Toward natural language semantics in learned representations

Sam Bowman
Asst. Prof. of Linguistics and Data Science, NYU

IPAM Workshop: *New Deep Learning Techniques*

Context: Deep learning in NLP

As in vision and elsewhere, deep learning techniques have yielded very fast progress on a few important data-rich tasks:

- **Reading comprehension questions**
    - Near human performance (but brittle)
- **Translation**
    - Large, perceptually obvious improvements over past systems.
- **Syntactic parsing**
    - Measurable improvements on a longstanding state of the art

The Question

Can current neural network methods learn to do anything that resembles *compositional semantics*?

The Question

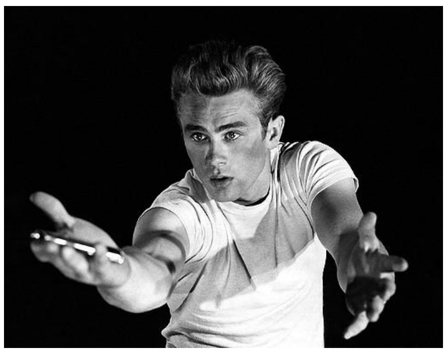
Can current neural network methods learn to do anything that resembles *compositional semantics*?

If we take this as *a goal to work toward*, what's our metric?

Proposal: 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

Example: MacCartney thesis '09

Judging Understanding with NLI

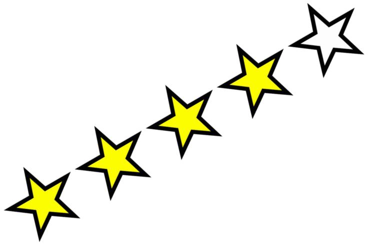
To reliably perform well at NLI, your representations of meaning must handle with the full complexity of compositional semantics:*

- Lexical entailment (*cat* vs. *animal*, *cat* vs. *dog*)
- Quantification (*all*, *most*, *fewer than eight*)
- Lexical ambiguity and scope ambiguity (*bank*, ...)
- Modality (*might*, *should*, ...)
- Common sense background knowledge

  …

* without grounding to the outside world.

Why not Other Tasks?

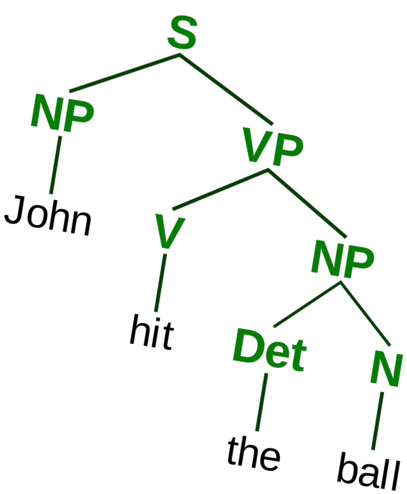


Many tasks that have been used to evaluate sentence representation models don’t require all that much language understanding:

- Sentiment analysis
- Sentence similarity

…

Why not Other Tasks?



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

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

  …

But it's the easiest of these.

Outline

- **Background:** NLI as a research task for NLU
- **Part 1** Data and early results
- **Part 2** More data, more results
- **Part 3** Next steps: Discovering structure
- **Conclusion**

Part I

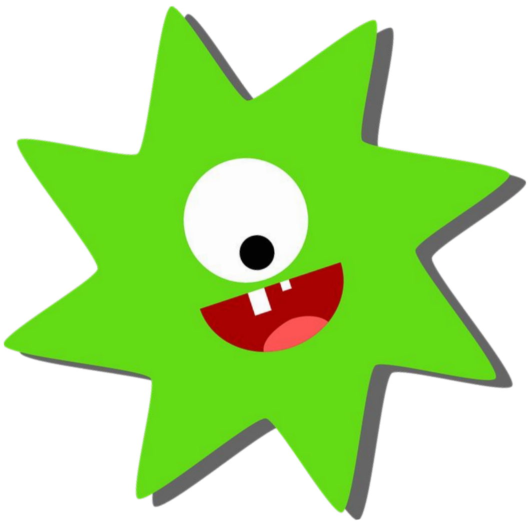
The Stanford NLI Corpus

Samuel R. Bowman
Gabor Angeli
Christopher Potts
Christopher D. Manning

Natural Language Inference Data

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

Natural Language Inference Data



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

Our data collection prompt

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL ‘14

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption. **Contradiction**

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL ‘14

What we got

Some Sample Results

Premise: *Two women are embracing while holding to go packages.*

Hypothesis: *Two woman are holding packages.*

Label: Entailment

Some Sample Results

Premise: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*

Hypothesis: *A man is repainting a garage*

Label: Neutral

Some Sample Results

Premise: *A man selling donuts to a customer during a world exhibition event held in the city of Angeles*

Hypothesis: *A woman drinks her coffee in a small cafe.*

Label: Contradiction

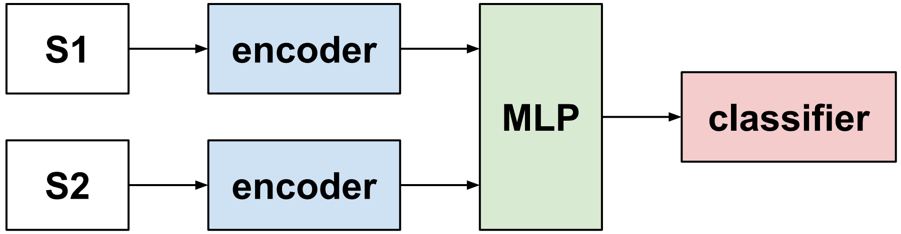
Results on SNLI

Some Results on SNLI

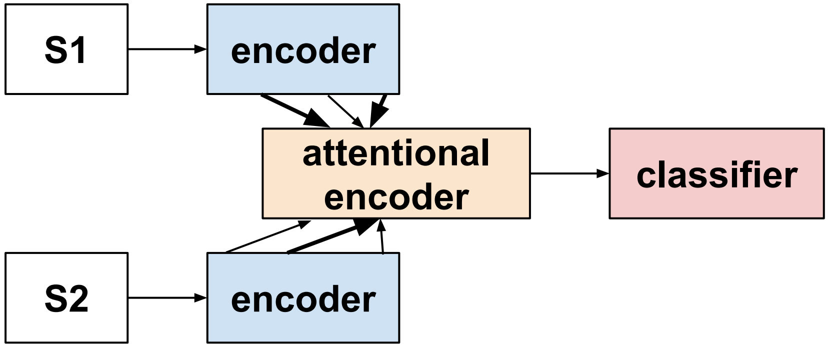
Model	Test accuracy
Most frequent class	34.2%
Big lexicalized classifier	78.2%

Two Classes of Neural Network

- Sentence encoder-based models



- Attention and memory models



Some Results on SNLI

Model	Test accuracy
Most frequent class	34.2%
Big lexicalized classifier	78.2%
300D CBOW	80.6%
300D BiLSTM	81.5%

Some Results on SNLI

Model	Test accuracy
Most frequent class	34.2%
Big lexicalized classifier	78.2%
300D CBOW	80.6%
300D BiLSTM	81.5%
REINFORCE-Trained Self-Attention (Tao Shen et al. '18)	86.3%
Self-Attention/Cross-Attention + Ensemble (Yi Tay et al. '18)	**89.3%**

Success?

- We're not at human performance yet…
- ...but with 100+ published experiments, the best systems rarely stray too far from the standard toolkit:
  - LSTMs
  - Attention
  - Pretrained word embeddings
  - Ensembling

Part II

The Multi-genre NLI Corpus

Adina Williams
Nikita Nangia
Samuel R. Bowman

SNLI is Showing its Limitations

- Little headroom left:
  - SotA: **89.3%**
  - Human performance: ~96%
- Many linguistic phenomena underattested or ignored
  - Tense
  - Beliefs
  - Modality (possibility/permission)

    ...

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

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**

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	*genre-matched evaluation*
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	*genre-mismatched evaluation*
OUP (Nonfiction Books)	0	2,000	2,000	
Verbatim (Magazine)	0	2,000	2,000	
Total	**392,702**	**20,000**	**20,000**	

What we got

Typical Dev Set Examples

Premise: *In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.*

Hypothesis: *The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.*

Label: Contradiction

Genre: Oxford University Press (Nonfiction books)

Typical Dev Set Examples

Premise: *someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny*

Hypothesis: *No one noticed and it wasn't funny at all.*

Label: Contradiction

Genre: Switchboard (Telephone Speech)

Typical Dev Set Examples

Premise: *The father can beget new offspring safe from Macbeth's hand; the son is the palpable threat.*

Hypothesis: *The son wants to kill him to marry his mom*

Label: Neutral

Genre: Verbatim (Magazine)

Key Findings

- Inter-annotator agreement measures are *identical* between SNLI and MultiNLI (within 0.5%):
    - MultiNLI is not hard for humans.
- State-of-the-art SNLI models perform around *15 percentage points worse* when re-trained and tested on MultiNLI.
    - MultiNLI *is* hard for machine learning models.

Key Figures

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

Early Results

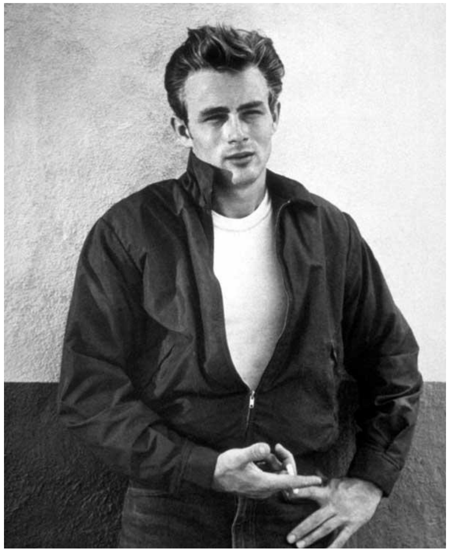
Model	Matched Test Acc.	Mismatched Test Acc.
Most frequent class	36.5%	35.6%
Deep BiLSTMs with gated skips (Chen et al. ‘17)	74.9%	74.9%
Attention+convolutions (Gong et al. ‘18)	80.0%	78.7%

NLI as a Pretraining Task

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS14
Unsupervised representation training (unordered sentences)										
Unigram-TFIDF	73.7	**79.2**	90.3	82.4	-	85.0	73.6/81.7	-	-	.58/.57
word2vec BOW	73.6	77.3	89.2	85.0	-	82.2	69.3/77.2	-	-	.58/.57
SIF	-	-	-	-	82.2	-	-	-	**84.6**	**.68**/ -
ParagraphVec (DBOW)	60.2	66.9	76.3	70.7	-	59.4	72.9/81.1	-	-	.42/.43
SDAE	74.6	78.0	**90.8**	86.9	-	78.4	**73.7**/80.7	-	-	.37/.38
GloVe BOW†	**78.7**	78.8	90.6	87.6	79.4	77.4	73.0/81.6	0.799	78.7	.46/.50
GloVe Positional Encoding†	76.3	77.4	90.4	87.1	80.6	80.8	72.5/81.2	0.789	77.9	.44/.48
BiLSTM-Max (untrained)†	77.5	**81.3**	89.6	**88.7**	80.7	**85.8**	73.2/81.6	**0.860**	83.4	.39/.48
Unsupervised representation training (ordered sentences)										
FastSent	70.8	78.4	88.7	80.6	-	76.8	72.2/80.3	-	-	**.63/.64**
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	71.2/79.1	-	-	.62/.62
SkipThought	76.5	80.1	93.6	87.1	82.0	**92.2**	**73.0/82.0**	**0.858**	82.3	.29/.35
SkipThought-LN	**79.4**	**83.1**	**93.7**	**89.3**	82.9	88.4	-	**0.858**	79.5	.44/.45
Supervised representation training										
CaptionRep (bow)	61.9	69.3	77.4	70.8	-	72.2	-	-	-	.46/.42
DictRep (bow)	76.7	78.7	90.7	87.2	-	81.0	68.4/76.8	-	-	**.67/.70**
NMT En-to-Fr	64.7	70.1	84.9	81.5	-	82.8	-	-		.43/.42
Paragram-phrase	-	-	-	-	79.7	-	-	0.849	83.1	- /**.71**
BiLSTM-Max (on SST)†	(*)	83.7	90.2	89.5	(*)	86.0	72.7/80.9	0.863	83.1	.55/.54
BiLSTM-Max (on SNLI)†	79.9	84.6	92.1	**89.8**	83.3	**88.7**	75.1/82.3	**0.885**	**86.3**	.66/.64
BiLSTM-Max (on AllNLI)†	81.1	86.3	92.4	90.2	84.6	88.2	76.2/83.1	0.884	86.3	.68/.65

Conneau et al. '17; see also Subramanian et al. '18

Discussion: NLI



- NLI lets you judge the degree to which models can learn to understand natural language sentences.
- With SNLI, it’s now possible to train low-bias machine learning models like NNs on NLI.
- MultiNLI makes it possible to test models’ ability to understand American English in nearly its full range of uses.
- Sentence encoders trained on NLI, like InferSent, are likely the best general-purpose encoders we have.

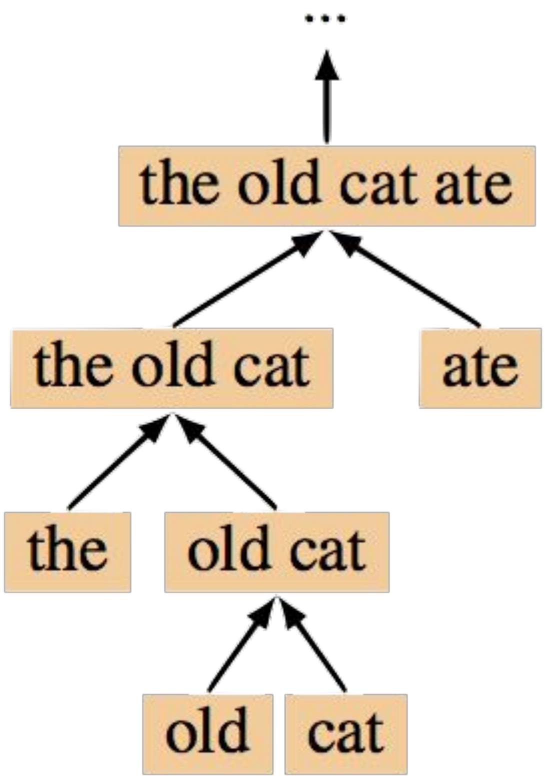
Part III

Next Steps:
Learning Syntax from Semantics

Adina Williams
Andrew Drozdov
Samuel R. Bowman

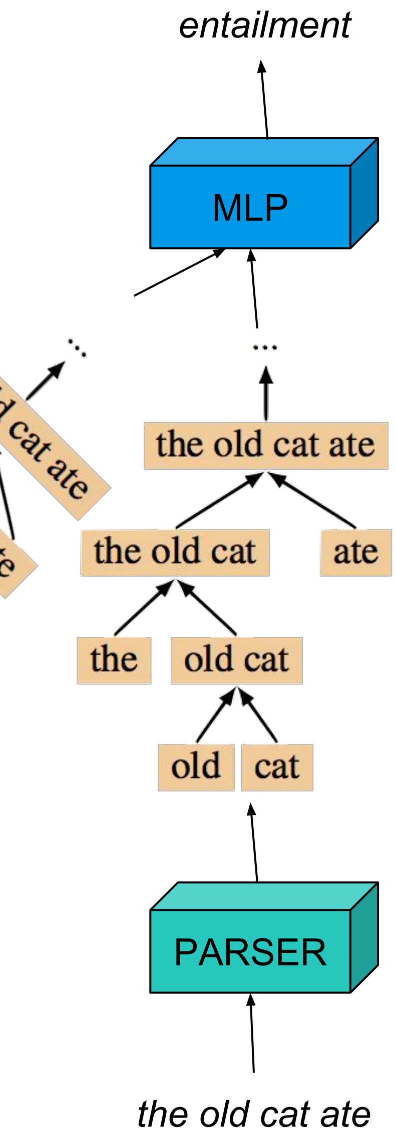
TACL 2018

Background: TreeLSTMs



TreeLSTMs replace the linear sequence of an LSTM RNN with a binary tree from a trained parser.

TreeLSTMs outperform comparable LSTM RNNs by small but consistent margins on tasks like sentiment, translation, and NLI.



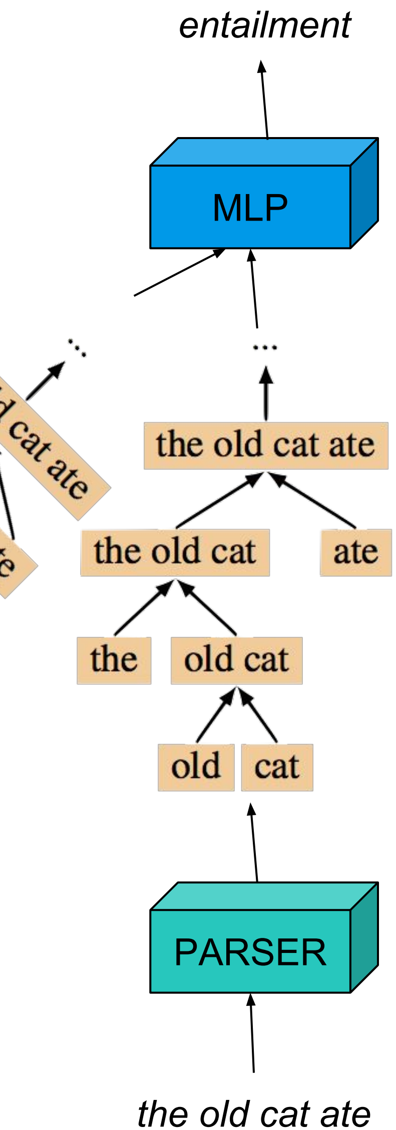
Goal: Learn syntax from semantics

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.

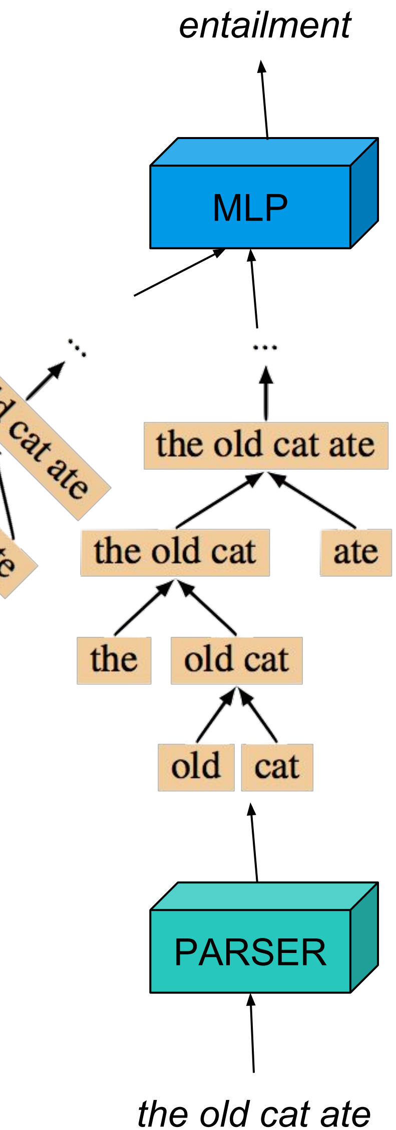


Results to Date

Three 2017 papers on SNLI report that TreeLSTMs learned trees outperform ones based on trees from an external parser:

- Yogatama et al.:
  - Shift-reduce parser + REINFORCE
- Maillard et al.:
  - Chart parser + soft gating
- Choi et al.:
  - Novel parser + Straight through Gumbel softmax

Limited analysis of the resulting parses so far.

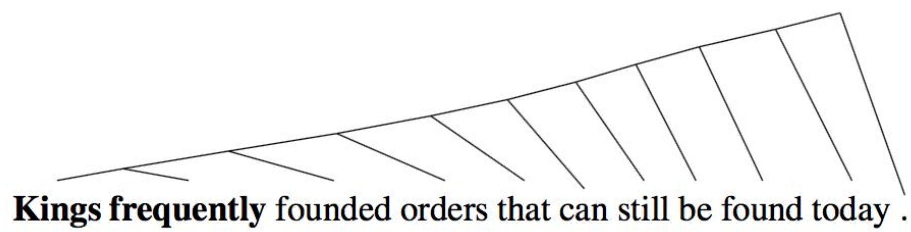


Our Findings

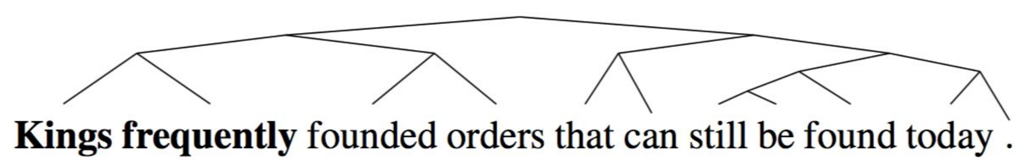
We reproduced the numeric results for the best two of these.

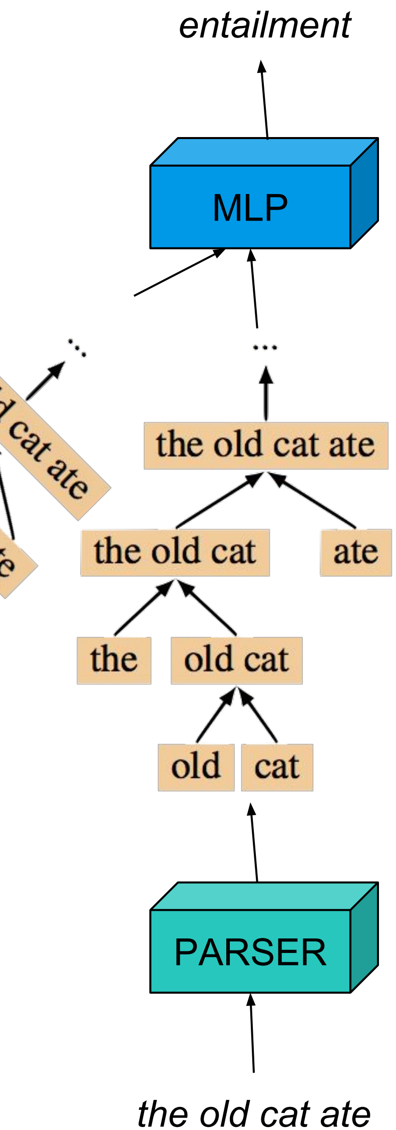
If thoroughly tuned for dev set performance, both learn:

- Trivial left- or right-branching trees (RNN-equivalent)



- ...or trivial balanced trees.





Our Findings

No categorical successes yet.

Open problems:

- The performance gain from discovering correct trees is small, and therefore difficult to optimize for with current tools. Could better models improve this?
- How do we explore possible parsing strategies when it may take many gradient updates to the rest of the model to know if any strategy helps?

Thanks!

Questions, code, & data:

nyu.edu/projects/bowman

Plus:

- **Adina Williams** is on the job market in cognitive science!
- **Nikita Nangia** and **Andrew Drozdov** are applying to PhD programs in NLP!