Sentence Understanding with Neural Networks and Natural Language Inference

Adina Williams & **Sam Bowman**
NYU Linguistics NYU Linguistics/Center for Data Science

Computational Intelligence, Learning, Vision, and Robotics
October 30th 2017

https://wp.nyu.edu/adinawilliams/work/

Goal:
Build neural network models that understand sentence meaning.

What?

Translating sentences to machine-readable representations of the speaker's intended meaning.

aka. learning to extract something like *semantics*

Why?

Key part of:

- Dialog
- Translation
- Summarization
- Question Answering
- Sentiment analysis
- Information extraction
- Information retrieval

...

The goal

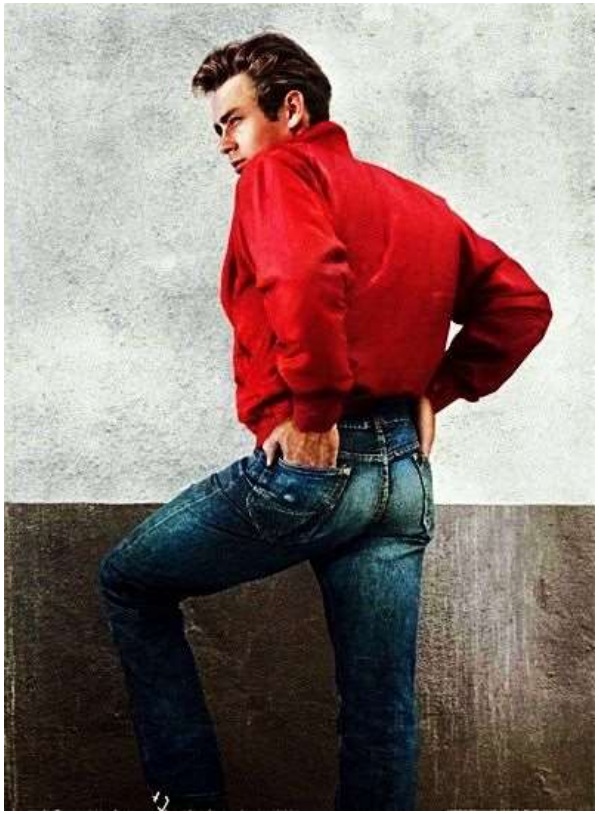
Build artificial neural network models that can understand and reason with the meanings of natural language sentences.

How do we measure success?

Natural Language Inference as a research task

Natural language inference (NLI)

also known as recognizing textual entailment (RTE)



James Byron Dean refused to move without blue jeans

{**entails**, contradicts, neither}

James Dean didn't dance without pants

The claim

If a model can learn to represent sentence meaning effectively:

> **The model should learn to do NLI relatively easily**.

The claim

If a model can learn to represent sentence meaning effectively:

The model should learn to do NLI relatively easily.

If a model can't:

The model will fail.

Judging understanding with NLI

To reliably perform well at NLI, your representations of meaning **must** handle the full complexity of compositional semantics:

- Lexical entailment
- Quantification
- Object and event coreference
- Lexical ambiguity and scope ambiguity
- Common sense background knowledge
- Modality (*might*, *should*, ...)

...

Why not other tasks?

Many popular evaluation tasks don't require much language understanding:

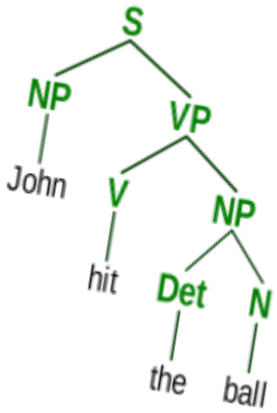
- Sentiment analysis
- Sentence similarity

...

NLI isn't the only task to require high-quality natural language understanding:

- Machine translation
- Question answering
- Goal-driven dialog
- Semantic parsing
- Syntactic parsing

...



But it's the easiest of these.

Part I

Multi-genre NLI Corpus

Adina Williams
Nikita Nangia
Samuel R. Bowman

Available for download now:
https://www.nyu.edu/projects/bowman/multinli

Natural language inference data

Our very own S. Bowman collected the **Stanford Natural Language Inference** Corpus in 2015

(w/ Gabor Angeli
Christopher Potts
Christopher D. Manning)

Corpus	Size	Natural	Validated
FraCaS	.3k	~	✓
RTE	7k	✓	✓
SICK	10k	✓	✓
SNLI	**570k**	✓	✓
DG	728k	~	
Levy	1,500k		
PPDB2	100,000k	~	

SNLI is showing its limitations

from a modelling perspective

- Little headroom left:
  - SotA: **88.6%**
  - Human performance: ~96%
- Short sentences (mean premise length 14)
  - Attention models can succeed without word order (Parikh et al. ‘16)

SNLI is showing its limitations

from a linguistics perspective

Falls short of goal of representing full range of sentence meaning:

Homogenous source sentences:
image captions from Flickr30k
(Young, Lai, Hodosh, and Hockenmaier, TACL '14)

Mostly **Imageable** sentences:

- What does "He has no honor" look like?
- Or "He can speak Spanish", or "He is my nephew", or "He died last August"?

…

What about spoken language?
What about formal v. informal speech?

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

The MultiGenre NLI Corpus

Genre	Train	Dev	Test
Captions (SNLI Corpus)	(550,152)	(10,000)	(10,000)
Fiction	77,348	2,000	2,000
Government	77,350	2,000	2,000
Slate	77,306	2,000	2,000
Switchboard (Telephone Speech)	83,348	2,000	2,000
Travel Guides	77,350	2,000	2,000
9/11 Report	0	2,000	2,000
Face-to-Face Speech	0	2,000	2,000
Letters	0	2,000	2,000
OUP (Nonfiction Books)	0	2,000	2,000
Verbatim (Magazine)	0	2,000	2,000
Total	**392,702**	**20,000**	**20,000**

genre-matched evaluation

genre-mismatched evaluation

MultiNLI is harder than SNLI

from a linguistics perspective

Pronouns (PTB)	34	**68**	34
Quantifiers	33	**63**	30
Modals (PTB)	<1	**28**	28
Negation (PTB)	5	**31**	26
'Wh' Words (PTB)	5	**30**	25
Belief Verbs	<1	**19**	18
Time Terms	19	**36**	17
Conversational Pivots	<1	**14**	14
Presupposition Triggers	8	**22**	14
Comparatives/Superlatives (PTB)	3	**17**	14
Conditionals	4	**15**	11
Tense Match (PTB)	62	**69**	7
Interjections (PTB)	<1	**5**	5
>20 Words	<1	**5**	5
Existentials (PTB)	5	**8**	3

from a modelling perspective

Genre	#Wds. Prem.	'S' parses Prem.	Hyp.	Agrmt.	Model Acc. ESIM	CBOW
SNLI	*14.1*	*74%*	*88%*	*89.0%*	*86.7%*	*80.6 %*
FICTION	14.4	94%	97%	89.4%	73.0%	67.5%
GOVERNMENT	24.4	90%	97%	87.4%	74.8%	67.5%
SLATE	21.4	94%	98%	87.1%	67.9%	60.6%
TELEPHONE	25.9	71%	97%	88.3%	72.2%	63.7%
TRAVEL	24.9	97%	98%	89.9%	73.7%	64.6%
9/11	20.6	98%	99%	90.1%	71.9%	63.2%
FACE-TO-FACE	18.1	91%	96%	89.5%	71.2%	66.3%
LETTERS	20.0	95%	98%	90.1%	74.7%	68.3%
OUP	25.7	96%	98%	88.1%	71.7%	62.8%
VERBATIM	28.3	93%	97%	87.3%	71.9%	62.7%
MultiNLI Overall	**22.3**	**91%**	**98%**	**88.7%**	**72.2%**	**64.7%**

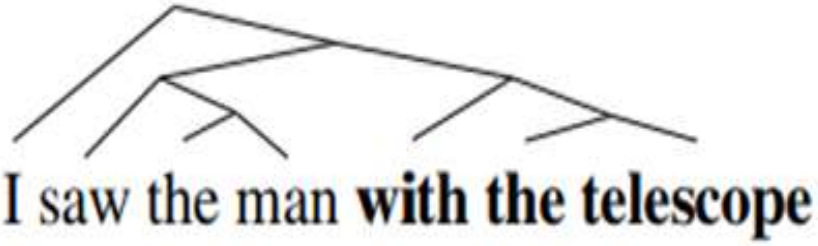
Part II

Learning Syntax from Semantics

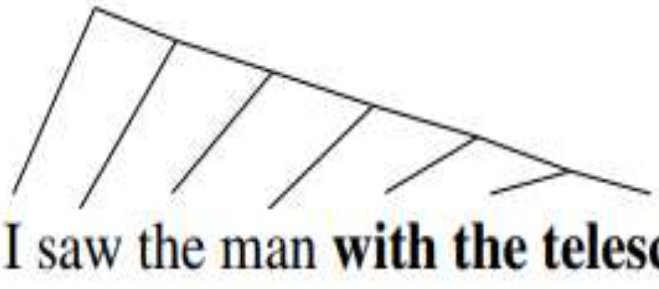
Adina Williams
Andrew Drozdov
Samuel R. Bowman

Paper Available now:
https://arxiv.org/abs/1709.01121

Syntax Matters for Interpretation

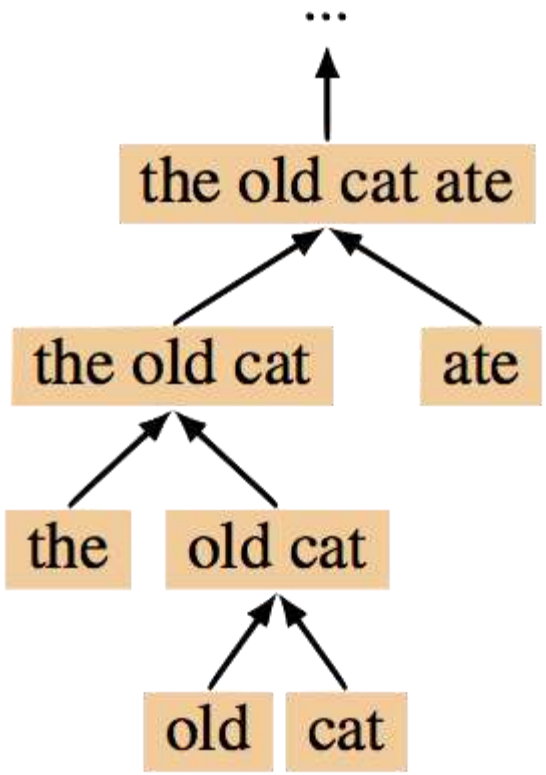


(a) I [saw the man] [with the telescope]
↪ I used the telescope to view the man



(b) I saw the [man [with the telescope]]
↪ I saw the man who had a telescope

Background: TreeRNNs



What?

- Run a (binary constituency) parser
- Use parse tree as computation graph

Why?

- *Modelling*: Some empirical advantage
- *Linguistics*: Theoretically appealing

But...

- Relies on an external parser
- Prohibitively slow

Socher et al. ‘10, Tai et al. ‘15, Bowman et al. ‘16, Liu et al. ‘17

Goal: Learn syntax from semantics

What?

Build one model that can:

- Parse sentences
- Use resulting parses in a TreeRNN text classifier

Train the full model (including the parser!) on SNLI or MultiNLI

Why?

Engineering objective:

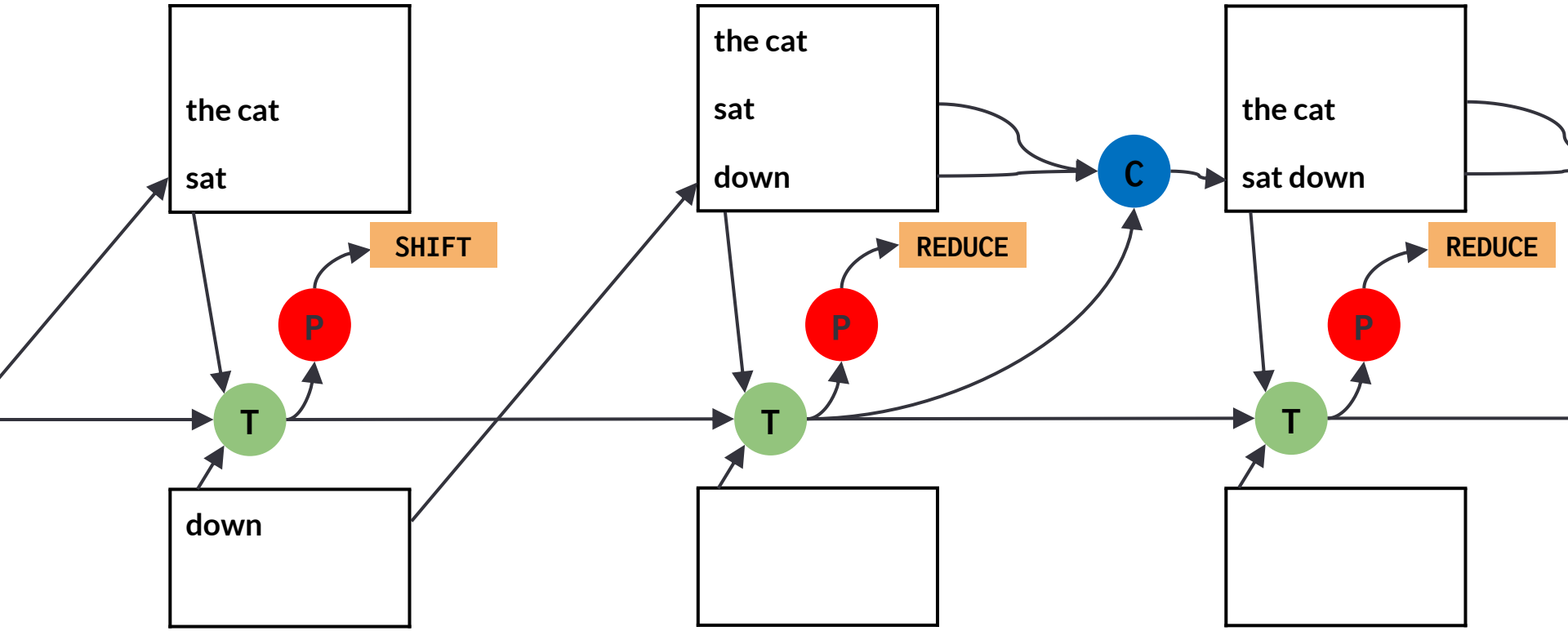
Identify *better* parsing strategies for NLU

Scientific objective:

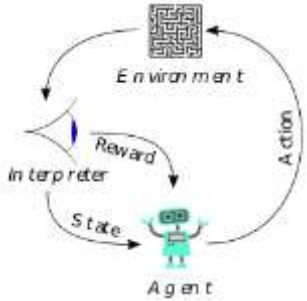
Discover what syntactic structures are both valuable and learnable.

SPINN

- Shift-reduce parser and TreeRNN share representations
- Supervised by *existing parses* at training time

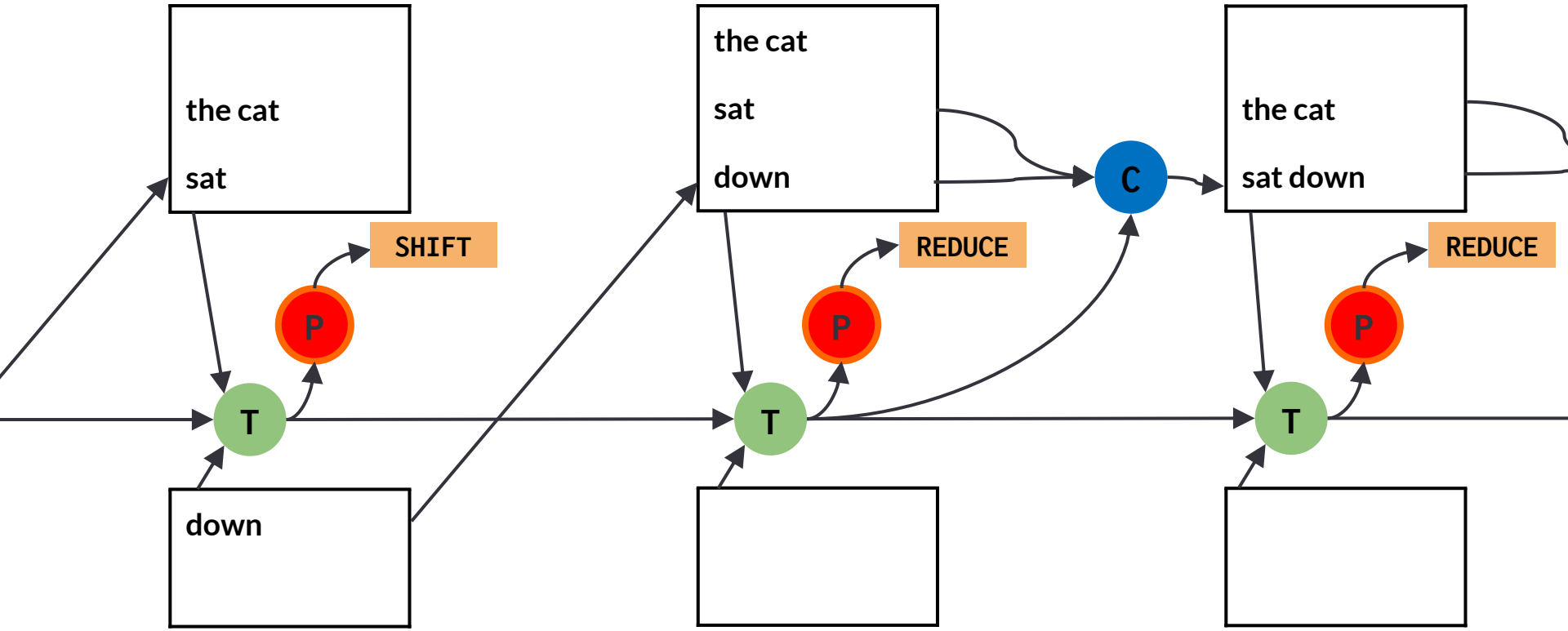


Bowman et al. ‘16



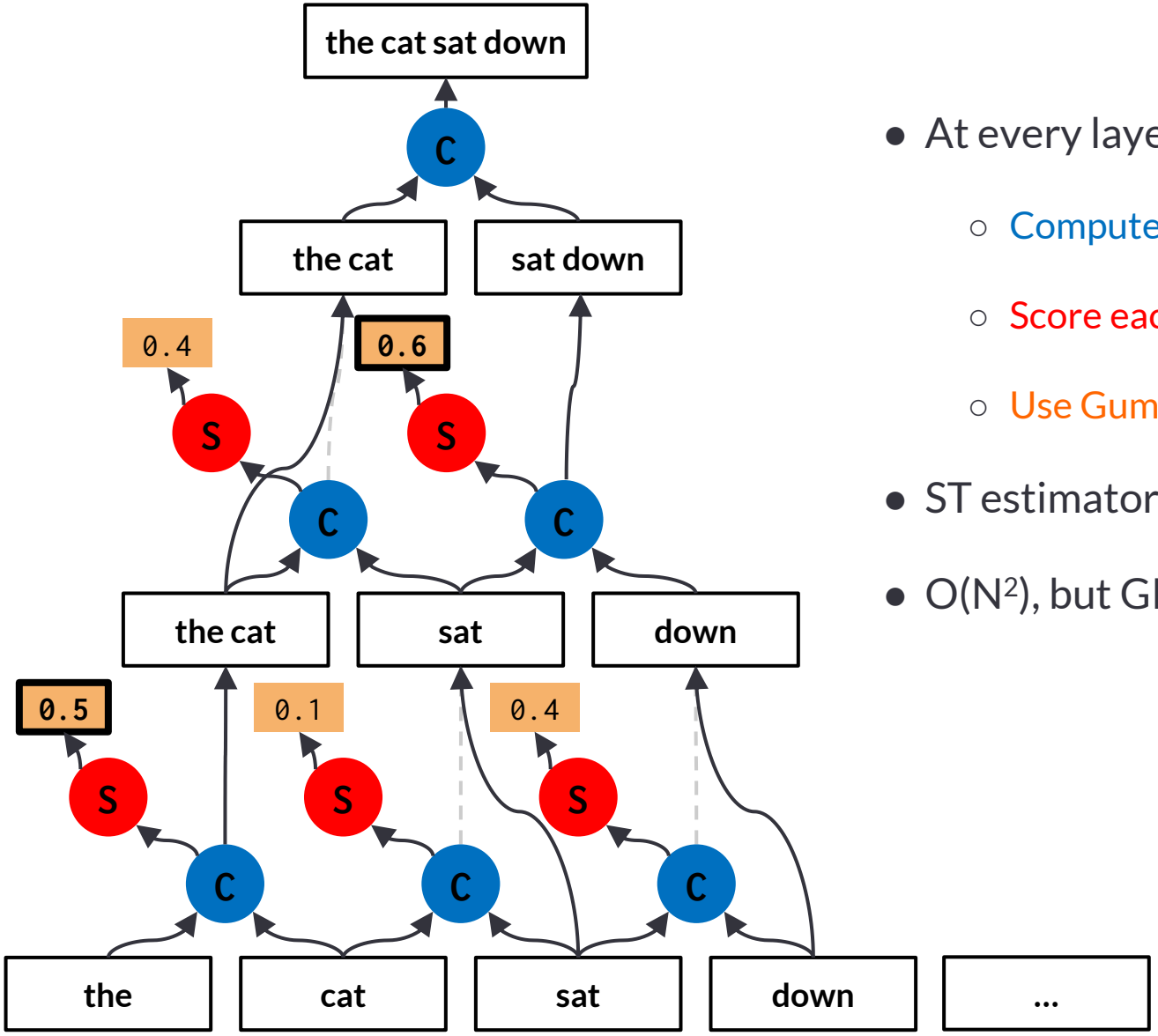
RL-SPINN

- Shift-reduce parser and TreeRNN share representations
- Parser trained using RL on NLU objective
- 100D model only
- Improvements from **latent** trees!



Yogatama et al. '17

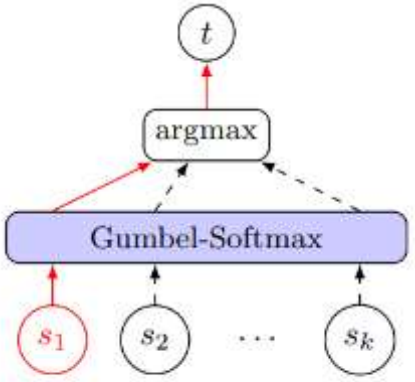
Work to date: Straight Through-Gumbel



- At every layer:
  - Compute every possible merge
  - Score each merge
  - Use Gumbel Softmax to select best
- ST estimator for gradients
- $O(N^2)$, but GPU-friendly

Choi et al. '17

Work to date: ST-Gumbel



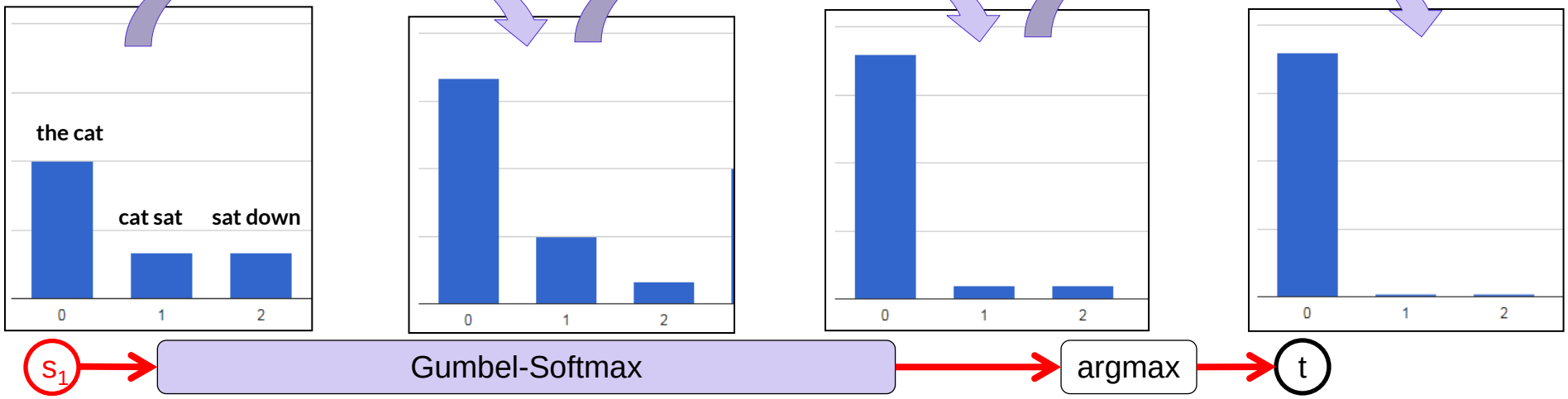
(a) Forward

Gumbel-Softmax

Softmax + Soft Sampling

Temperature

argmax



s_n : score for the n^{th} possible merge

t : the decision of what to merge together (range: [0, (n-1)])

Choi et al. '17

Work to date: ST-Gumbel

t is different for forward and backward passes!

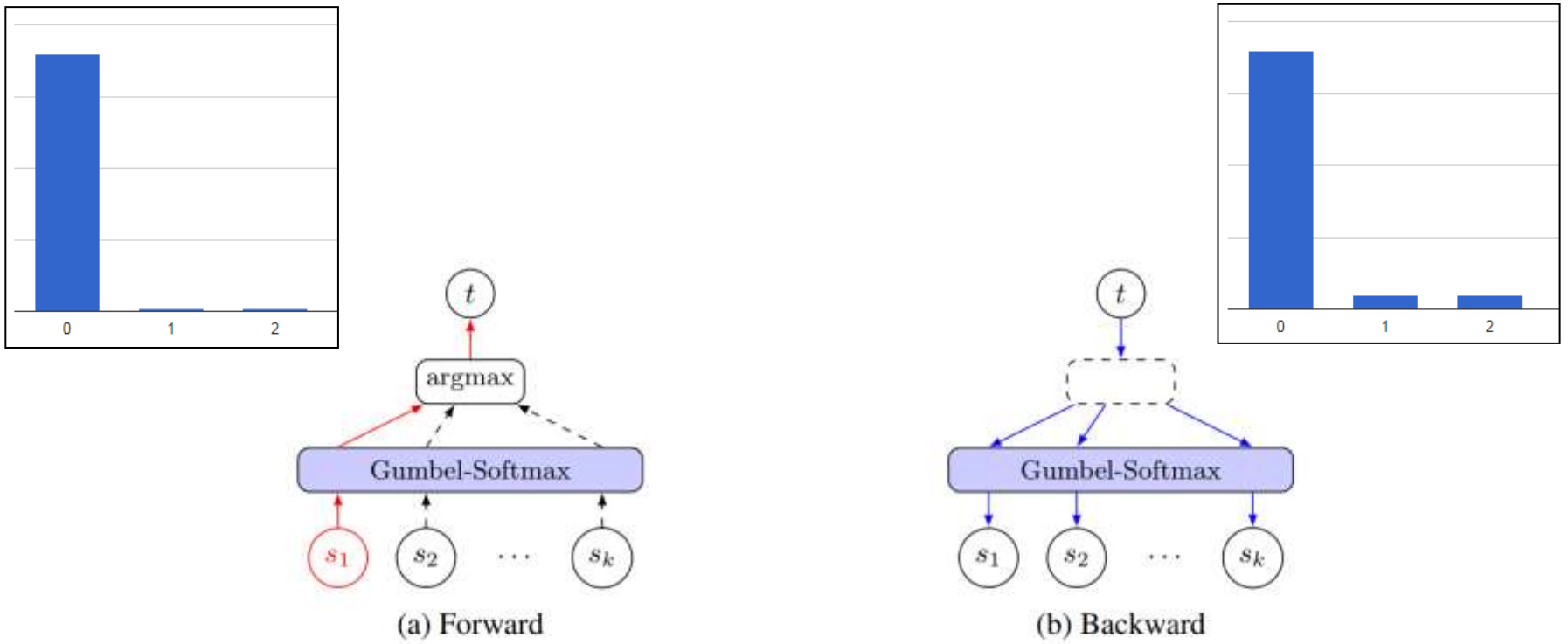


Figure 1: Visualization of forward and backward computation path of ST Gumbel-Softmax. In the forward pass, a model can maintain sparseness due to arg max operation. In the backward pass, since there is no discrete operation, the error signal can backpropagate.

Choi et al. '17

This project:
What grammar do these models learn?

Our findings: Task performance

- 300D runs on MultiNLI and SNLI, extensively tuned:
  - Absolute performance on SNLI:
    - Outperform published RL-SPINN (+1.8%)
    - Slightly underperform published ST-Gumbel (-0.9%)
  - Against our own baselines:
    - RL-SPINN **worse** with latent trees
    - ST-Gumbel **better with latent trees than parser trees**

Our findings: Consistency

Across ten random restarts, measuring F1 between runs on the MultiNLI Dev Set:

- RL-SPINN produces highly consistent trees
- ST-Gumbel produces inconsistent trees, but better than chance
- Both models see some variation in accuracy ($\sigma \cong 0.7\%$)

F1 is the harmonic mean of **precision** and **recall:**

- **Precision**: how many of selected were relevant
- **Recall**: how many of relevant were selected

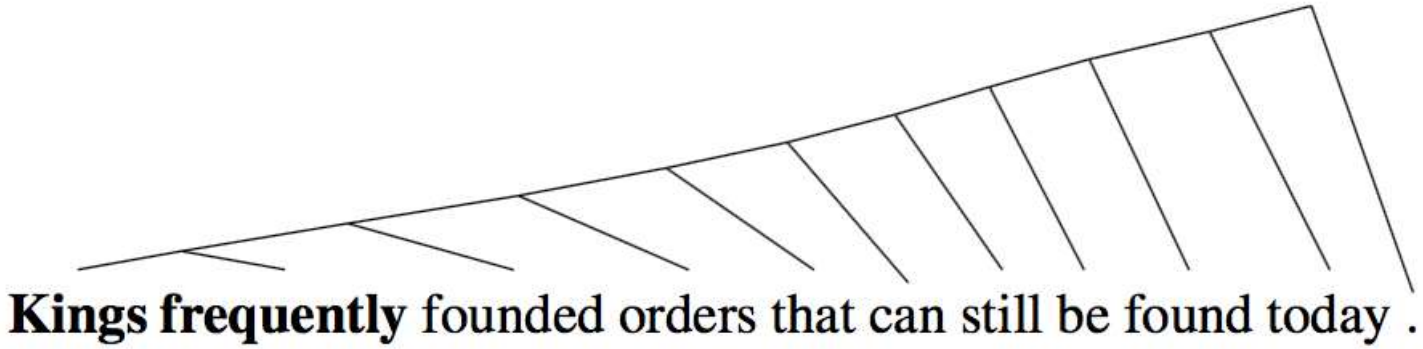
Our findings: latent vs. PTB

Unlabeled F1 score on the full Wall Street Journal section of Penn Tree Bank:

- Baseline models: 45–60%
- ST-Gumbel: 25%
- RL-SPINN: 13%
- *Random trees:* 21%

Our findings: Qualitative

- The RL-SPINN runs that perform best use *strictly left-branching* parses!
- Some runs are less strict, but variation from this trend appears random.

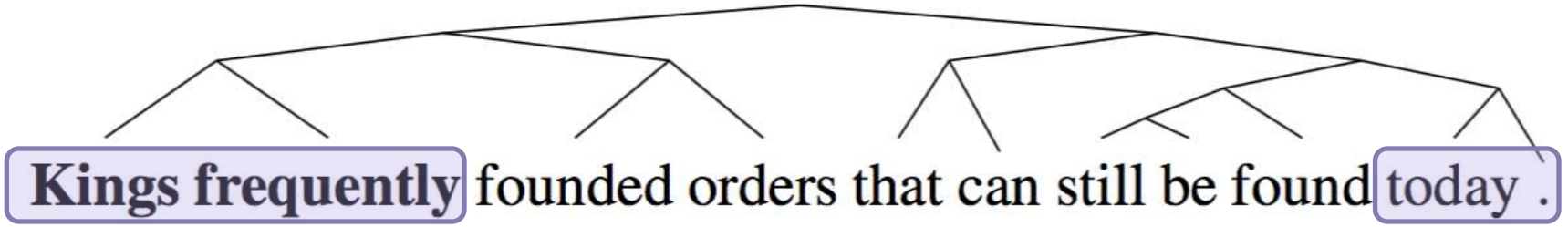


- Explains worse-than-chance parsing performance: English prefers ***right*-branching trees**.
- Model is *equivalent to* RNN, task performance shows that.

Our findings: Qualitative

ST-Gumbel parses tend to be balanced and shallow.

- The first two and last two words nearly always form constituents.



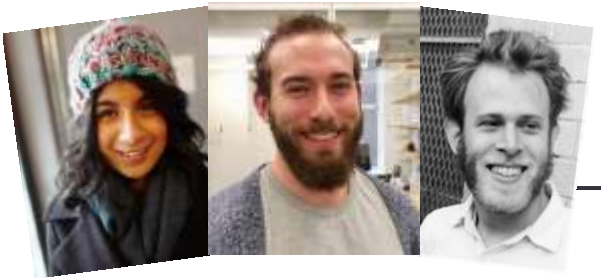
- Disappointing, but others have found these trees to work better than sequences: Munkhdalai and Yu ‘16

Closing thoughts

- MultiNLI improves on SNLI in scope and difficulty
- NLI is an ideal research task for work on modeling or learning with NN models in NLU.
- Latent tree learning is **harder than it looks**:
  Don't take performance numbers at face value!

I'm on the job market!

Let me know if you hear of anything!



Thanks!

Questions?
adinawilliams@nyu.edu

Data
nyu.edu/projects/bowman/multinli