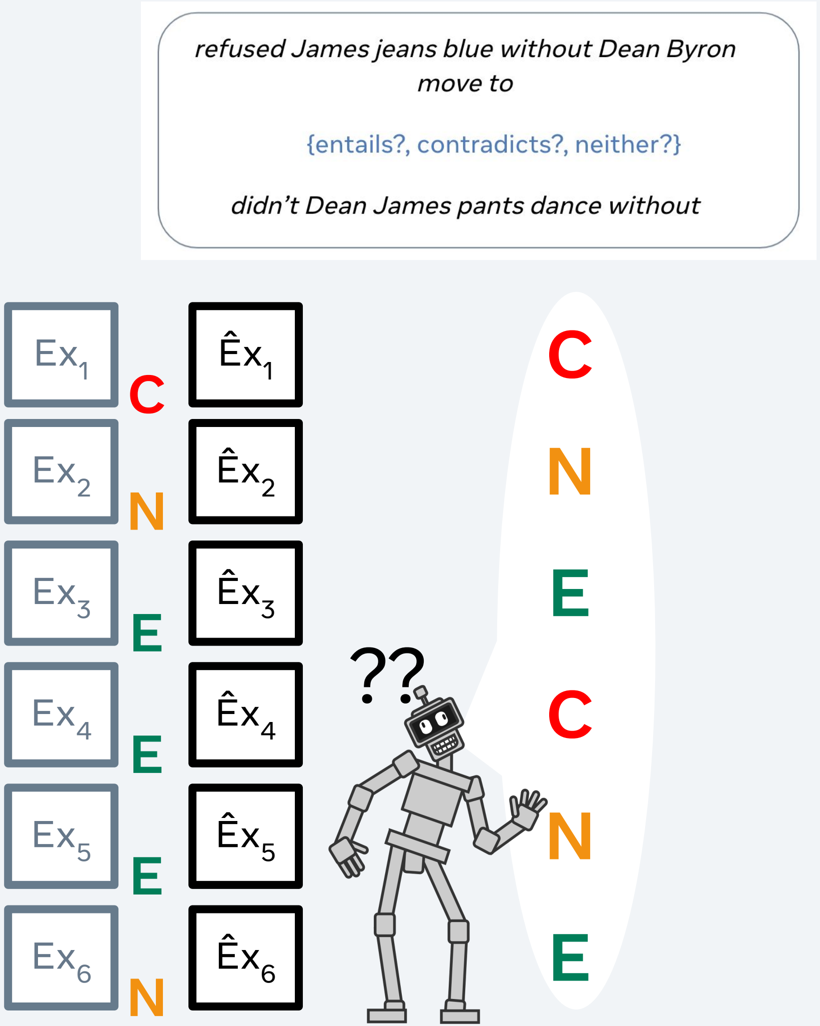


63

(2) Maybe it will just be very uncertain…

Perhaps it will just have no idea…then it should get roughly equal probability mass on all predictions.

This is approximately the **most frequent class** baseline.



64

Spoiler! **It's neither!**

State-of-the-art NLI **models are largely** ***invariant*** **to word order!**

Models often *accept* permuted examples (i.e. assign the original gold label to them).

Same for pre-Transformer era neural models, too!

P: Boats in daily use lie within feet of the fashionable bars and restaurants .
H: There are boats close to bars and restaurants .

Gold Label	Premise	Hypothesis
E	Boats in daily use lie within feet of the fashionable bars and restaurants.	There are boats close to bars and restaurants.
E	restaurants and use feet of fashionable lie the in Boats within bars daily .	bars restaurants are There and to close boats .
C	He and his associates weren't operating at the level of metaphor.	He and his associates were operating at the level of the metaphor.
C	his at and metaphor the of were He operating associates n't level .	his the and metaphor level the were He at associates operating of .

Concurrently, similar findings on GLUE and QA has been shown by Pham et al 2021, Gupta et al 2021

65

Constructing permutation function

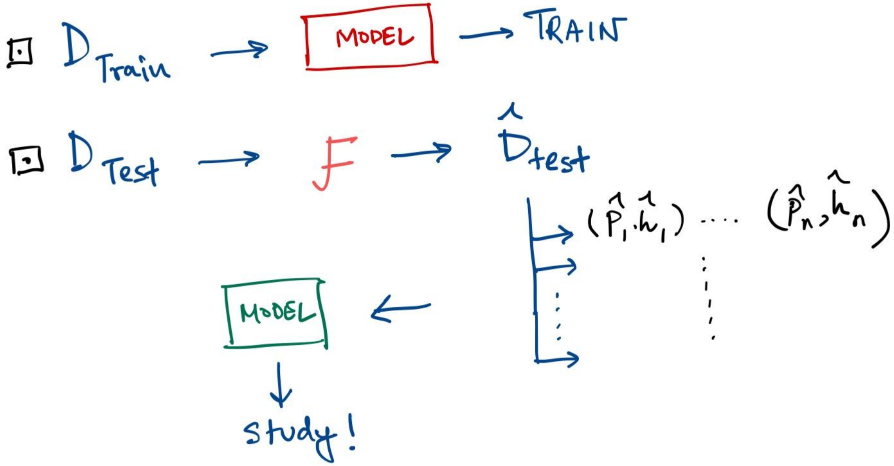
No word should appear in its original position

A sentence of length *n* has *(n-1)!* possible permutations

We select only *unique* permutations from this operation

P: Boats in daily use lie within feet of the fashionable bars and restaurants .

H: There are boats close to bars and restaurants .



66

Experimental Setup:

Trained models (RoBERTa, BART, DistilBERT, InferSent, ConvNet, BiLSTM) on MNLI to SOTA levels.

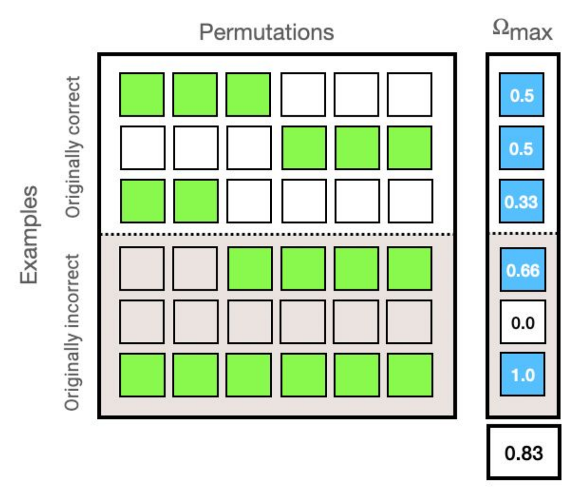
Fine-tuned on (normal) MNLI.

Evaluated on permuted MNLI, SNLI (in domain), ANLI (out of domain).

How many examples have at least one permutation predicting the gold label?

Model	Eval Dataset	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
RoBERTa (large)	MNLI_m_dev	0.906	0.987			
	MNLI_mm_dev	0.901	0.987			
	SNLI_dev	0.879	0.988			
	SNLI_test	0.883	0.988			
	A1_dev	0.456	0.897			
	A2_dev	0.271	0.889			
	A3_dev	0.268	0.902			
	Mean	0.652	0.948			
	Harmonic Mean	0.497	0.946			
BART (large)	MNLI_m_dev	0.902	0.989			
	MNLI_mm_dev	0.900	0.986			
	SNLI_dev	0.886	0.991			
	SNLI_test	0.888	0.990			
	A1_dev	0.455	0.894			
	A2_dev	0.316	0.887			
	A3_dev	0.327	0.931			
	Mean	**0.668**	**0.953**			
	Harmonic Mean	**0.543**	**0.951**			
DistilBERT	MNLI_m_dev	0.800	0.968			
	MNLI_mm_dev	0.811	0.968			
	SNLI_dev	0.732	0.956			
	SNLI_test	0.738	0.950			
	A1_dev	0.251	0.750			
	A2_dev	0.300	0.760			
	A3_dev	0.312	0.830			
	Mean	0.564	0.883			
	Harmonic Mean	0.445	0.873			

98.9%



E_1: 3 gold label assignments **(50%)**
E_2: 3 gold label assignments **(50%)**
E_3: 2 gold label assignments **(33%)**
E_4: 4 gold label assignments **(66%)**
E_5: 0 gold label assignments ~~(00%)~~
E_6: 6 gold label assignments **(100%)**

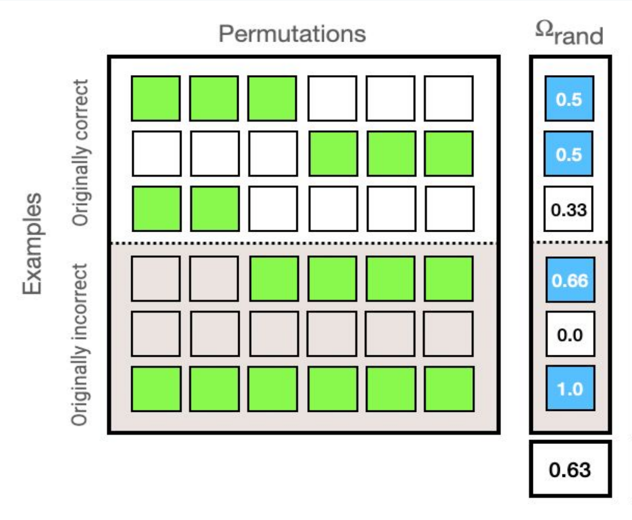
Ω_{max} = ⅚ examples = 83%

68

How many examples have at least 1/3rd permutations predicting the gold label?

Model	Eval Dataset	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
RoBERTa (large)	MNLI_m_dev	0.906	0.987			0.794
	MNLI_mm_dev	0.901	0.987			0.790
	SNLI_dev	0.879	0.988			0.826
	SNLI_test	0.883	0.988			0.828
	A1_dev	0.456	0.897			0.364
	A2_dev	0.271	0.889			0.359
	A3_dev	0.268	0.902			0.397
	Mean	0.652	0.948			0.623
	Harmonic Mean	0.497	0.946			0.539
BART (large)	MNLI_m_dev	0.902	0.989			0.784
	MNLI_mm_dev	0.900	0.986			0.788
	SNLI_dev	0.886	0.991			0.834
	SNLI_test	0.888	0.990			0.836
	A1_dev	0.455	0.894			0.374
	A2_dev	0.316	0.887			0.397
	A3_dev	0.327	0.931			0.424
	Mean	**0.668**	**0.953**			**0.634**
	Harmonic Mean	**0.543**	**0.951**			**0.561**
DistilBERT	MNLI_m_dev	0.800	0.968			0.779
	MNLI_mm_dev	0.811	0.968			0.786
	SNLI_dev	0.732	0.956			0.731
	SNLI_test	0.738	0.950			0.725
	A1_dev	0.251	0.750			0.300
	A2_dev	0.300	0.760			0.343
	A3_dev	0.312	0.830			0.363
	Mean	0.564	0.883			0.575
	Harmonic Mean	0.445	0.873			0.490

83.6%



E_1: 3 gold label assignments (50%)
E_2: 3 gold label assignments (50%)
E_3: 2 gold label assignments ~~(33%)~~
E_4: 4 gold label assignments (66%)
E_5: 0 gold label assignments ~~(00%)~~
E_6: 6 gold label assignments (100%)

Ω_{rand} = ⅔ examples = 63%

69

How many examples have ALL permutations predicting the gold label?

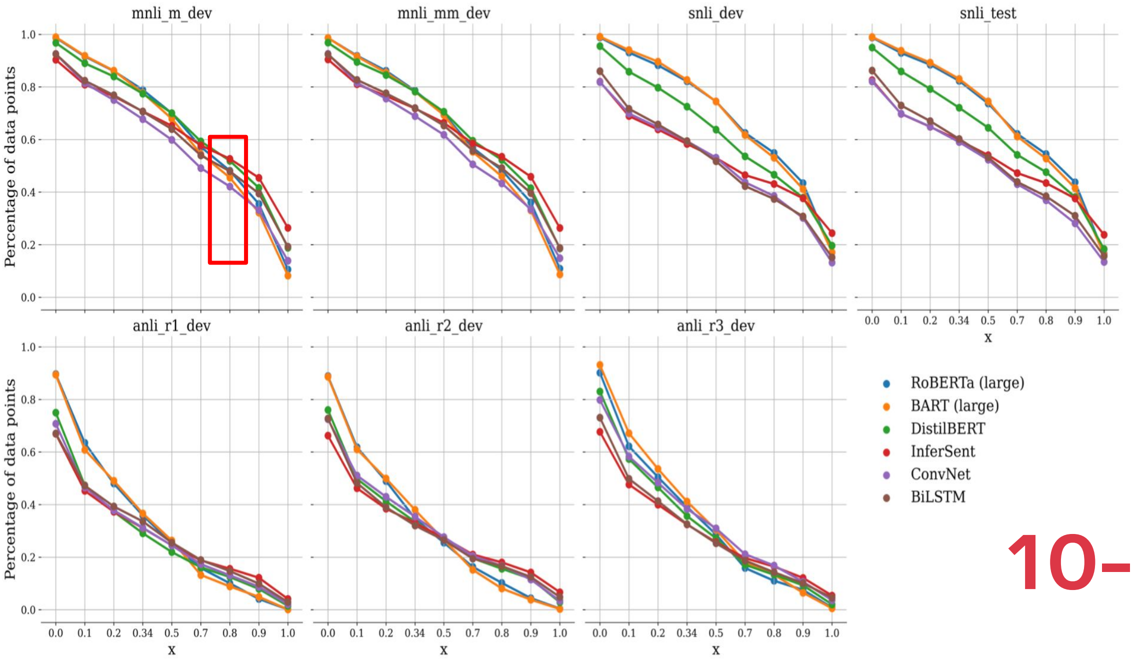
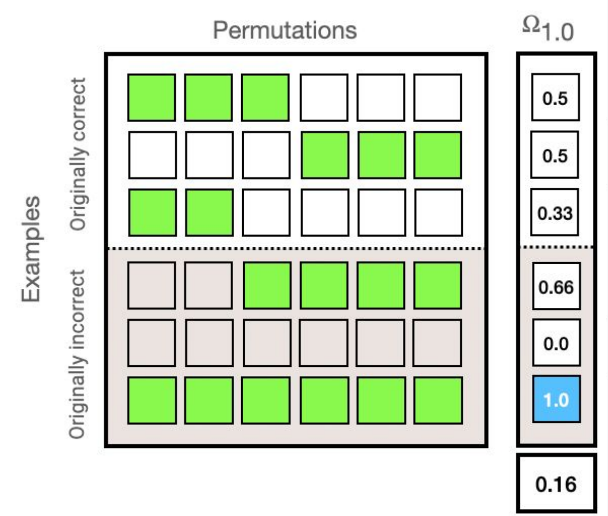


Figure 7: Ω_x threshold for all datasets with varying x and computing the percentage of examples that fall within the threshold. The top row consists of in-distribution datasets (MNLI, SNLI) and the bottom row contains out-of-distribution datasets (ANLI)



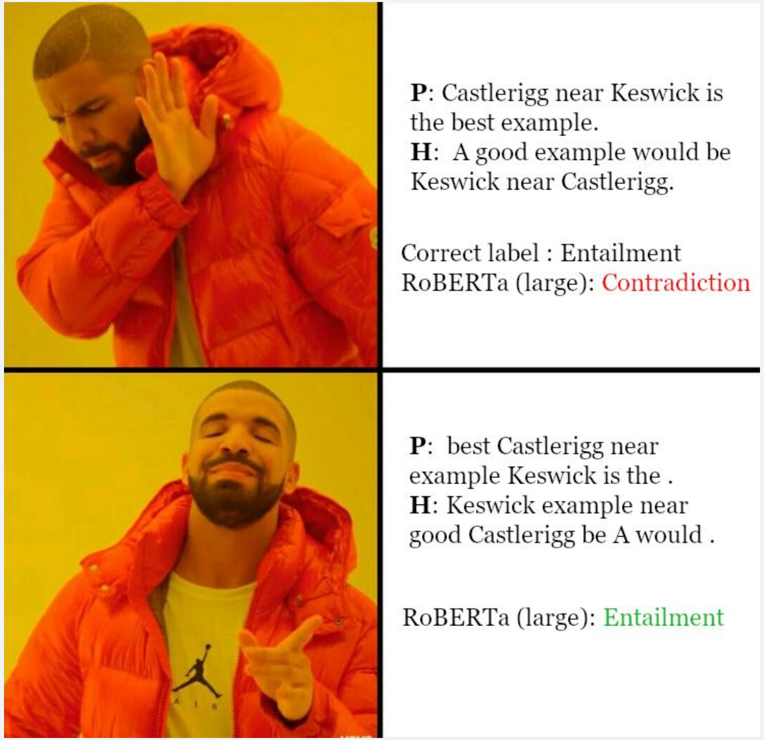
10-20%

E_1: 3 gold label assignments (50%)
E_2: 3 gold label assignments (50%)
E_3: 2 gold label assignments ~~(33%)~~
E_4: 4 gold label assignments (66%)
E_5: 0 gold label assignments ~~(00%)~~
E_6: 6 gold label assignments (100%)

$\Omega_{1.0}$ = ⅙ examples = 16%

70

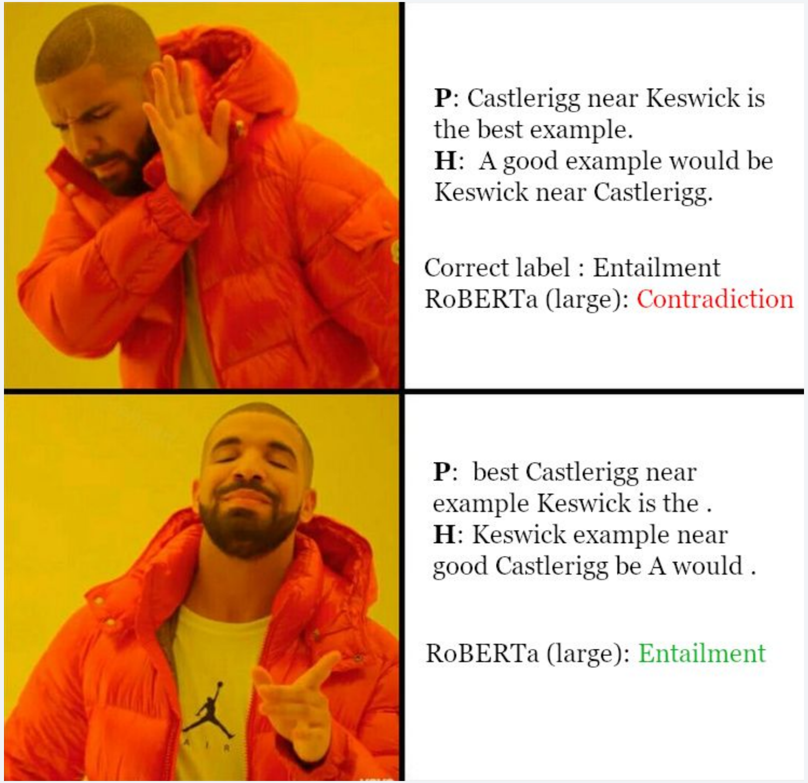
We observed that **for some examples the models initially got wrong**, there exists (a) permutation(s) that receive(s) the **gold label**!



71

What do we find?

- We also find, **for examples the models initially got wrong**, there exists a word-ordering that can make it correct!



UnNatural Language Inference

Koustuv Sinha[1,2,3], Prasanna Parthasarathi[1,2], Joelle Pineau[1,2,3] and Adina Williams[3]
[1] School of Computer Science, McGill University, Canada
[2] Montreal Institute of Learning Algorithms (Mila), Canada
[3] Facebook AI Research (FAIR)
{koustuv.sinha, prasanna.parthasarathi, jpineau, adinawilliams}
@{mail.mcgill.ca, mail.mcgill.ca, cs.mcgill.ca, fb.com}

72

FLIPS: *What percentage of permutations predict gold label, whose original pairs were INCORRECTLY predicted?*

Model	Eval Dataset	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
RoBERTa (large)	MNLI_m_dev	0.906	0.987	0.707	0.383	0.794
	MNLI_mm_dev	0.901	0.987	0.707	0.387	0.790
	SNLI_dev	0.879	0.988	0.768	0.393	0.826
	SNLI_test	0.883	0.988	0.760	0.407	0.828
	A1_dev	0.456	0.897	0.392	0.286	0.364
	A2_dev	0.271	0.889	0.465	0.292	0.359
	A3_dev	0.268	0.902	0.480	0.308	0.397
	Mean	0.652	0.948	0.611	0.351	0.623
	Harmonic Mean	0.497	0.946	0.572	0.344	0.539
BART (large)	MNLI_m_dev	0.902	0.989	0.689	0.393	0.784
	MNLI_mm_dev	0.900	0.986	0.695	0.399	0.788
	SNLI_dev	0.886	0.991	0.762	0.363	0.834
	SNLI_test	0.888	0.990	0.762	0.370	0.836
	A1_dev	0.455	0.894	0.379	0.295	0.374
	A2_dev	0.316	0.887	0.428	0.303	0.397
	A3_dev	0.327	0.931	0.428	0.333	0.424
	Mean	0.668	0.953	0.592	**0.351**	0.634
	Harmonic Mean	0.543	0.951	0.546	**0.347**	0.561
DistilBERT	MNLI_m_dev	0.800	0.968	0.775	0.343	0.779
	MNLI_mm_dev	0.811	0.968	0.775	0.346	0.786
	SNLI_dev	0.732	0.956	0.767	0.307	0.731
	SNLI_test	0.738	0.950	0.770	0.312	0.725
	A1_dev	0.251	0.750	0.511	0.267	0.300
	A2_dev	0.300	0.760	0.619	0.265	0.343
	A3_dev	0.312	0.830	0.559	0.259	0.363
	Mean	0.564	0.883	**0.682**	0.300	0.575
	Harmonic Mean	0.445	0.873	**0.664**	0.296	0.490

35-40%

Note: for a classic Bag-of-Words, P^c would be 100% and P^f would be 0%!

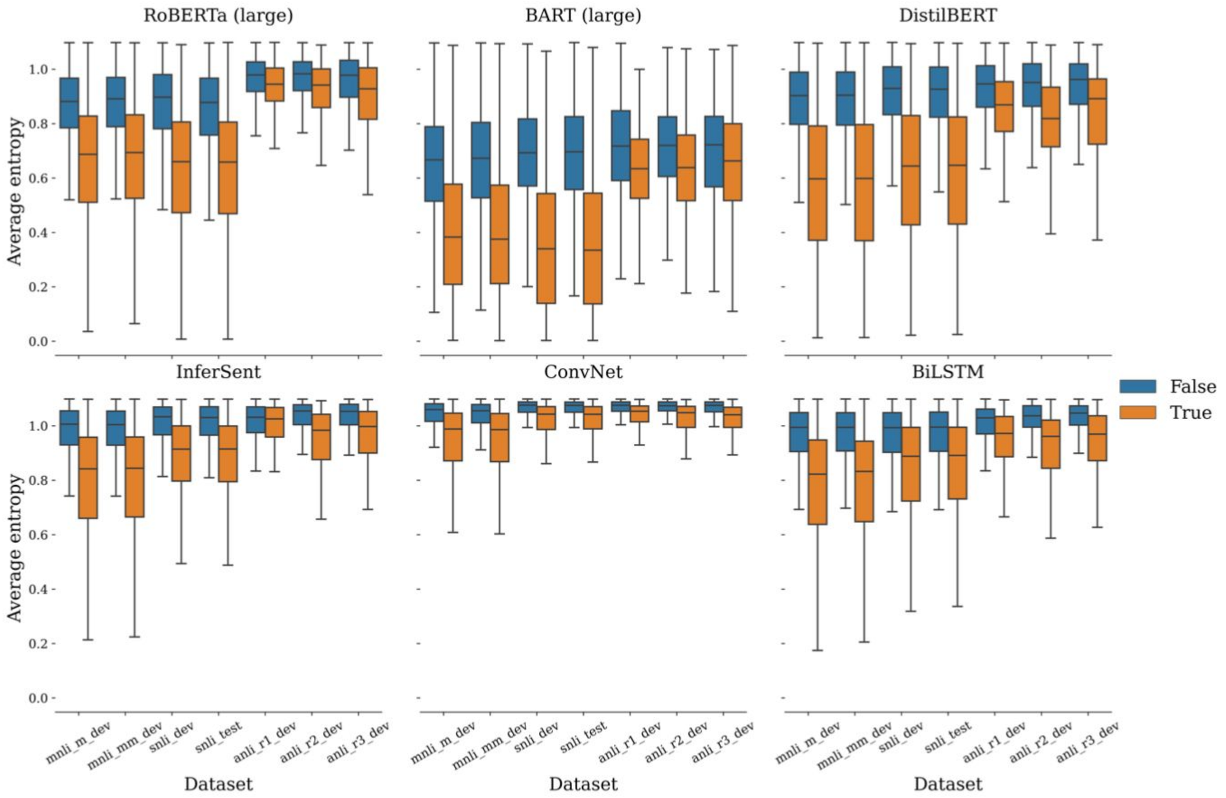
73

Is it just for Transformers? No!

- *Weaker models, weaker effect.*
- P^f for non-Transformers is approximately the same as for transformers.
- Both architectures are similarly bag-of-words-y (though no investigated model is a strict BOW).

Model	Eval Dataset	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
InferSent	MNLI_m_dev	0.658	0.904	0.842	0.359	0.712
	MNLI_mm_dev	0.669	0.905	0.844	0.368	0.723
	SNLI_dev	0.556	0.820	0.821	0.323	0.587
	SNLI_test	0.560	0.826	0.824	0.321	0.600
	A1_dev	0.316	0.669	0.425	0.395	0.313
	A2_dev	0.310	0.662	0.689	0.249	0.330
	A3_dev	0.300	0.677	0.675	0.236	0.332
	Mean	**0.481**	0.780	0.731	**0.322**	0.514
	Harmonic Mean	0.429	0.767	0.694	**0.311**	0.455
ConvNet	MNLI_m_dev	0.631	0.926	0.773	0.340	0.684
	MNLI_mm_dev	0.640	0.926	0.782	0.343	0.694
	SNLI_dev	0.506	0.819	0.813	0.339	0.597
	SNLI_test	0.501	0.821	0.809	0.341	0.596
	A1_dev	0.271	0.708	0.648	0.218	0.316
	A2_dev	0.307	0.725	0.703	0.224	0.356
	A3_dev	0.306	0.798	0.688	0.234	0.388
	Mean	0.452	**0.817**	**0.745**	0.291	0.519
	Harmonic Mean	0.404	**0.810**	**0.740**	0.279	**0.473**
BiLSTM	MNLI_m_dev	0.662	0.925	0.800	0.351	0.711
	MNLI_mm_dev	0.681	0.924	0.809	0.344	0.724
	SNLI_dev	0.547	0.860	0.762	0.351	0.598
	SNLI_test	0.552	0.862	0.771	0.363	0.607
	A1_dev	0.262	0.671	0.648	0.271	0.340
	A2_dev	0.297	0.728	0.672	0.209	0.328
	A3_dev	0.304	0.731	0.656	0.219	0.331
	Mean	0.472	0.814	0.731	0.301	**0.520**
	Harmonic Mean	0.410	0.803	0.725	0.287	0.463

Wait a minute! The labels must be chosen by chance!



- Unfortunately, no. The average entropy for Transformers is pretty low, suggesting overconfidence*!
- BART has the lowest entropy/highest confidence!
- Pre-Transformer models are somewhat better, but probably due to their lower capacity.

Recall: highest entropy for 3-labels is ~1.58

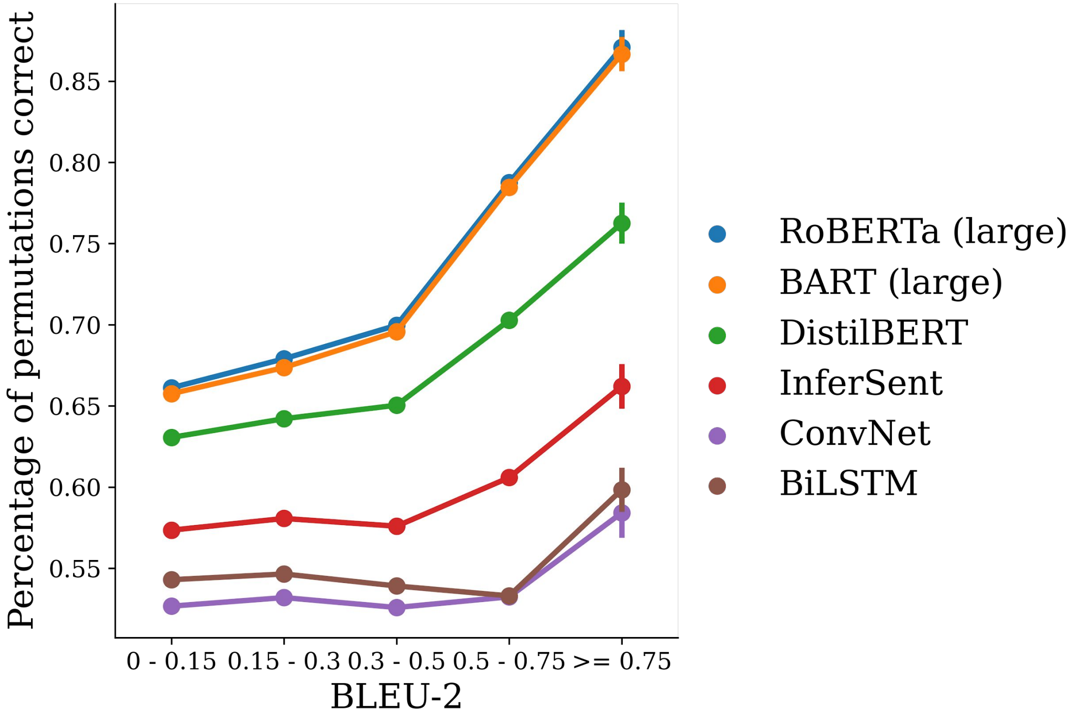
*although miscalibration might also come into play.

75

Which permutations do our models accept?

Preserving local word order leads to accepted permutations

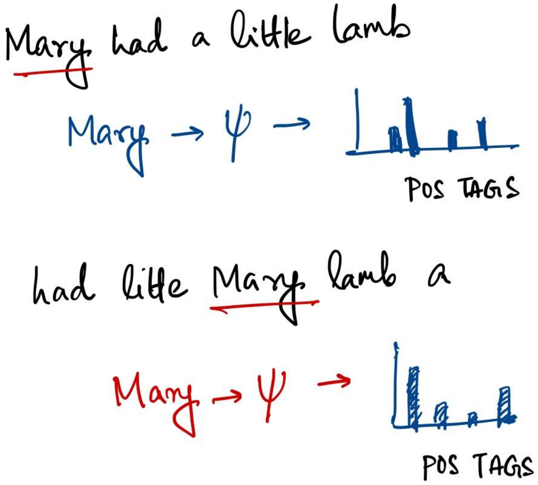
Percentage of permutations correct increases with more bi-gram overlap!

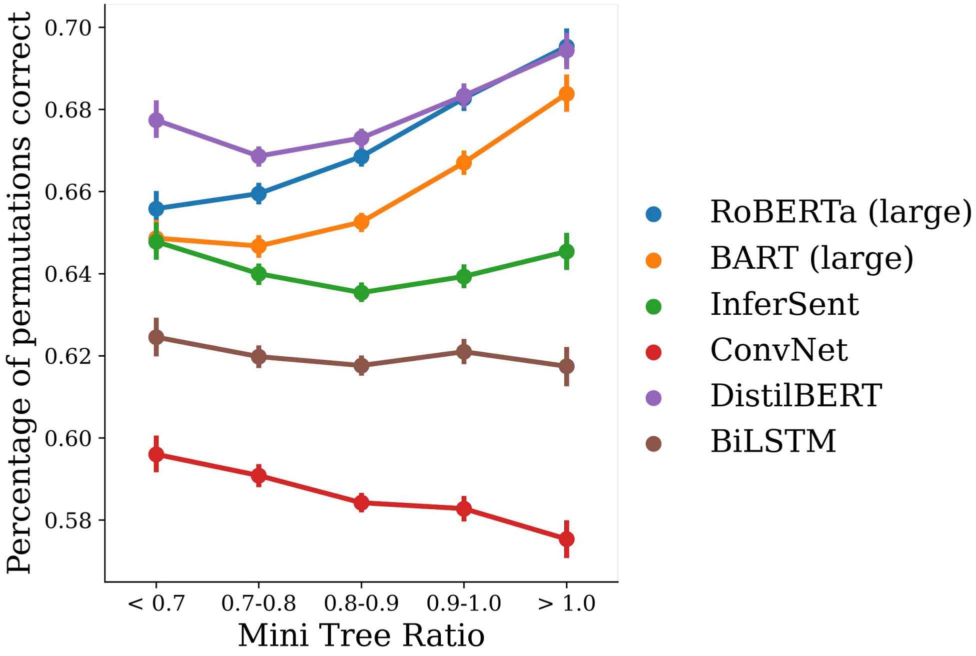


(BLEU-3 and BLEU-4 were too low to compare)

77

Transformer LMs aren't *entirely* BOW, they can handle *some* more abstract syntactic information





78

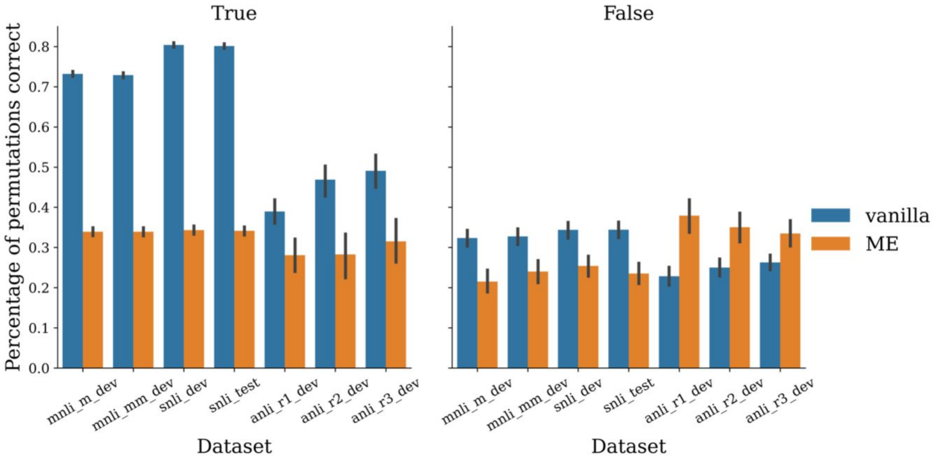
Initial Attempt: Max Entropy Training

A simple technique, but it **works**!

$$\mathcal{L} = \underset{\theta}{\operatorname{argmin}} \sum_{((p,h),y)} y \log(p(y|(p,h);\theta)) + \sum_{i=1}^{n} \mathbf{H}\left(y|(\hat{p}_i, \hat{h}_i);\theta\right)$$

- Accuracy is constant while the percentage of accepted permutations reduced considerably!

- However, there's still room to improve!



Similar approach concurrently by Gupta et al 2021

Human Analysis

Evaluator	Accuracy	Macro F1	Acc on D^c	Acc on D^f
X	0.581 $\pm$0.068	0.454	0.649 $\pm$0.102	0.515 $\pm$0.089
Y	0.378 $\pm$0.064	0.378	0.411 $\pm$0.098	0.349 $\pm$0.087

- 200 permuted sentences of varying length
- Annotators are “experts” in NLI

80

Human Analysis

Evaluator	Accuracy	Macro F1	Acc on D^c	Acc on D^f
X	0.581 $\pm$0.068	0.454	0.649 $\pm$0.102	0.515 $\pm$0.089
Y	0.378 $\pm$0.064	0.378	0.411 $\pm$0.098	0.349 $\pm$0.087

- 200 permuted sentences of varying length - **RoBERTa gets all of them "correct"!**
- Annotators are "experts" in NLI

Note: concurrent work on various perturbations of the GLUE Benchmark finds "turkers can only 'predict' the correct label for invalid examples in 35%" of cases (Gupta et al 2021; AAAI)

81

Once again, this time, in Chinese!

Just to verify this, we looked into another language…

Similar issue in Chinese OCNLI corpus!

This isn't a tokenization complication, or some quirk of English.

Model	$\mathcal{A}$	Ω_{max}	$\mathcal{P}^c$	$\mathcal{P}^f$	Ω_{rand}
RoBERTa (large)	**0.784**	0.988	0.726	0.339	0.773
InferSent	0.573	0.931	0.771	0.265	0.615
ConvNet	0.407	0.752	0.808	0.199	0.426
BiLSTM	0.566	0.963	0.701	0.271	0.611

Hu, Richardson, Xu, Li, Kuebler, Moss 2020 (EMNLP)
OCNLI: Original Chinese Natural Language Inference

82

What can we do about it?

Preliminary attempt : Entropy maximization

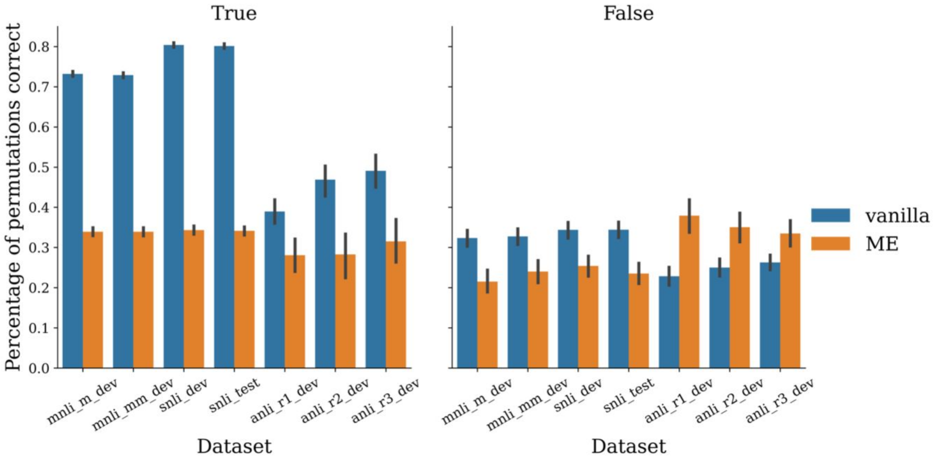
Initial Attempt: Max Entropy Training

A simple technique, but it **works**!

$$\mathcal{L} = \underset{\theta}{\operatorname{argmin}} \sum_{((p,h),y)} y \log(p(y|(p,h);\theta)) + \sum_{i=1}^{n} \mathbf{H}\left(y|(\hat{p}_i,\hat{h}_i);\theta\right)$$

- Accuracy is constant while the percentage of accepted permutations reduced considerably!

- However, there's still room to improve!



Similar approach concurrently by Gupta et al 2021

84

Thank You

https://arxiv.org/abs/2101.00010
https://github.com/facebookresearch/unlu

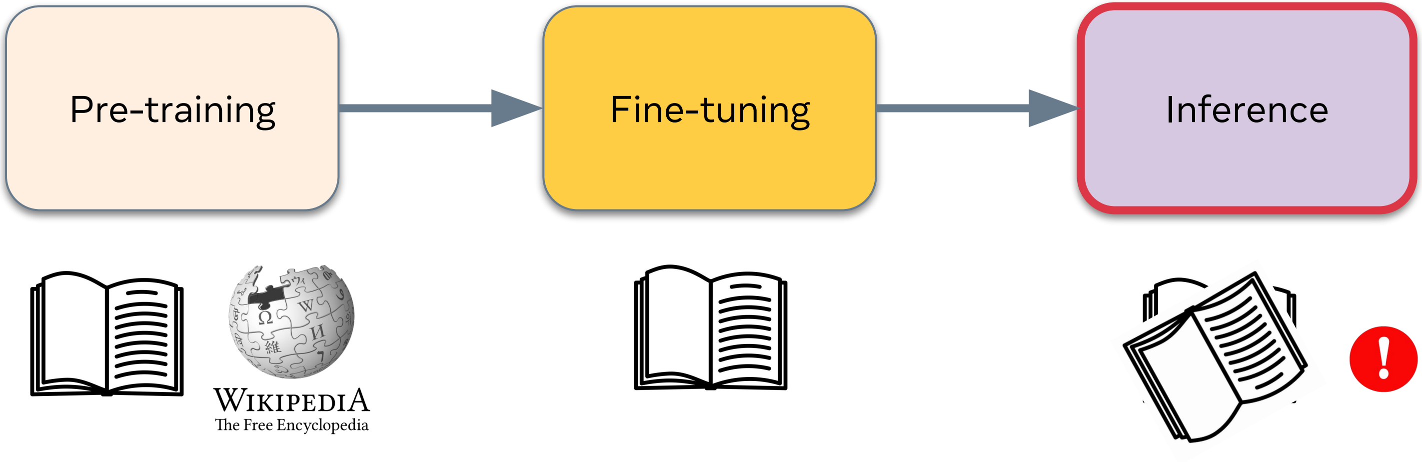
It is not enough that models should succeed where humans succeed, they should also fail where humans fail.

McGill | School of Computer Science

Mila

MLM & Distributional Hypothesis

“BERT rediscovers the classical NLP pipeline”



Alternative Hypothesis

Success of large scale models might just be explained by **Distributional Hypothesis** instead of internal representation of “NLP Pipeline”

“A word is characterized by the company it keeps”

Harris, 1954; Firth, 1957

BERT rediscovers the NLP pipeline

- **Tenney et al 2019** uses various probing tasks and conclude that BERT appears to have recreated an NLP pipeline in the expected sequence: POS tagging, parsing, NER, semantic roles, coreference

- **Manning et al 2020, Hewitt et al 2019** show evidence that BERT's MLM self-supervision learns syntactic grammatical structures and coreference resolution

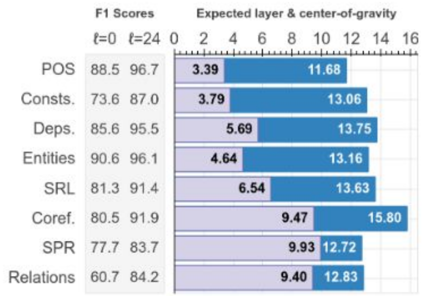
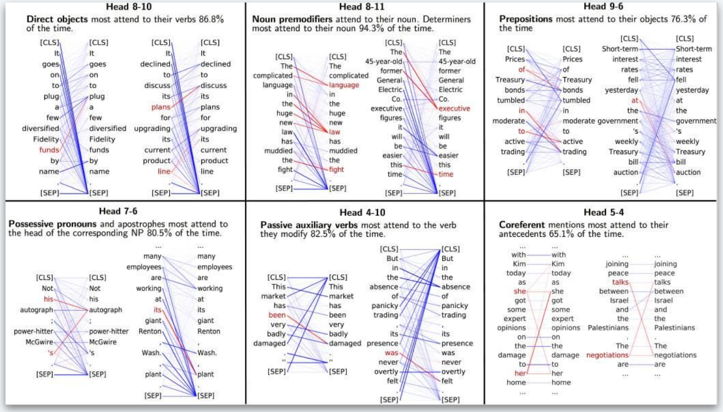


Figure 1: Summary statistics on BERT-large. Columns on left show F1 dev-set scores for the baseline ($P_\tau^{(0)}$) and full-model ($P_\tau^{(L)}$) probes. Dark (blue) are the mixing weight center of gravity (Eq. 2); light (purple) are the expected layer from the cumulative scores (Eq. 4).



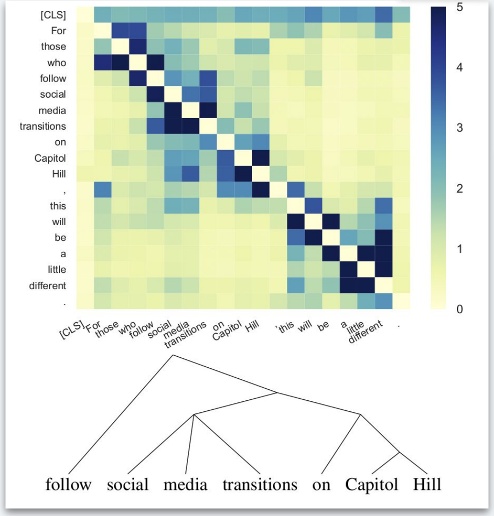
89

Do MLMs understand syntax?

- Recently large scale Transformer-based Language models (TLMs) have exceeded RNN’s performance on almost all NLU tasks

- Several papers claim these TLMs “*understand syntax*” [1] [2] [3]

- *“BERT rediscovers the classical NLP pipeline”* [4]



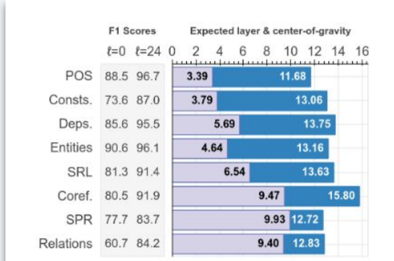
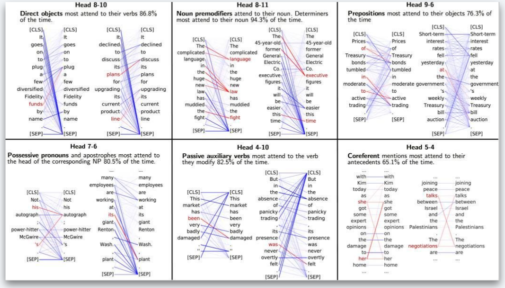


Figure 1: Summary statistics on BERT-large. Columns on left show F1 dev-set scores for the baseline ($P_\tau^{(0)}$) and full-model ($P_\tau^{(L)}$) probes. Dark (blue) are the mixing weight center of gravity (Eq. 2); light (purple) are the expected layer from the cumulative scores (Eq. 4).



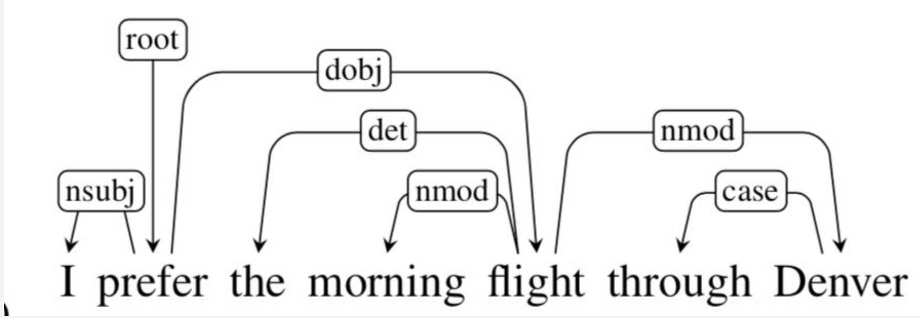
[1] John Hewitt and Christopher D Manning. *A structural prove for finding syntax in word representations*. NAACL 2019
[2] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. *Emergent linguistic structure in artificial neural networks trained by self-supervision*, PNAS 2020
[3] Ganesh Jawahar, Benoit Sagot and Djame Seddah. *What does BERT learn about structure of language?* ACL 2019
[4] Ian Tenney, Dipanjan Das, and Ellie Pavlick. *Bert rediscovers the classical nlp pipeline*. ACL 2019

90

Is syntactic understanding necessary for language understanding?

- A natural and common perspective in most formal theories of linguistics is that knowing natural language requires you to know the syntax

- Knowing the syntax of a sentence = being sensitive to at least the “order of the words” in the sentence



91

The sentence superiority effect

- Humans are known to exhibit a sentence superiority effect

- Given a normal sentence and a scrambled sentence, humans are found to perform significantly worse on the latter, with worse task performance.

	position 1	*position 2*
normal	our fox can fly	that was not red
scrambled	our can fly fox	not was red that

	position 3	*position 4*
normal	she can work now	the guy did this
scrambled	now she work can	guy did the this

JOURNAL ARTICLE

The Time it Takes to See and Name Objects

James McKeen Cattell

Mind

Vol. 11, No. 41 (Jan., 1886), pp. 63-65 (3 pages)

Published By: Oxford University Press

Joshua Snell, Jonathan Grainger, The sentence superiority effect revisited, Cognition, Volume 168, 2017, Pages 217-221

92

Investigation of distributional hypothesis through word order

- RoBERTa (base) - 125M parameters, 768 hidden size, 12 layers
- Data: BookWiki corpus (16GB)
- “no word should appear in its original position”
- N-gram shuffled corpus, where n=1,2,3,4

They are commonly known as daturas, but also known as devil's trumpets, not to be confused with angel's trumpets, its closely related genus "Brugmansia"

be They angel's also but trumpets, genus related devil's as commonly closely known its daturas, trumpets, as "Brugmansia". confused with known are to not

Word-order as proxy for syntax

- Word order information should be crucial for any syntactic pipeline
- Without syntax, ***many linguistic constructions are undefined*** (Chomsky, 1957)

They are commonly known as daturas, but also known as devil's trumpets, not to be confused with angel's trumpets, its closely related genus "Brugmansia"

be They angel's also but trumpets, genus related devil's as commonly closely known its daturas, trumpets, as "Brugmansia". confused with known are to not

94

Alternative Hypothesis

Distributional Hypothesis

“A word is characterized by the company it keeps”

Harris, 1954; Firth, 1957

Alternative Hypothesis

If an MLM has learned the "*the kind of abstractions we intuitively believe are important for representing natural language*":

- ✅ It should be sensitive to syntactic perturbations
- ✅ It should not learn the NLP pipeline if trained on un-syntactic data

Sentence randomization

- N-gram based randomization

- Given n, sample n-grams from a given sentence

- Convert n-grams to joined tokens

- Randomly shuffle the tokens in the sentence, such that “no word should appear in its original position”

> *They are commonly known as daturas, but also known as devil's trumpets, not to be confused with angel's trumpets, its closely related genus "Brugmansia"*

> *be They angel's also but trumpets, genus related devil's as commonly closely known its daturas, trumpets, as "Brugmansia". confused with known are to not*

97

Experimental Setup

Control

- RoBERTa (base) - 125M parameters, 768 hidden size, 12 layers
- Data: BookWiki corpus (16GB)
- Training: 100K updates, 8K batch size, 20k warmup steps, 6e-4 LR

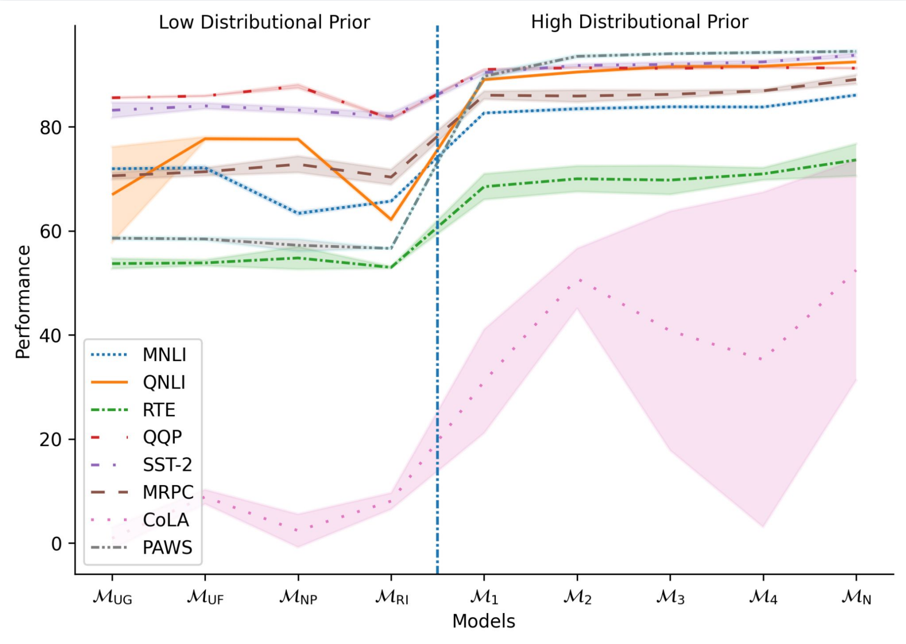
Ro  a

FAIRSEQ

 98

Results: Downstream Tasks

- Subset of GLUE benchmark : RTE, MRPC, MNLI, CoLA, QNLI, QQP, SST-2
- Paragraph adversaries for Word scrambling (PAWS)



99

Results: Downstream Tasks

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\mathrm{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21.2
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10.0
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.80
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23.0
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32.2
$\mathcal{M}_{\mathrm{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.60
$\mathcal{M}_{\mathrm{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.20
$\mathcal{M}_{\mathrm{UF}}$	77.69 +/- 0.4	53.84 +/- 0.6	85.92 +/- 0.1	84.00 +/- 0.6	71.35 +/- 0.8	58.43 +/- 0.3	72.10 +/- 0.4 / 72.58 +/- 0.4	8.89 +/- 1.40
$\mathcal{M}_{\mathrm{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.10

Table 1: GLUE and PAWS-Wiki dev set results on different RoBERTa (base) models trained on variants of the BookWiki corpus (with mean and std). The top row is the original model, the middle half contains our primary models under investigation, and the bottom half contains the ablations.

Inductive bias only

Results: Downstream Tasks

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\mathrm{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21.2
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10.0
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.80
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23.0
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32.2
$\mathcal{M}_{\mathrm{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.60
$\mathcal{M}_{\mathrm{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.20
$\mathcal{M}_{\mathrm{UF}}$	77.69 +/- 0.4	53.84 +/- 0.6	85.92 +/- 0.1	84.00 +/- 0.6	71.35 +/- 0.8	58.43 +/- 0.3	72.10 +/- 0.4 / 72.58 +/- 0.4	8.89 +/- 1.40
$\mathcal{M}_{\mathrm{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.10

Table 1: GLUE and PAWS-Wiki dev set results on different RoBERTa (base) models trained on variants of the BookWiki corpus (with mean and std). The top row is the original model, the middle half contains our primary models under investigation, and the bottom half contains the ablations.

Advantage with word phrases

Results: Downstream Tasks

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\text{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21.2
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10.0
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.80
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23.0
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32.2
$\mathcal{M}_{\text{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.60
$\mathcal{M}_{\text{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.20
$\mathcal{M}_{\text{UF}}$	77.69 +/- 0.4	53.84 +/- 0.6	85.92 +/- 0.1	84.00 +/- 0.6	71.35 +/- 0.8	58.43 +/- 0.3	72.10 +/- 0.4 / 72.58 +/- 0.4	8.89 +/- 1.40
$\mathcal{M}_{\text{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.10

Table 1: GLUE and PAWS-Wiki dev set results on different RoBERTa (base) models trained on variants of the BookWiki corpus (with mean and std). The top row is the original model, the middle half contains our primary models under investigation, and the bottom half contains the ablations.

Huge improvement, just with distributional prior

Results: Downstream Tasks

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\text{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21.2
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10.0
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.80
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23.0
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32.2
$\mathcal{M}_{\text{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.60
$\mathcal{M}_{\text{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.20
$\mathcal{M}_{\text{UF}}$	77.69 +/- 0.4	53.84 +/- 0.6	85.92 +/- 0.1	84.00 +/- 0.6	71.35 +/- 0.8	58.43 +/- 0.3	72.10 +/- 0.4 / 72.58 +/- 0.4	8.89 +/- 1.40
$\mathcal{M}_{\text{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.10

Table 1: GLUE and PAWS-Wiki dev set results on different RoBERTa (base) models trained on variants of the BookWiki corpus (with mean and std). The top row is the original model, the middle half contains our primary models under investigation, and the bottom half contains the ablations.

Almost equivalent to the original model pre-training!

Results: Downstream Tasks

Model	QNLI	RTE	QQP	SST-2	MRPC	PAWS	MNLI-m/mm	CoLA
$\mathcal{M}_{\text{N}}$	92.45 +/- 0.2	73.62 +/- 3.1	91.25 +/- 0.1	93.75 +/- 0.4	89.09 +/- 0.9	94.49 +/- 0.2	86.08 +/- 0.2 / 85.4 +/- 0.2	52.45 +/- 21.2
$\mathcal{M}_1$	89.05 +/- 0.2	68.48 +/- 2.5	91.01 +/- 0.0	90.41 +/- 0.4	86.06 +/- 0.8	89.69 +/- 0.6	82.64 +/- 0.1 / 82.67 +/- 0.2	31.08 +/- 10.0
$\mathcal{M}_2$	90.51 +/- 0.1	70.00 +/- 2.5	91.33 +/- 0.0	91.78 +/- 0.3	85.90 +/- 1.2	93.53 +/- 0.3	83.45 +/- 0.3 / 83.54 +/- 0.3	50.83 +/- 5.80
$\mathcal{M}_3$	91.56 +/- 0.4	69.75 +/- 2.8	91.22 +/- 0.1	91.97 +/- 0.5	86.22 +/- 0.8	94.03 +/- 0.1	83.83 +/- 0.2 / 83.71 +/- 0.1	40.78 +/- 23.0
$\mathcal{M}_4$	91.65 +/- 0.1	70.94 +/- 1.2	91.39 +/- 0.1	92.46 +/- 0.3	86.90 +/- 0.3	94.26 +/- 0.2	83.79 +/- 0.2 / 83.94 +/- 0.3	35.25 +/- 32.2
$\mathcal{M}_{\text{RI}}$	62.17 +/- 0.4	52.97 +/- 0.2	81.53 +/- 0.2	82.0 +/- 0.7	70.32 +/- 1.5	56.62 +/- 0.0	65.70 +/- 0.2 / 65.75 +/- 0.3	8.06 +/- 1.60
$\mathcal{M}_{\text{NP}}$	77.59 +/- 0.3	54.78 +/- 2.2	87.78 +/- 0.4	83.21 +/- 0.6	72.78 +/- 1.6	57.22 +/- 1.2	63.35 +/- 0.4 / 63.63 +/- 0.2	2.37 +/- 3.20
$\mathcal{M}_{\text{UF}}$	77.69 +/- 0.4	53.84 +/- 0.6	85.92 +/- 0.1	84.00 +/- 0.6	71.35 +/- 0.8	58.43 +/- 0.3	72.10 +/- 0.4 / 72.58 +/- 0.4	8.89 +/- 1.40
$\mathcal{M}_{\text{UG}}$	66.94 +/- 9.2	53.70 +/- 1.0	85.57 +/- 0.1	83.17 +/- 1.5	70.57 +/- 0.7	58.59 +/- 0.3	71.93 +/- 0.2 / 71.33 +/- 0.5	0.92 +/- 2.10

Table 1: GLUE and PAWS-Wiki dev set results on different RoBERTa (base) models trained on variants of the BookWiki corpus (with mean and std). The top row is the original model, the middle half contains our primary models under investigation, and the bottom half contains the ablations.

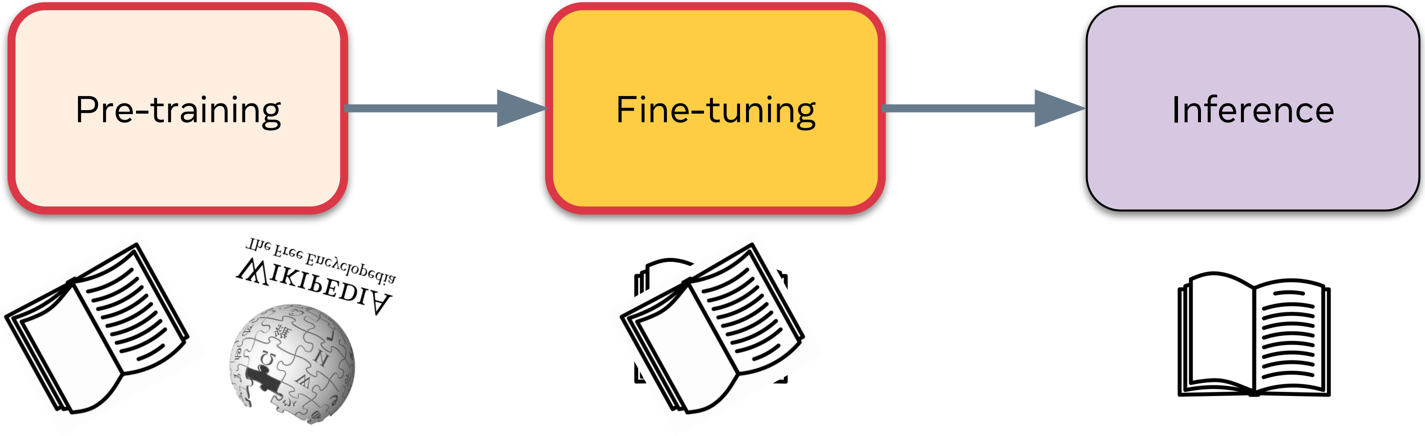
Word order reliant - CoLA (Matthews correlation)

What is the source of word order?

Possible explanations:

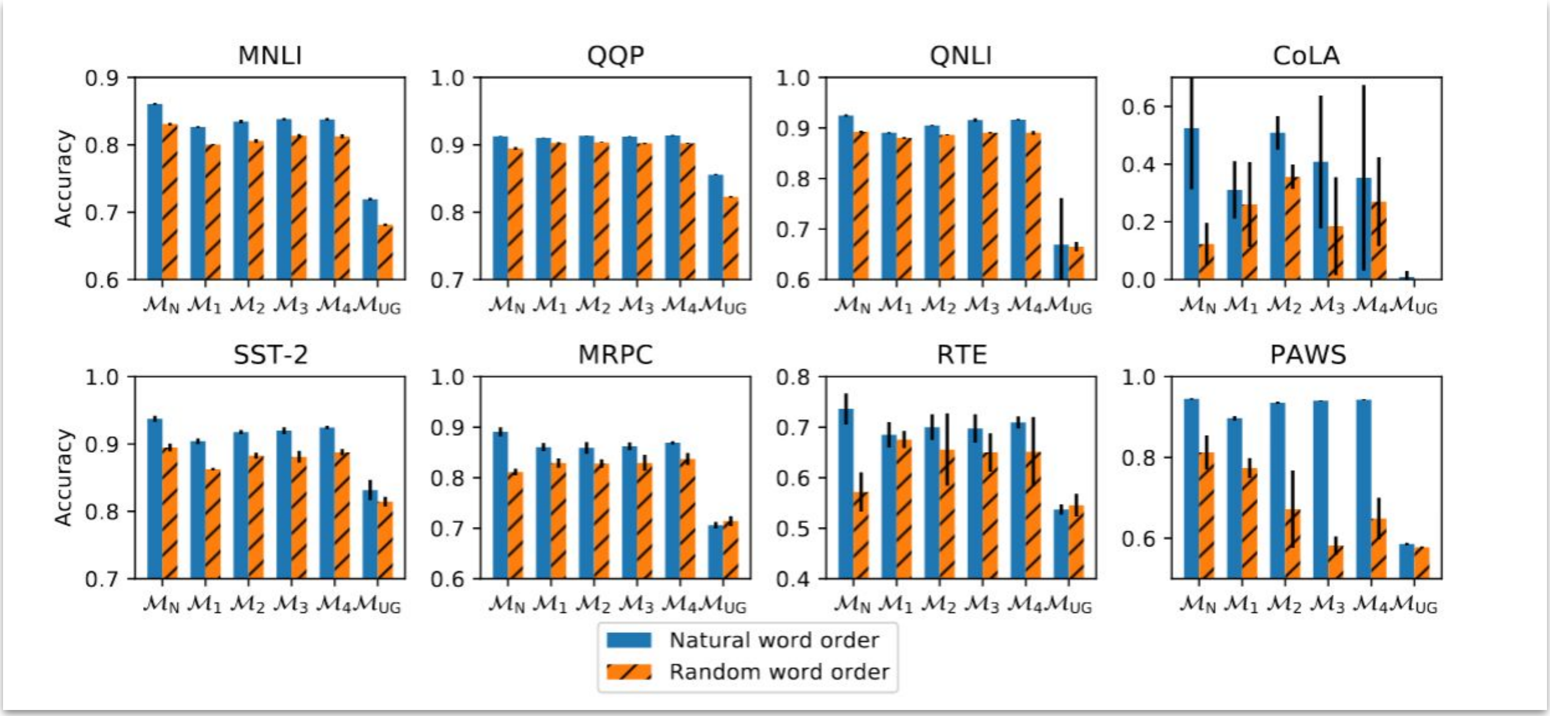
1. Tasks do not need word order information to solve them
2. The tasks need some word order information, but can be largely learned from fine-tuning

Where does BERT learn word order?



What is the source of word order?

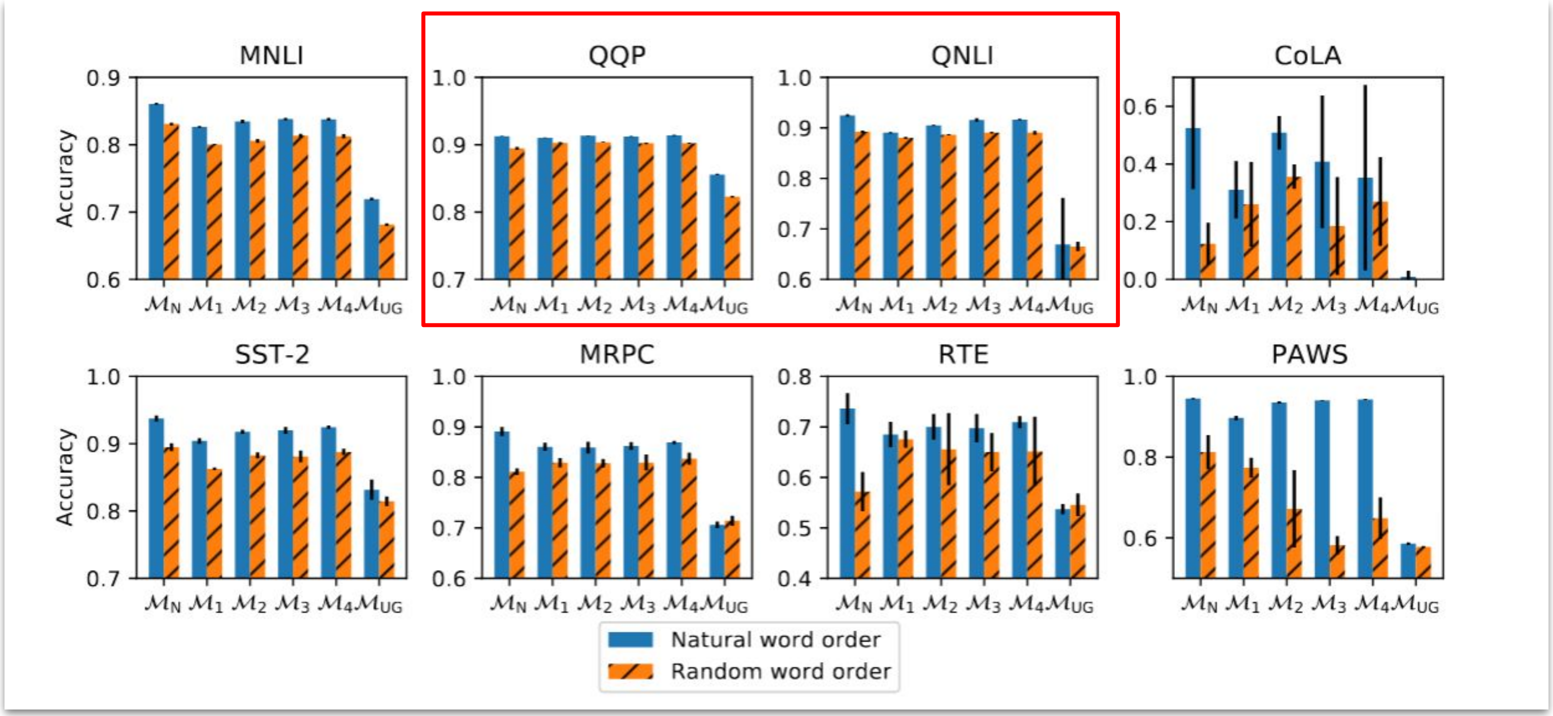
Evidence for both hypothesis!



What is the source of word order?

Evidence for both hypothesis!

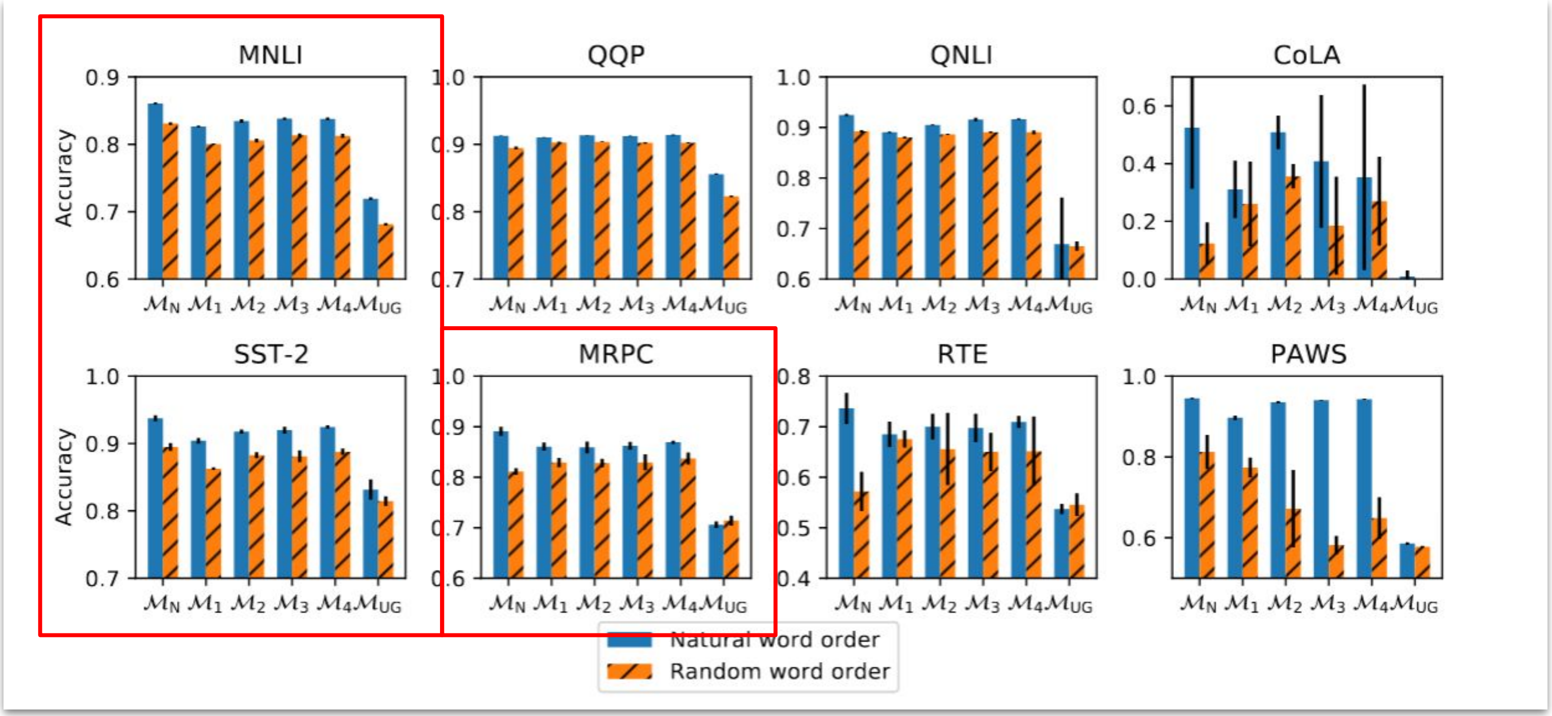
Word order not important



What is the source of word order?

Evidence for both hypothesis!

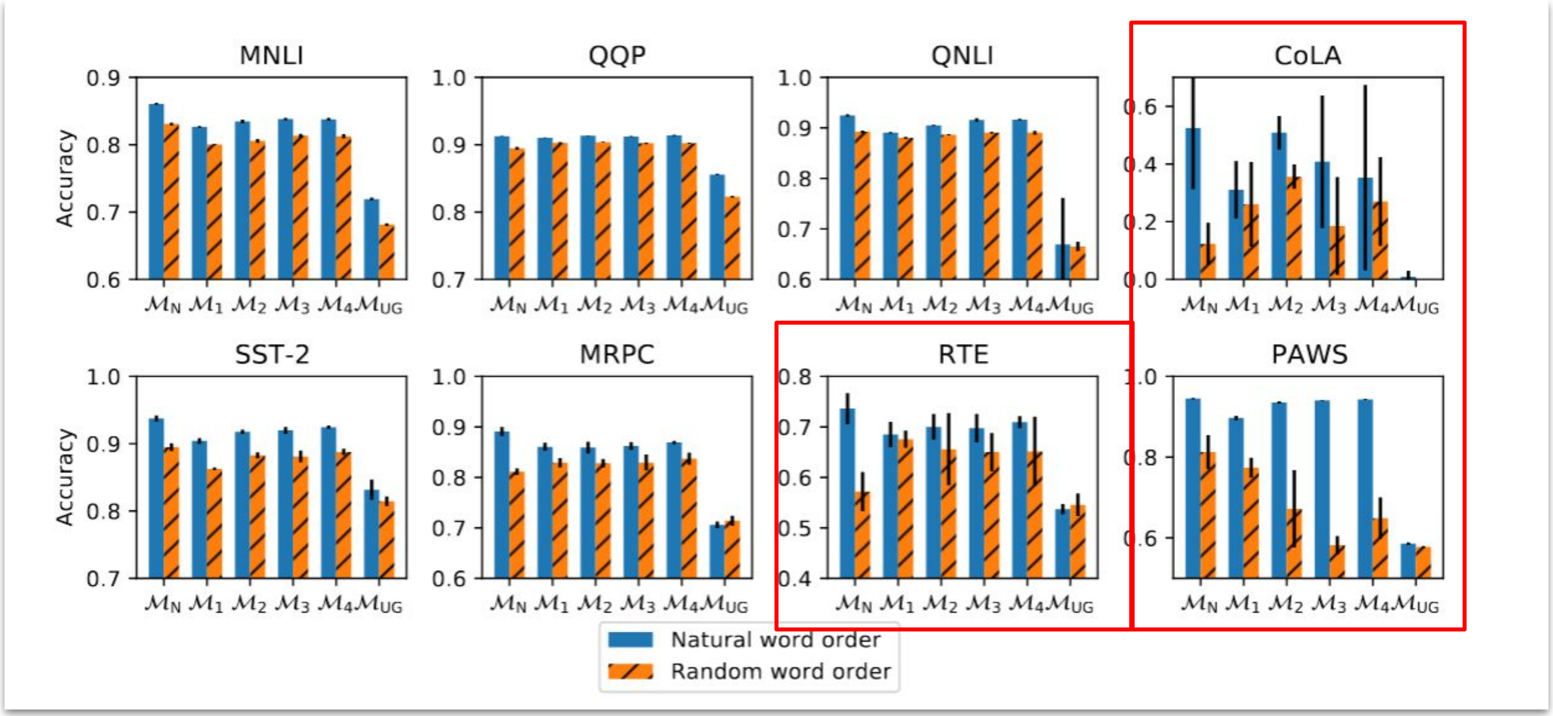
Lexical information is enough



What is the source of word order?

Evidence for both hypothesis!

Word order is important, but learned during fine-tuning



Fine-tuning experiments

name	fine-tune-train	fine-tune-eval	MNLI	QNLI	RTE	CoLA	MRPC	SST-2	PAWS
$\mathcal{M}_{\mathrm{N}}$	natural	natural	86.08 +/- 0.15	92.45 +/- 0.24	73.62 +/- 3.09	52.44 +/- 21.22	89.09 +/- 0.88	93.75 +/- 0.44	94.49 +/- 0.18
	natural	shuffled	68.11 +/- 0.52	81.08 +/- 0.38	56.72 +/- 3.29	4.77 +/- 1.82	75.94 +/- 1.01	80.78 +/- 0.37	62.22 +/- 0.09
	shuffled	natural	82.99 +/- 0.16	89.32 +/- 0.23	57.9 +/- 4.71	0.0 +/- 0.0	79.71 +/- 2.57	89.12 +/- 0.5	72.03 +/- 13.79
	shuffled	shuffled	79.96 +/- 0.1	87.51 +/- 0.09	59.07 +/- 3.2	1.4 +/- 2.17	79.17 +/- 0.35	86.11 +/- 0.5	65.15 +/- 0.48
$\mathcal{M}_1$	natural	natural	82.64 +/- 0.15	89.05 +/- 0.15	68.48 +/- 2.51	31.07 +/- 9.97	85.97 +/- 0.89	90.41 +/- 0.43	89.69 +/- 0.59
	natural	shuffled	76.67 +/- 0.34	87.21 +/- 0.17	65.8 +/- 6.11	23.06 +/- 5.3	81.84 +/- 0.43	83.94 +/- 0.33	62.86 +/- 0.19
	shuffled	natural	79.87 +/- 0.1	87.81 +/- 0.36	65.65 +/- 2.33	24.53 +/- 13.63	82.51 +/- 0.82	86.45 +/- 0.41	73.34 +/- 6.88
	shuffled	shuffled	79.75 +/- 0.0	88.21 +/- 0.24	64.88 +/- 6.32	22.43 +/- 10.79	82.65 +/- 0.42	86.25 +/- 0.4	63.15 +/- 2.2
$\mathcal{M}_{\mathrm{UG}}$	natural	natural	71.93 +/- 0.21	66.94 +/- 9.21	53.7 +/- 1.01	0.92 +/- 2.06	70.57 +/- 0.66	83.17 +/- 1.5	58.59 +/- 0.33
	natural	shuffled	62.27 +/- 0.57	63.13 +/- 7.13	52.42 +/- 2.77	0.09 +/- 0.21	70.56 +/- 0.33	79.41 +/- 0.63	56.91 +/- 0.16
	shuffled	natural	67.62 +/- 0.3	66.49 +/- 0.49	52.17 +/- 1.26	0.0 +/- 0.0	70.37 +/- 0.93	79.93 +/- 1.01	57.59 +/- 0.29
	shuffled	shuffled	67.02 +/- 0.33	66.24 +/- 0.33	53.44 +/- 0.53	0.08 +/- 0.18	70.28 +/- 0.62	80.05 +/- 0.4	57.38 +/- 0.16

Table 9: Fine-tuning evaluation by varying different sources of word order (with mean and std dev). We vary the word order contained in the pre-trained model ($\mathcal{M}_{\mathrm{N}}$,$\mathcal{M}_1$,$\mathcal{M}_{\mathrm{UG}}$); in fine-tuning training set (natural and shuffled); and in fine-tuning evaluation (natural and shuffled). Here, *shuffled* corresponds to unigram shuffling of words in the input. In case of fine-tune evaluation containing shuffled input, we evaluate on a sample of 100 unigram permutations for each data point in the dev set of the corresponding task.

Perplexity scores

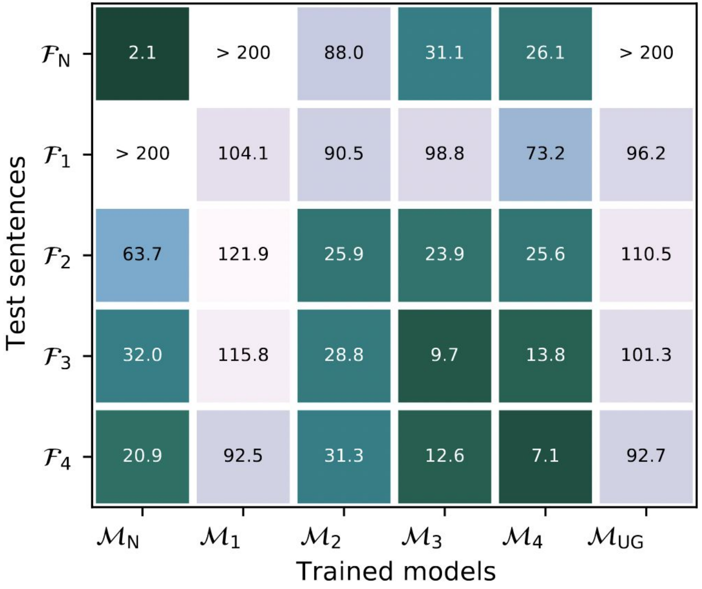


Figure 4: BPPL scores per model per test scenario.

RDA Analysis

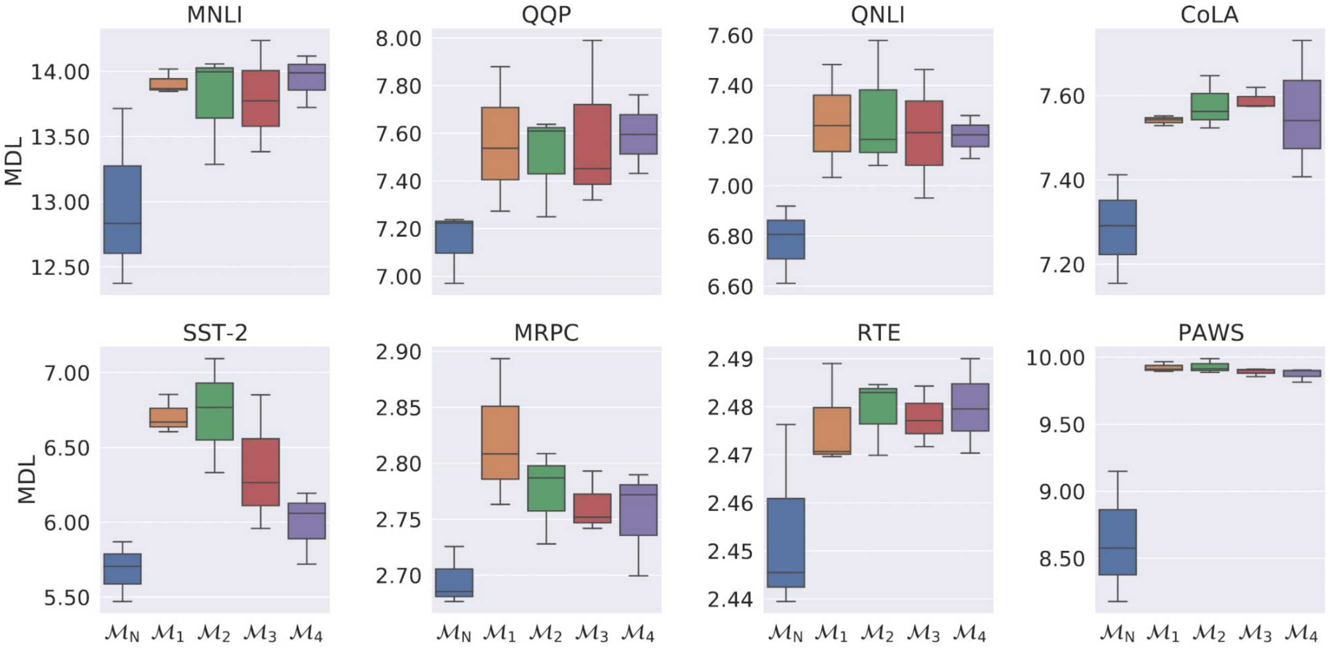


Figure 5: Rissanen Data Analysis (Perez et al., 2021) on the GLUE benchmark and PAWS datasets. The lower minimum description length (MDL, measured in kilobits), the better the learning ability of the model.

GLUE improvement during pre-training

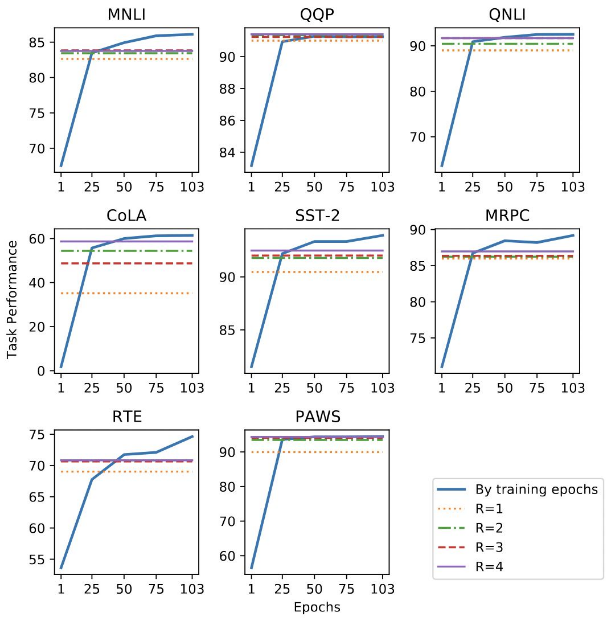


Figure 6: Comparison among GLUE task performance from different steps in pre-training of RoBERTa on BookWiki Corpus.

What do we learn by Probing?

- Parametric Probing
  - Dependency arc labelling
  - POS Tagging
  - Dependency parsing
  - SentEval - 10 probes
- Non-parametric Probing
  - Singular/Plural inflection verb stimuli

Parametric Probing

- **POS Tagging**
- Pareto Probing framework (Pimentel et al, 2020)
- Linear and MLP probe
- UD EWT and PTB corpus

Model	UD EWT		PTB	
	MLP	Linear	MLP	Linear
$\mathcal{M}_{\mathrm{N}}$	93.74 +/- 0.15	88.82 +/- 0.42	97.07 +/- 0.38	93.1 +/- 0.65
$\mathcal{M}_1$	88.60 +/- 3.43	80.76 +/- 3.38	95.33 +/- 0.37	87.83 +/- 1.86
$\mathcal{M}_2$	93.39 +/- 0.45	87.58 +/- 1.06	96.96 +/- 0.15	91.80 +/- 0.50
$\mathcal{M}_3$	92.89 +/- 0.65	86.78 +/- 1.32	97.03 +/- 0.13	91.70 +/- 0.70
$\mathcal{M}_4$	92.83 +/- 0.61	87.23 +/- 0.77	96.96 +/- 0.12	92.08 +/- 0.39
$\mathcal{M}_{\mathrm{UG}}$	89.10 +/- 0.21	79.75 +/- 0.5	94.12 +/- 0.01	84.15 +/- 0.51

Table 3: Accuracy on the part-of-speech labelling task (POS) on two datasets, UD EWT and PTB, using the Pareto Probing framework (Pimentel et al., 2020a).

Parametric Probing

- **Dependency Arc labelling**
- Pareto Probing framework (Pimentel et al, 2020)
- Linear and MLP probe
- UD EWT and PTB corpus

Model	UD EWT		PTB	
	MLP	Linear	MLP	Linear
$\mathcal{M}_{\mathrm{N}}$	89.63 +/- 0.60	84.35 +/- 0.78	93.96 +/- 0.63	88.35 +/- 1.00
$\mathcal{M}_1$	83.55 +/- 3.31	75.26 +/- 3.08	91.10 +/- 0.38	82.34 +/- 1.37
$\mathcal{M}_2$	88.57 +/- 0.68	82.05 +/- 1.10	93.27 +/- 0.26	86.88 +/- 0.87
$\mathcal{M}_3$	88.69 +/- 1.09	82.37 +/- 1.26	93.46 +/- 0.29	87.12 +/- 0.72
$\mathcal{M}_4$	88.66 +/- 0.76	82.58 +/- 1.04	93.49 +/- 0.33	87.30 +/- 0.79
$\mathcal{M}_{\mathrm{UG}}$	84.93 +/- 0.34	76.30 +/- 0.52	89.98 +/- 0.43	78.59 +/- 0.68

Table 4: Accuracy on the dependency arc labelling task (DAL) on two datasets, UD EWT and PTB, using the Pareto Probing framework (Pimentel et al., 2020a).

117

Parametric Probing

- **Dependency Parsing**
- Pareto Probing framework (Pimentel et al, 2020)
- Linear and MLP probe
- UD EWT and PTB corpus

Model	UD EWT		PTB	
	MLP	Linear	MLP	Linear
$\mathcal{M}_{\mathrm{N}}$	80.41 +/- 0.85	66.26 +/- 1.59	86.99 +/- 1.49	66.47 +/- 2.77
$\mathcal{M}_1$	69.26 +/- 6.00	56.24 +/- 5.05	79.43 +/- 0.96	57.20 +/- 2.76
$\mathcal{M}_2$	78.22 +/- 0.88	64.96 +/- 2.32	84.72 +/- 0.55	64.69 +/- 2.50
$\mathcal{M}_3$	77.80 +/- 3.09	64.89 +/- 2.63	85.89 +/- 1.01	66.11 +/- 1.68
$\mathcal{M}_4$	78.04 +/- 2.06	65.61 +/- 1.99	85.62 +/- 1.09	66.49 +/- 2.02
$\mathcal{M}_{\mathrm{UG}}$	74.15 +/- 0.93	65.69 +/- 7.35	80.07 +/- 0.79	57.28 +/- 1.42

Table 2: Unlabeled Attachment Score (UAS) on the dependency parsing task (DEP) on two datasets, UD EWT and PTB, using the Pareto Probing framework (Pimentel et al., 2020a)

118

Parametric Probing

"BERT embeds a rich hierarchy of linguistic signals: surface information at the bottom, syntactic information in the middle, semantic information at the top"

Jawahar et al, 2019

- **SentEval**
- 10 probing tasks ranging from Lexical (surface), Syntactic and Semantic

Model	Length (Surface)	WordContent (Surface)	TreeDepth (Syntactic)	TopConstituents (Syntactic)	BigramShift (Syntactic)	Tense (Semantic)	SubjNumber (Semantic)	ObjNumber (Semantic)	OddManOut (Semantic)	CoordInversion (Semantic)
$\mathcal{M}_{N}$	78.92 +/- 1.91	31.83 +/- 1.75	35.97 +/- 1.38	**78.26** +/- 4.08	**81.82** +/- 0.55	87.83 +/- 0.51	85.05 +/- 1.23	75.94 +/- 0.68	58.40 +/- 0.33	**70.87** +/- 2.46
$\mathcal{M}_1$	88.33 +/- 0.14	**64.03** +/- 0.34	40.24 +/- 0.20	70.94 +/- 0.38	58.37 +/- 0.40	87.88 +/- 0.08	83.49 +/- 0.12	**83.44** +/- 0.06	56.51 +/- 0.26	56.98 +/- 0.50
$\mathcal{M}_2$	**93.54** +/- 0.29	62.52 +/- 0.21	**41.40** +/- 0.32	74.31 +/- 0.29	75.44 +/- 0.14	**87.91** +/- 0.35	84.88 +/- 0.11	83.98 +/- 0.14	57.60 +/- 0.36	59.46 +/- 0.37
$\mathcal{M}_3$	91.52 +/- 0.16	48.81 +/- 0.26	38.63 +/- 0.61	70.29 +/- 0.31	77.36 +/- 0.12	86.74 +/- 0.12	83.83 +/- 0.38	80.99 +/- 0.26	57.01 +/- 0.21	60.00 +/- 0.26
$\mathcal{M}_4$	92.88 +/- 0.15	57.78 +/- 0.36	40.05 +/- 0.29	72.50 +/- 0.51	76.12 +/- 0.29	88.32 +/- 0.13	**85.65** +/- 0.13	82.95 +/- 0.05	**58.89** +/- 0.30	61.31 +/- 0.19
$\mathcal{M}_{UG}$	86.69 +/- 0.33	36.60 +/- 0.33	32.53 +/- 0.76	61.54 +/- 0.60	57.42 +/- 0.04	68.45 +/- 0.23	71.25 +/- 0.12	66.63 +/- 0.21	50.06 +/- 0.40	56.26 +/- 0.17

Table 5: SentEval Probing (Conneau et al., 2018; Conneau and Kiela, 2018) results on different model variants.

Syntactic : 2/3

Lexical : 0/2

Semantic: 1/5

119

Non - Parametric Probing

- No learnable parameters!
- Stimuli to predict the correct inflection (singular/plural) of focus verb
- Three datasets: Linzen et al 2016, Marvin & Linzen 2018, Gulordava et al 2018

Can identify syntax-related modeling failures that parametric ones do not!

A 13-year boy named Toby Lolness , who is just one and a half millimetres tall , <mask> in a civilization nestled in an oak tree .

lives live

P(good) > P(bad)

Model	Linzen et al. (2016) *	Gulordava et al. (2018b) *	Marvin and Linzen (2018)
$\mathcal{M}_{\mathrm{N}}$	91.17 +/- 2.6	68.66 +/- 11.6	88.05 +/- 6.5
$\mathcal{M}_4$	66.93 +/- 3.2	69.47 +/- 4.9	70.66 +/- 12.5
$\mathcal{M}_3$	64.60 +/- 2.7	66.10 +/- 5.9	73.82 +/- 15.7
$\mathcal{M}_2$	61.27 +/- 3.1	60.20 +/- 7.6	73.95 +/- 14.3
$\mathcal{M}_1$	58.96 +/- 1.8	68.10 +/- 14.4	70.69 +/- 11.6
$\mathcal{M}_{\mathrm{UG}}$	65.36 +/- 7.1	60.88 +/- 24.3	50.10 +/- 0.2

120

Future Work

- Investigate broader amount of tasks with unnatural pre-trained models
- Investigate NLG using a an unnatural pre-trained model
- Benefits in privacy: we can therefore release models trained with random word order with a little bit of performance loss but no way to recover original word order!

https://cs.mcgill.ca/~ksinha4

https://arxiv.org/abs/2104.06644

@koustuvsinha

Thanks for Listening!

Looking forwards to discuss more @ EMNLP 2021

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, Douwe Kiela


Always has been
Wait it's a bag-of-words model?


Adam Lerer @adamlerer · Apr 15
My hotel wall a good pretraining example is?
This KITCHEN is Love Seasoned with


HELLO, I'M PAPER BOTTLE
BERT models
distributional semantics

Alternate Hypothesis

Distributional Hypothesis - BERT may not be that different from Word2Vec

$$p(t \mid w; \theta) = \frac{e^{f(t,w)}}{\sum_{t' \in V} e^{f(t',w)}}$$

BPE

Defenestration

Non-linearity

Data + Compute

$$p(t \mid C(t); \theta) = \frac{e^{g(t,C(t))}}{\sum_{t' \in U} e^{g(t',C(t))}}$$

 123

Source of Word Order

name	fine-tune-train	fine-tune-eval	MNLI	QNLI	RTE	CoLA	MRPC	SST-2	PAWS
$\mathcal{M}_{\mathrm{N}}$	natural	natural	86.08 +/- 0.15	92.45 +/- 0.24	73.62 +/- 3.09	52.44 +/- 21.22	89.09 +/- 0.88	93.75 +/- 0.44	94.49 +/- 0.18
	natural	shuffled	68.11 +/- 0.52	81.08 +/- 0.38	56.72 +/- 3.29	4.77 +/- 1.82	75.94 +/- 1.01	80.78 +/- 0.37	62.22 +/- 0.09
	shuffled	natural	82.99 +/- 0.16	89.32 +/- 0.23	57.9 +/- 4.71	0.0 +/- 0.0	79.71 +/- 2.57	89.12 +/- 0.5	72.03 +/- 13.79
	shuffled	shuffled	79.96 +/- 0.1	87.51 +/- 0.09	59.07 +/- 3.2	1.4 +/- 2.17	79.17 +/- 0.35	86.11 +/- 0.5	65.15 +/- 0.48
$\mathcal{M}_{1}$	natural	natural	82.64 +/- 0.15	89.05 +/- 0.15	68.48 +/- 2.51	31.07 +/- 9.97	85.97 +/- 0.89	90.41 +/- 0.43	89.69 +/- 0.59
	natural	shuffled	76.67 +/- 0.34	87.21 +/- 0.17	65.8 +/- 6.11	23.06 +/- 5.3	81.84 +/- 0.43	83.94 +/- 0.33	62.86 +/- 0.19
	shuffled	natural	79.87 +/- 0.1	87.81 +/- 0.36	65.65 +/- 2.33	24.53 +/- 13.63	82.51 +/- 0.82	86.45 +/- 0.41	73.34 +/- 6.88
	shuffled	shuffled	79.75 +/- 0.0	88.21 +/- 0.24	64.88 +/- 6.32	22.43 +/- 10.79	82.65 +/- 0.42	86.25 +/- 0.4	63.15 +/- 2.2
$\mathcal{M}_{\mathrm{UG}}$	natural	natural	71.93 +/- 0.21	66.94 +/- 9.21	53.7 +/- 1.01	0.92 +/- 2.06	70.57 +/- 0.66	83.17 +/- 1.5	58.59 +/- 0.33
	natural	shuffled	62.27 +/- 0.57	63.13 +/- 7.13	52.42 +/- 2.77	0.09 +/- 0.21	70.56 +/- 0.33	79.41 +/- 0.63	56.91 +/- 0.16
	shuffled	natural	67.62 +/- 0.3	66.49 +/- 0.49	52.17 +/- 1.26	0.0 +/- 0.0	70.37 +/- 0.93	79.93 +/- 1.01	57.59 +/- 0.29
	shuffled	shuffled	67.02 +/- 0.33	66.24 +/- 0.33	53.44 +/- 0.53	0.08 +/- 0.18	70.28 +/- 0.62	80.05 +/- 0.4	57.38 +/- 0.16

Table 9: Fine-tuning evaluation by varying different sources of word order (with mean and std dev). We vary the word order contained in the pre-trained model ($\mathcal{M}_{\mathrm{N}}$,$\mathcal{M}_{1}$,$\mathcal{M}_{\mathrm{UG}}$); in fine-tuning training set (natural and shuffled); and in fine-tuning evaluation (natural and shuffled). Here, *shuffled* corresponds to unigram shuffling of words in the input. In case of fine-tune evaluation containing shuffled input, we evaluate on a sample of 100 unigram permutations for each data point in the dev set of the corresponding task.

124

More contributions …

NLU & NLG

- Learning an Unreferenced Metric for Online Dialogue Evaluation

  K Sinha, P Parthasarathi, J Wang, R Lowe, W Hamilton, J Pineau. ACL 2020

- Evaluating Gender Bias in Natural Language Inference

  S Sharma, M Dey, **K Sinha**. NeurIPS 2020 Workshop on Dataset Security

- Do translation systems fix their bias with more context? Mitigating gender bias in Neural Machine Translation models using extra-sentential information

  S Sharma, M Dey, **K Sinha**. NAACL 2022 Submission

Graph Representation Learning

- Evaluating Logical Generalization in Graph Neural Networks

  K Sinha, S Sodhani, J Pineau, W Hamilton. Arxiv Pre-print 2020

Vision

- COVID-19 Deterioration Prediction via Self-Supervised Representation Learning and Multi-Image Prediction

  A Sriram, M Muckley, **K Sinha**, F Shamout, J Pineau, K Geras, L Azour, Y Aphinyanaphongs, N Yakubova, W Moore. 2021, under review

Reproducibility

- Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program

  J Pineau, P Vincent-Lamarre, **K Sinha**, V Lariviere, A Beygelzimer, F d'Alche-Buc, E Fox, H Larochelle. JMLR 2020

- ML Reproducibility Challenge (2018 to present)

  K Sinha, J Dodge, S Luccioni, J Forde, S Raparthy, J Pineau, R Stojnic. 2021