How much linguistics is needed for NLP?

Ed Grefenstette
etg@google.com

Based on work with: **Karl Moritz Hermann, Phil Blunsom, Tim Rocktäschel, Tomáš Kočiský, Lasse Espeholt, Will Kay, and Mustafa Suleyman**

An Identity Crisis in NLP?

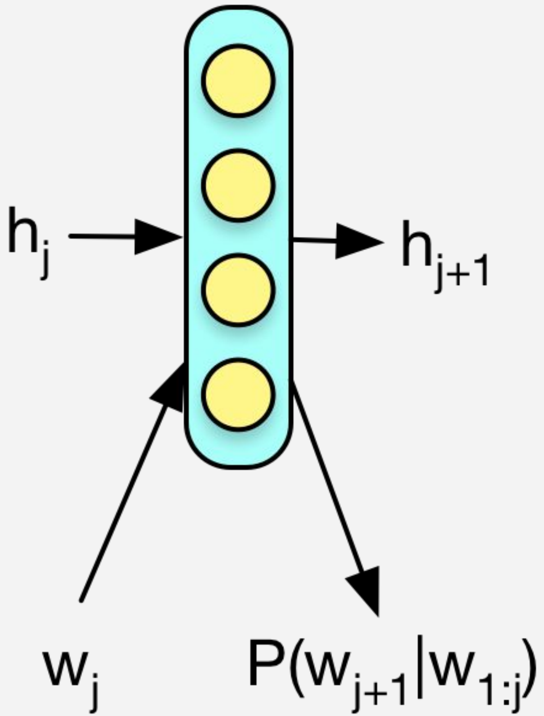
yoav goldberg
@yoavgo

This new wave of "we solved all of language with neural-nets" papers is not a step forward but back to naive 1960s-style Eliza-like work.

Today's Topics

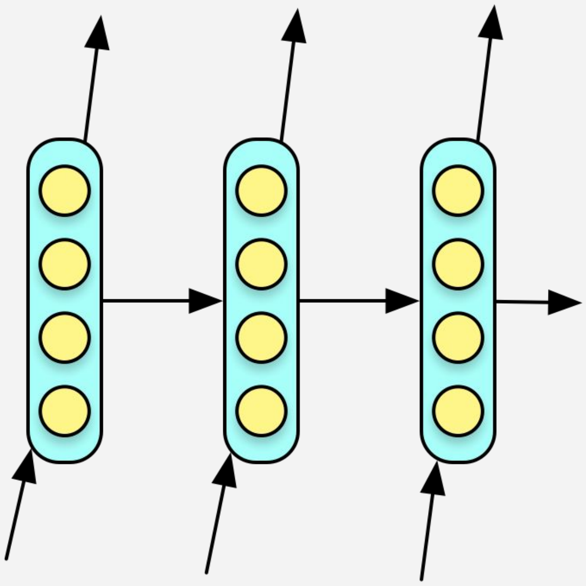
1. **Sequence-to-Sequence Modelling with RNNs**
2. Transduction with Unbounded Neural Memory
3. Machine Reading with Attention
4. Recognising Entailment with Attention

Some Preliminaries: RNNs



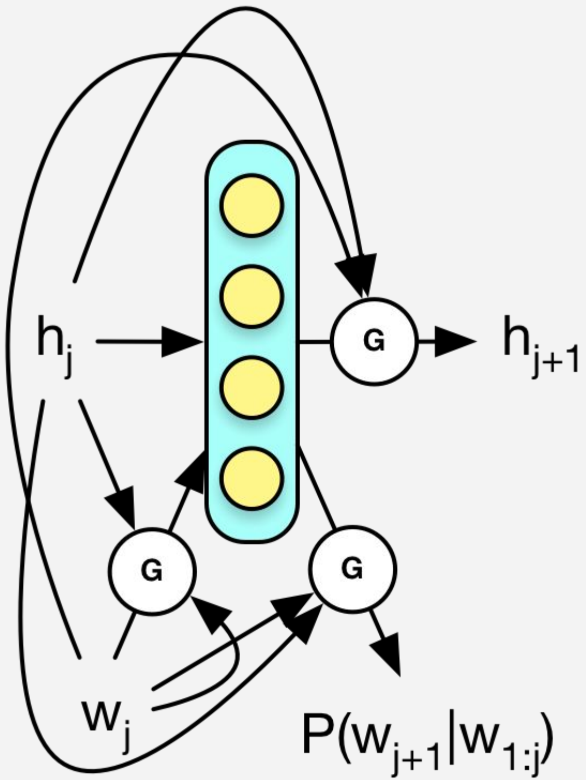
- Recurrent hidden layer outputs distribution over next symbol
- Connects "back to itself"
- Conceptually: hidden layer models history of the sequence.

Some Preliminaries: RNNs



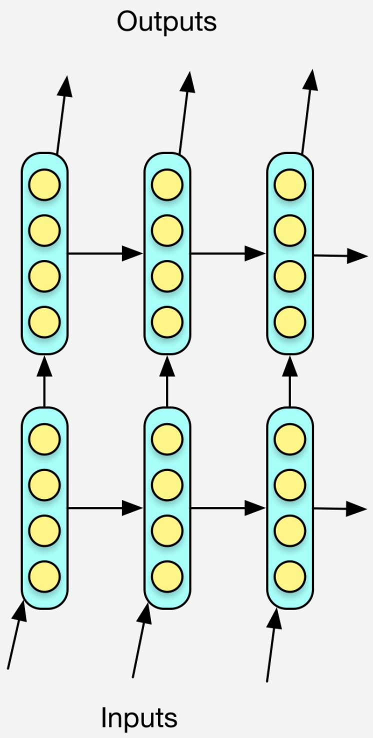
- RNNs fit variable width problems well
- Unfold to feedforward nets with shared weights
- Can capture long range dependencies
- Hard to train (exploding / vanishing gradients)

Some Preliminaries: LSTM RNNs



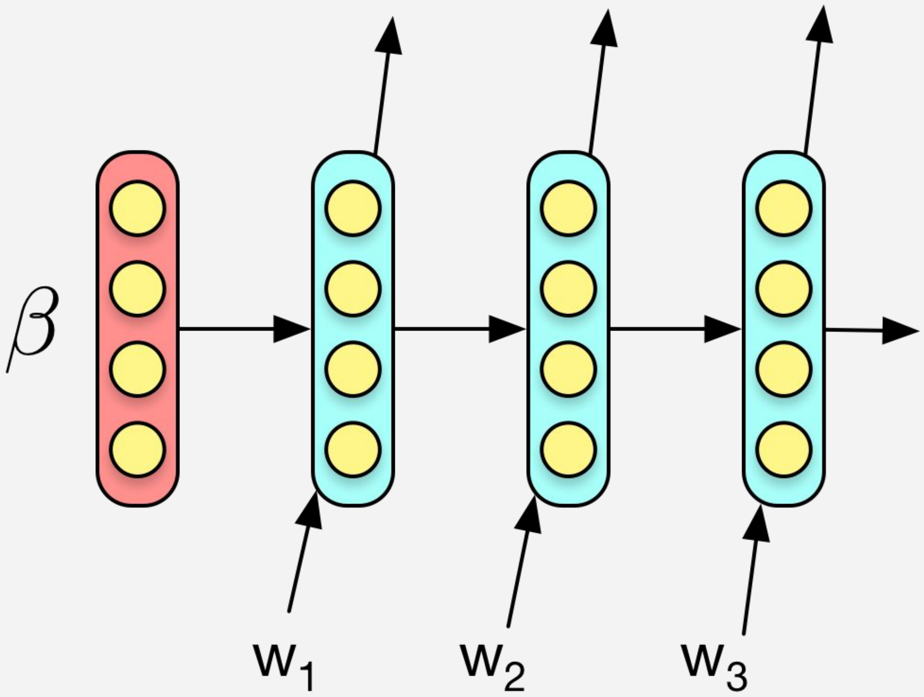
Network state determines when information is read in/out of cell, and when cell is emptied.

Some Preliminaries: Deep RNNs

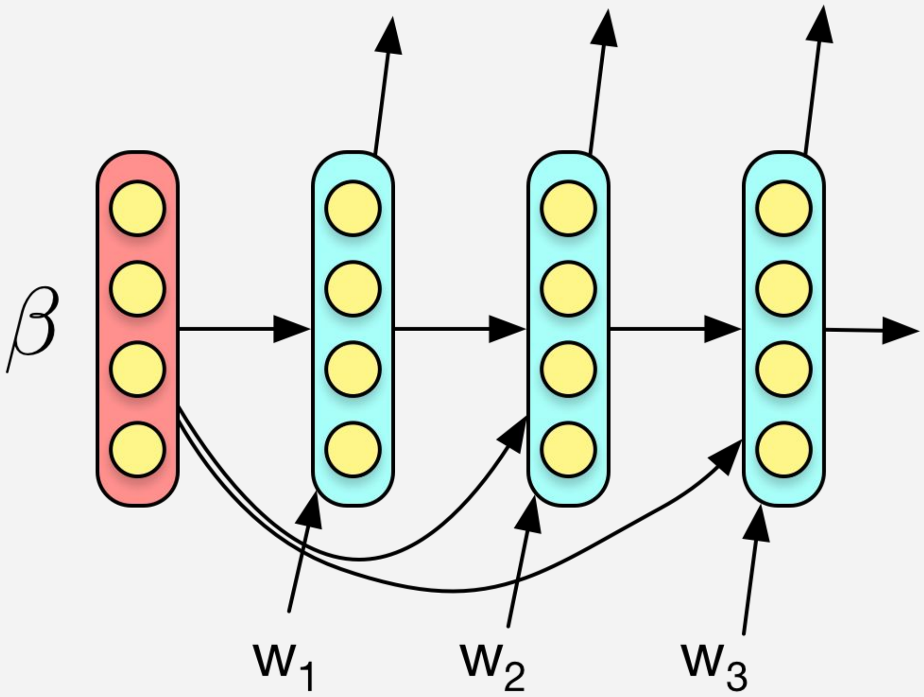


- RNNs can be layered: output of lower layers is input to higher layers
- Different interpretations: higher-order patterns, memory
- Generally needed for harder problems

Conditional Generation



Conditional Generation



Transduction and RNNs

Many NLP (and other!) tasks are castable as transduction problems. E.g.:

Translation: English to French transduction

Parsing: String to tree transduction

Computation: Input data to output data transduction

Transduction and RNNs

Generally, goal is to transform some source sequence

$$S = s_1 s_2 \dots s_m$$

into some target sequence

$$T = t_1 t_2 \dots t_n$$

Transduction and RNNs

Approach:

1. Model $P(t_{i+1}|t_1 \ldots t_n; S)$ with an RNN
2. Read in source sequences
3. Generate target sequences (greedily, beam search, etc).

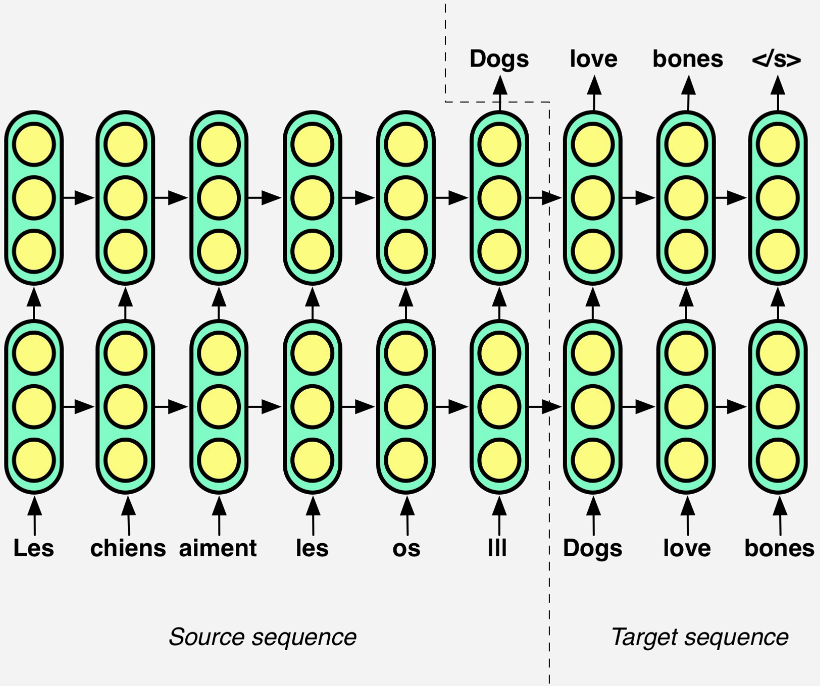
Encoder-Decoder Model

- Concatenate source and target sequences into joint sequences:

$$s_1\ s_2\ \ldots\ s_m\ |||\ t_1\ t_2\ \ldots\ t_n$$

- Train a single RNN over joint sequences
- Ignore RNN output until separator symbol (e.g. "|||")
- Jointly learn to compose source and generate target sequences

Deep LSTMs for Translation



(Sutskever *et al.* NIPS 2014)

Learning to Execute

Task (Zaremba and Sutskever, 2014):

- Read simple python scripts character-by-character
- Output numerical result character-by-character.

Input:

```
j=8584
for x in range(8):
  j+=920
b=(1500+j)
print((b+7567))
```

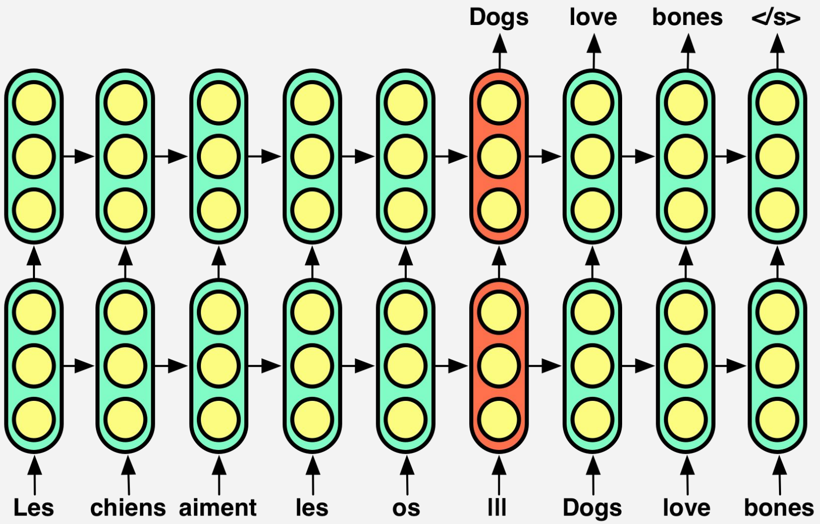
Target: 25011.

Input:

```
i=8827
c=(i-5347)
print((c+8704) if 2641<8500 else 5308)
```

Target: 12184.

The Transduction Bottleneck



Today's Topics

1. Sequence-to-Sequence Modelling with RNNs
2. **Transduction with Unbounded Neural Memory**
3. Machine Reading with Attention
4. Recognising Entailment with Attention

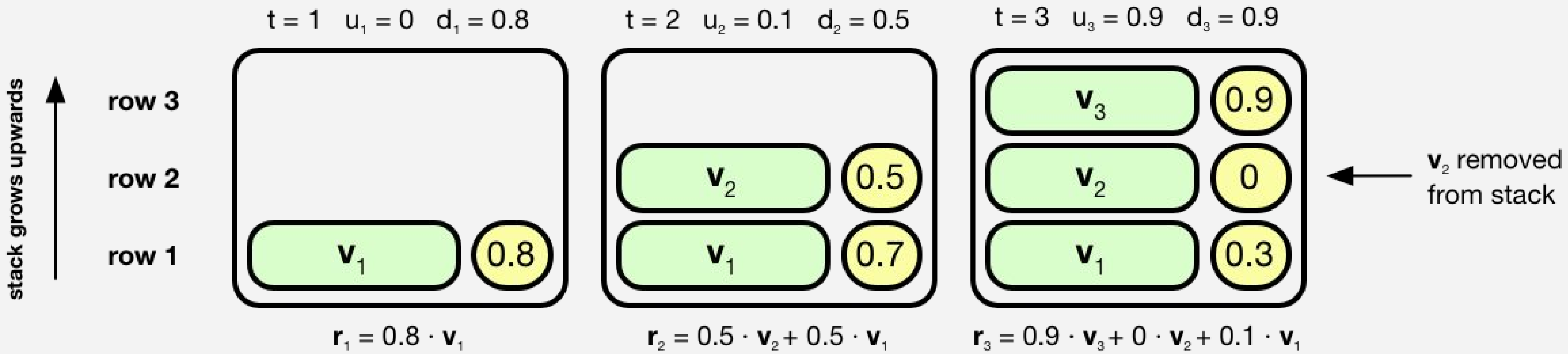
Solution: Unbounded Neural Memory

We introduce memory modules that act like Stacks/Queues/DeQues:

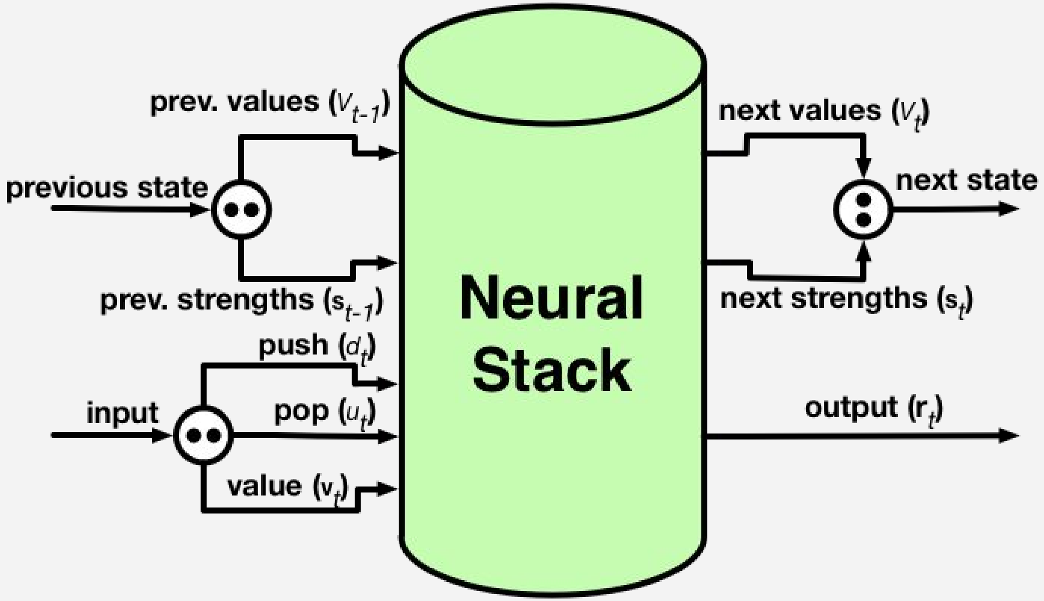
- Memory "size" grows/shrinks dynamically
- Continuous push/pop not affected by number of objects stored
- Can capture unboundedly long range dependencies*
- Propagates gradient flawlessly*

(* if operated correctly: see paper's appendix)

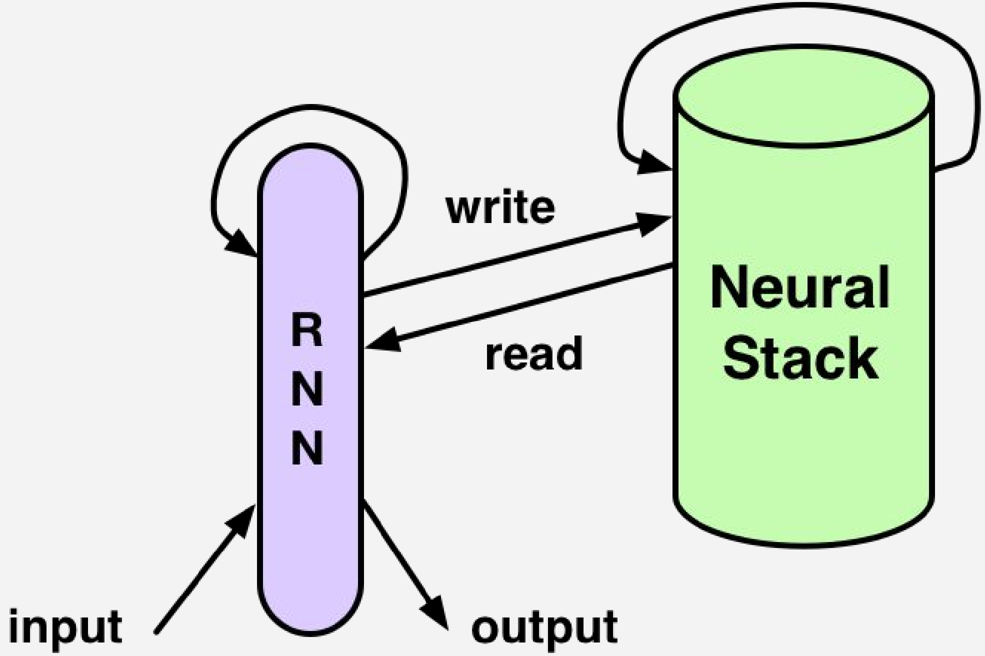
Example: A Continuous Stack



Example: A Continuous Stack



Controlling a Neural Stack



Synthetic Transduction Tasks

Copy

$$a_1a_2a_3 \ldots a_n \rightarrow a_1a_2a_3 \ldots a_n$$

Reversal

$$a_1a_2a_3 \ldots a_n \rightarrow a_n \ldots a_3a_2a_1$$

Bigram Flipping

$$a_1a_2a_3a_4 \ldots a_{n-1}a_n \rightarrow a_2a_1a_4a_3 \ldots a_na_{n-1}$$

Synthetic ITG Transduction Tasks

Subject-Verb-Object to Subject-Object-Verb Reordering

si1 vi28 oi5 oi7 si15 rpi si19 vi16 oi10 oi24 → so1 oo5 oo7 so15 rpo so19 vo16 oo10 oo24 vo28

Genderless to Gendered Grammar

we11 the en19 and the em17 → wg11 das gn19 und der gm17

Coarse- and Fine-Grained Accuracy

- **Coarse-grained accuracy**

  Proportion of entirely correctly predicted sequences in test set

- **Fine-grained accuracy**

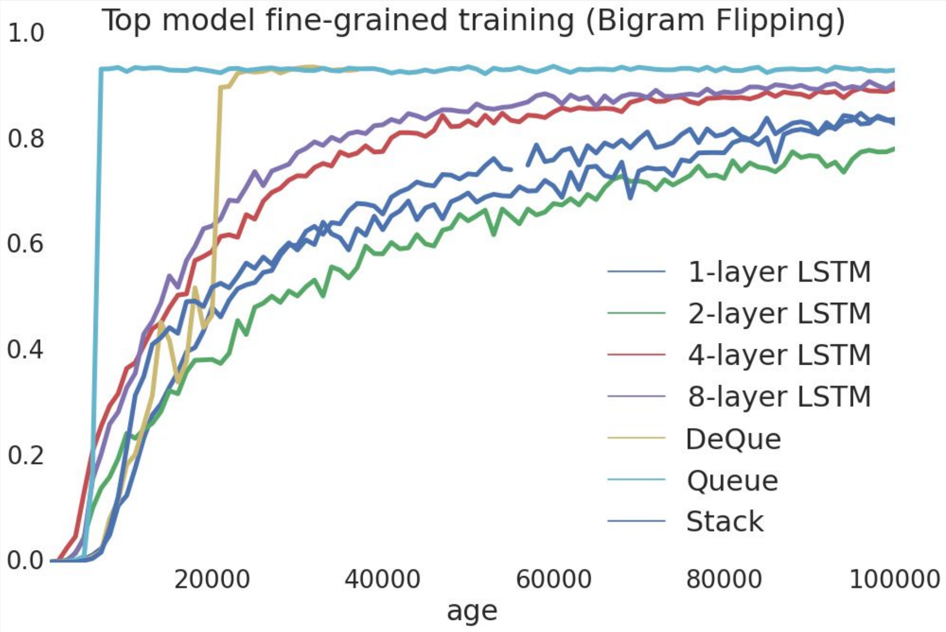
  Average proportion of sequence correctly predicted before first error

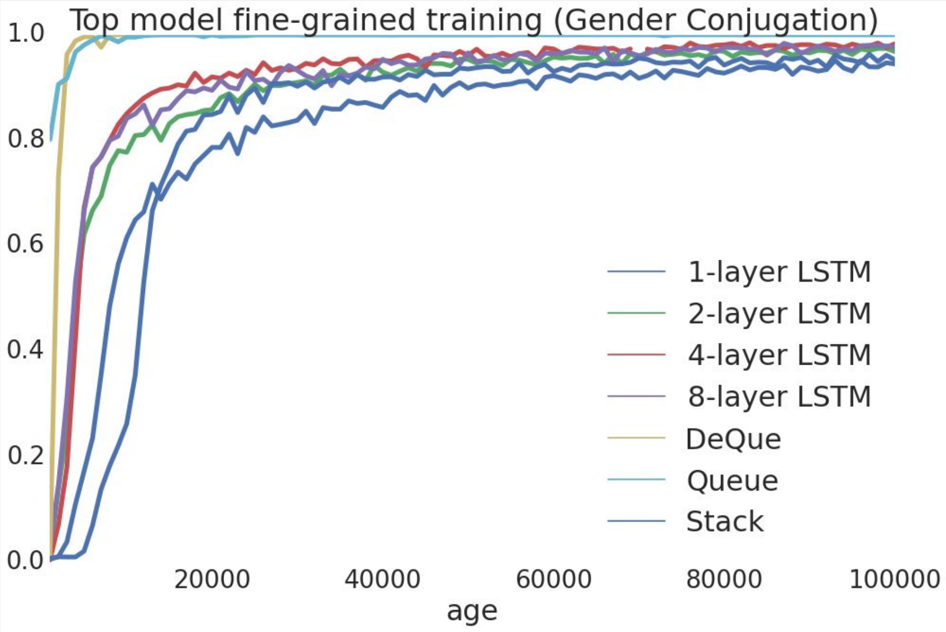
Results

Experiment	Stack	Queue	DeQue	Deep LSTM
Copy	Poor	**Solved**	**Solved**	Poor
Reversal	**Solved**	Poor	**Solved**	Poor
Bigram Flip	Converges	**Best Results**	**Best Results**	Converges
SVO-SOV	**Solved**	**Solved**	**Solved**	Converges
Conjugation	Converges	**Solved**	**Solved**	Converges

Every Neural Stack/Queue/DeQue that solves a problem preserves the solution for longer sequences (tested up to 2x length of training sequences).

Rapid Convergence





Today's Topics

1. Sequence-to-Sequence Modelling with RNNs
2. Transduction with Unbounded Neural Memory
3. **Machine Reading with Attention**
4. Recognising Entailment with Attention

Natural Language Understanding

1. Read text
2. Synthesise its information
3. Reason on basis of that information
4. Answer questions based on steps 1–3

We want to build models that can read text and answer questions based on them!

So far we are very good at step 1!

For the other three steps we first need to solve the data bottleneck

Data (I) – Microsoft MCTest Corpus

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back. One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home. …

Where did James go after he went to the grocery store?

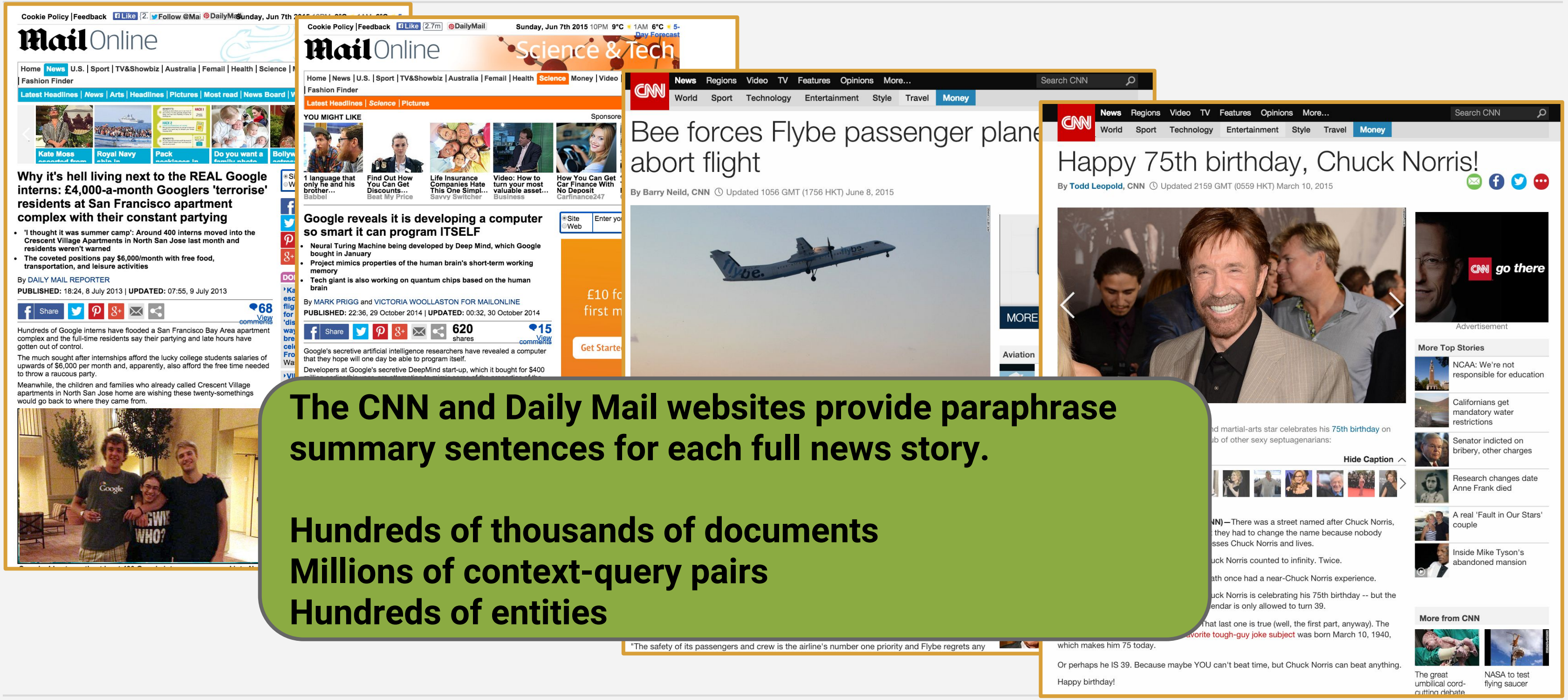
1. his deck
2. his freezer
3. a fast food restaurant
4. his room

Data (II) – Facebook Synthetic Data

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.

Query: Where was the apple before the kitchen?
Answer: office

A new source for Reading Comprehension data



The CNN and Daily Mail websites provide paraphrase summary sentences for each full news story.

Hundreds of thousands of documents
Millions of context-query pairs
Hundreds of entities

Large-scale Supervised Reading Comprehension

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …

Cloze-style question:

Query: Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer: Oisin Tymon

One catch: Avoid the Language Model trap

From the Daily Mail:

- The hi-tech bra that helps you beat breast **X**
- Could Saccharin help beat **X** ?
- Can fish oils help fight prostate **X** ?

Any n-gram language model train on the Daily Mail would correctly predict (**X** = cancer)

Anonymisation and permutation

Carefully designed problem to avoid shortcuts such as QA by LM:
⇛ We only solve this task if we solve it in the most general way possible:

The easy way ...

(CNN) New Zealand are on course for a first ever World Cup title after a thrilling semifinal victory over South Africa, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at Eden Park, Grant Elliott hit a six right at the death to confirm victory and send the Auckland crowd into raptures. It is the first time they have ever reached a world cup final.

Question:
_____ reach cricket Word Cup final?

Answer:
New Zealand

... our way

(*ent23*) *ent7* are on course for a first ever *ent15* title after a thrilling semifinal victory over *ent34*, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at *ent12*, *ent17* hit a six right at the death to confirm victory and send the *ent83* crowd into raptures. It is the first time they have ever reached a *ent15* final.

Question:
_____ reach *ent3* *ent15* final?

Answer:
ent7

Get the data now!

www.github.com/deepmind/rc-data

or follow "**Further Details**" link under the paper's entry on

www.deepmind.com/publications

Baseline Model Results

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	26.3	27.9	22.5	22.7
Exclusive frequency	30.8	32.6	27.3	27.7
Frame-semantic model	32.2	33.0	30.7	31.1
Word distance model	46.2	46.9	55.6	54.8

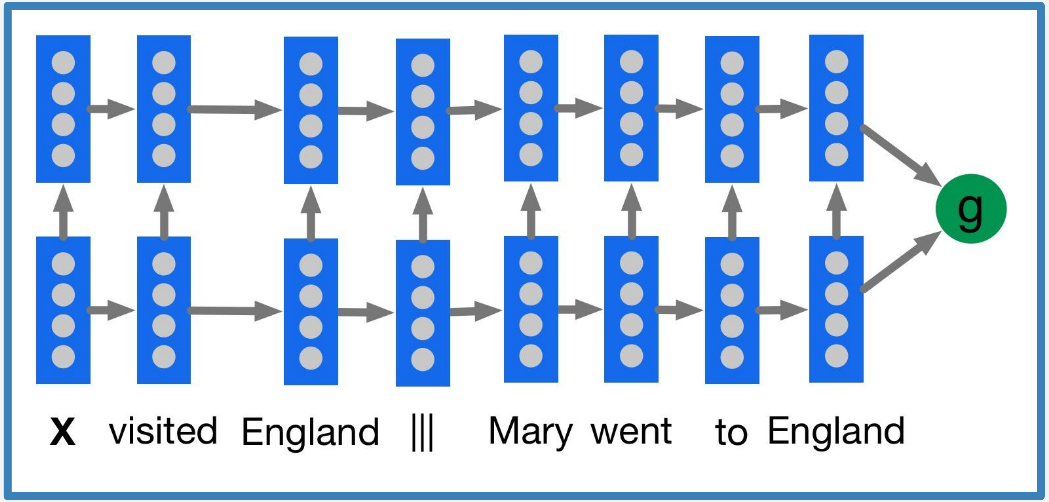
Neural Machine Reading

We estimate the probability of word type *a* from document d answering query *q*:

$$p(a|d,q) \propto \exp\left(W(a)g(d,q)\right), \quad \text{s.t. } a \in d.$$

where *W*(*a*) indexes row *a* of *W* and *g*(*d,q*) embeds of a document and query pair.

The Deep LSTM Reader

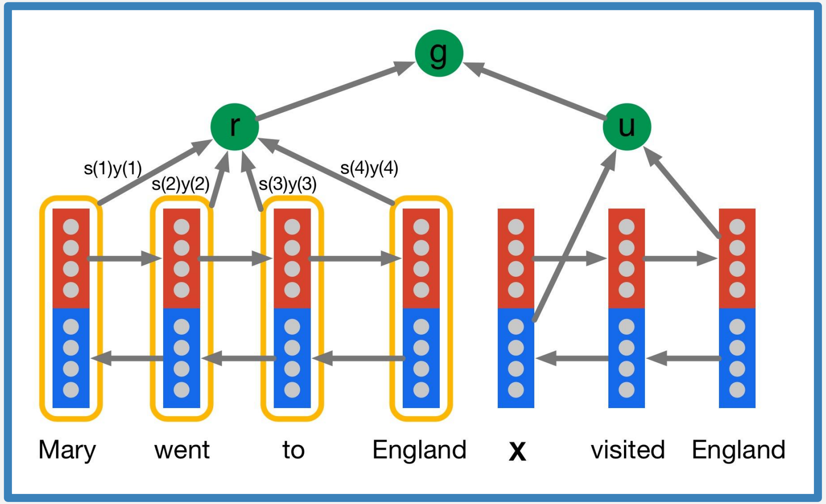


Achtung!

We can improve on this using an attention model over a bidirectional LSTM

- Separate encodings for query and context tokens
- Attend over context token encodings
- Predict based on joint weighted attention and query representation

The Attentive Reader

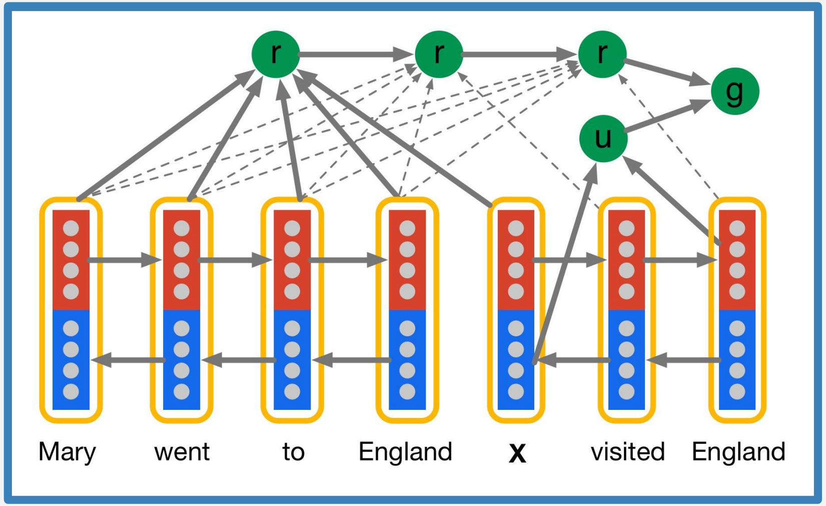


Impatience can be a virtue

We developed a nice iterative extension to the Attentive Reader as follows

- Read query word by word
- Attend over document at each step through query
- Iteratively combine attention distribution
- Predict answer with increased accuracy

The Impatient Reader



Impatience is a virtue - Results

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	26.3	27.9	22.5	22.7
Exclusive frequency	30.8	32.6	27.3	27.7
Frame-semantic model	32.2	33.0	30.7	31.1
Word distance model	46.2	46.9	55.6	54.8
Deep LSTM Reader	49.0	49.9	57.1	57.3
Uniform attention	31.1	33.6	31.0	31.7
Attentive Reader	56.5	58.9	64.5	63.7
Impatient Reader	**57.0**	**60.6**	**64.8**	**63.9**

The Attentive Reader - Correct Example

by *ent40* , *ent62* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent62*) a *ent88* was killed in a parachute accident in *ent87* , *ent28* , near *ent66* , a *ent47* official told *ent62* on wednesday . he was identified thursday as special warfare operator 3rd class *ent49* , 29 , of *ent44* , *ent13* . `` *ent49* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused commitment for posterity , '' the *ent47* said in a news release . *ent49* joined the seals in september after enlisting in the *ent47* two years earlier . he was married , the *ent47* said . initial indications are the parachute failed to open during a jump as part of a training exercise . *ent49* was part of a *ent57* - based *ent88* team .

ent47 identifies deceased sailor as **X** , who leaves behind a wife

Correct prediction (***ent49***) - Requires anaphora resolution

The Attentive Reader - Failed Prediction

by *ent37* , *ent61* updated 11:44 am et , tue march 10 , 2015 (*ent61*) a suicide attacker detonated a car bomb near a police vehicle in the capital of southern *ent12* 's *ent24* on tuesday , killing seven people and injuring 23 others , the province 's deputy governor said . the attack happened at about 6 p.m. in the *ent27* area of *ent2* city , said *ent66* , deputy governor of *ent24* . several children were among the wounded , and the majority of casualties were civilians , *ent66* said . details about the attacker 's identity and motive were n't immediately available .

car bomb detonated near police vehicle in **X** , deputy governor says

Correct entity ***ent2***, predicted ***ent24*** - Geographic ambiguity

Daily Mail News

Not so fast: Police find $200,000 Lamborghini with no license plates abandoned on Texas highway

- A yellow Lamborghini was discovered on the southbound side of the Dallas North Tollway
- Whoever was behind the wheel seemed to have ditched the car after slamming into a highway barrier
- The expensive vehicle was taken to a Dallas police impound lot

Authorities reportedly discovered a Lamborghini discarded on a Texas highway over the weekend. The vehicle was discovered on the southbound side of the Dallas North Tollway, local media reported.

Whoever was behind the wheel seemed to have ditched the car after slamming into a highway barrier, WFAA reported.

WFAA reported the Lamborghini did not contain any "identifying information" inside. The expensive vehicle was taken to a Dallas police impound lot, the television station reported.

Lamborghinis generally retail for hundreds of thousands of dollars. The Dallas Police Department did not immediately return a message seeking comment.

A yellow ____________ was discovered on the southbound side of the Dallas North Tollway


CNN
News Regions Video TV Features Opinions More...
'Most Interesting Man' cutout doesn't pass in HOV lane
Story highlights
A driver was caught in the x with a cutout of "Most Interesting Man"

Today's Topics

1. Sequence-to-Sequence Modelling with RNNs
2. Transduction with Unbounded Neural Memory
3. Machine Reading with Attention
4. **Recognising Entailment with Attention**

Recognizing Textual Entailment (RTE)

A man is crowd surfing at a concert

- The man is at a football game	**Contradiction**
- The man is drunk	**Neutral**
- The man is at a concert	**Entailment**

A wedding party is taking pictures

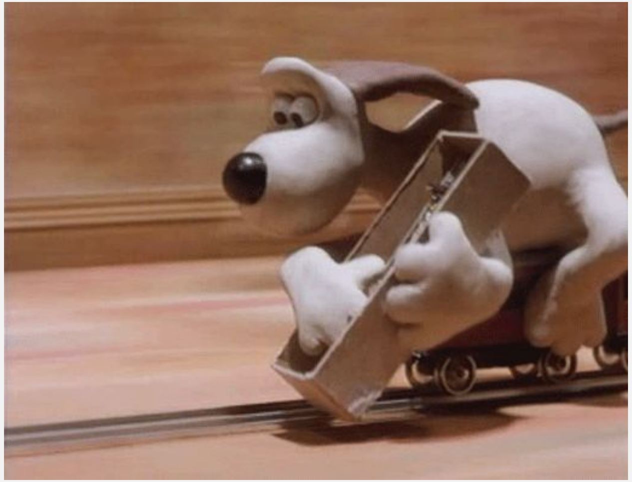
- There is a funeral	**Contradiction**
- They are outside	**Neutral**
- Someone got married	**Entailment**

Stanford Natural Language Inference Corpus

Project on RTE while working with SICK corpus (Marelli et al., SemEval 2014)

10k sentence pairs, partly synthetic

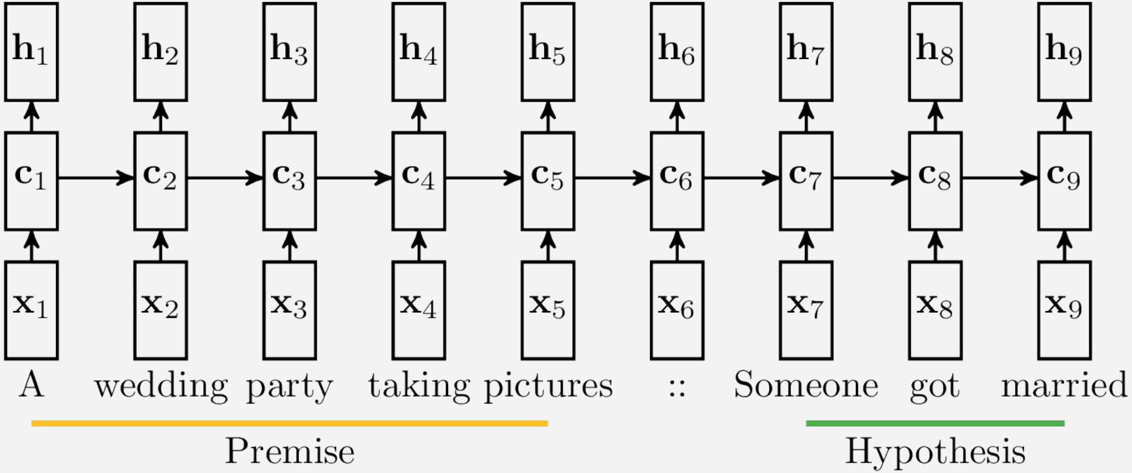
The last 1.5 months of Tim's internship, with the SNLI corpus (Bowman et al., EMNLP 2015)



570k sentence pairs from Mechanical Turkers

EMNLP 2015 “best data set or resource” award!

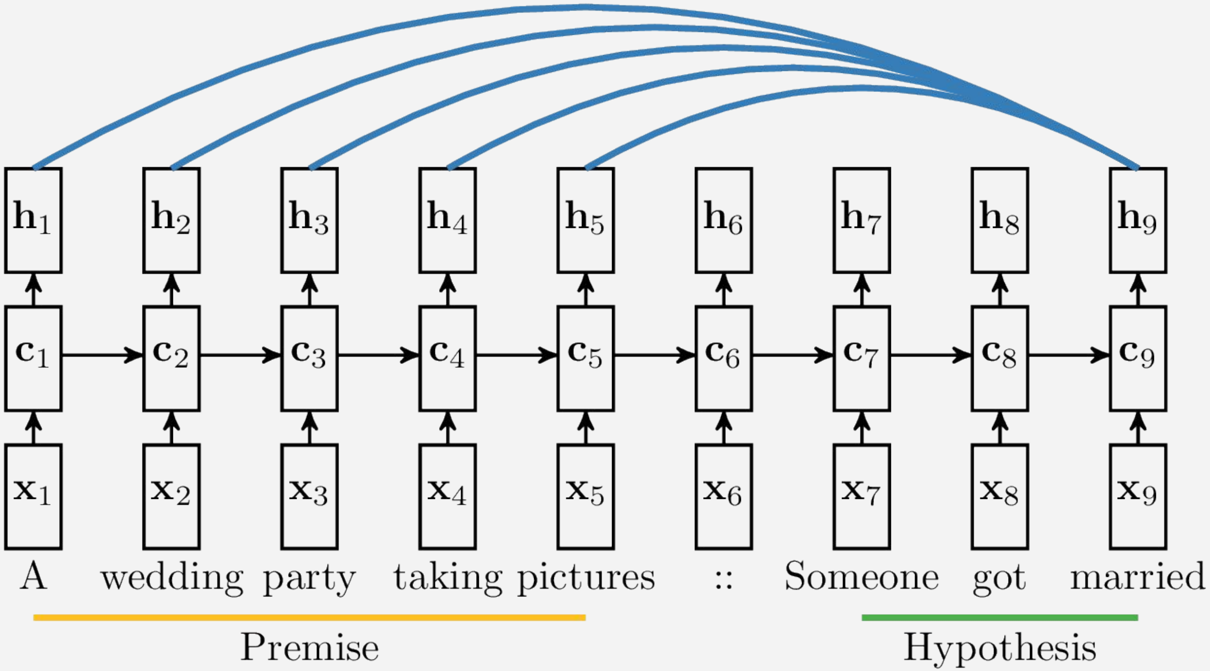
Model



$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{bmatrix}$$

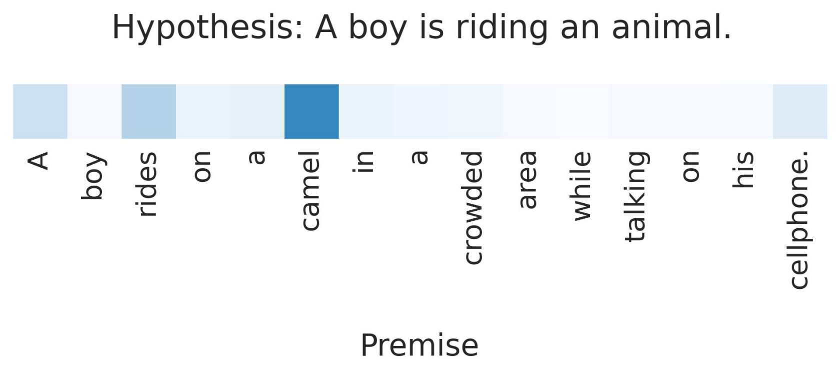
$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}^i\mathbf{H} + \mathbf{b}^i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}^f\mathbf{H} + \mathbf{b}^f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}^o\mathbf{H} + \mathbf{b}^o) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}^c\mathbf{H} + \mathbf{b}^c) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t).
\end{aligned}$$

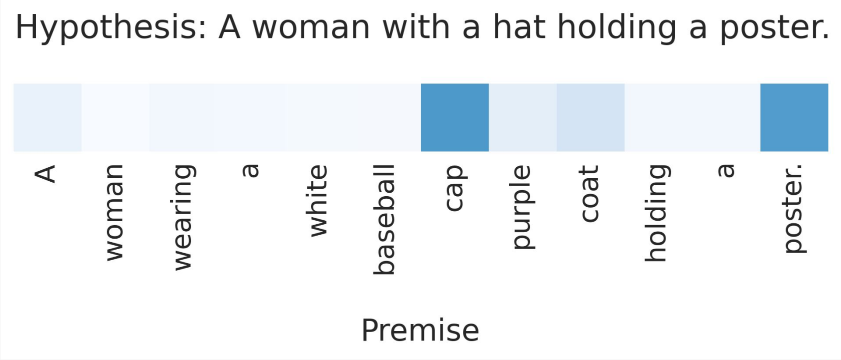
Attention (Bahdanau et al., 2014; Mnih et al., 2014)



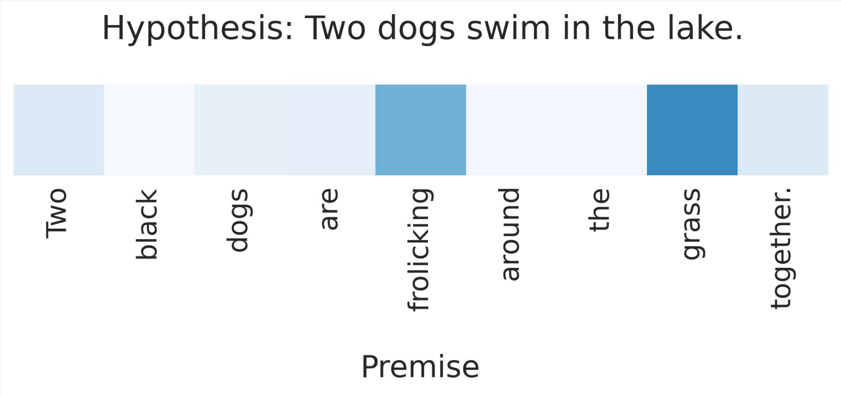
$$\mathbf{M} = \tanh(\mathbf{W}^y\mathbf{Y} + \mathbf{W}^h\mathbf{h}_N \otimes \mathbf{e}_L)$$
$$\alpha = \mathrm{softmax}(\mathbf{w}^T\mathbf{M})$$
$$\mathbf{r} = \mathbf{Y}\alpha.$$

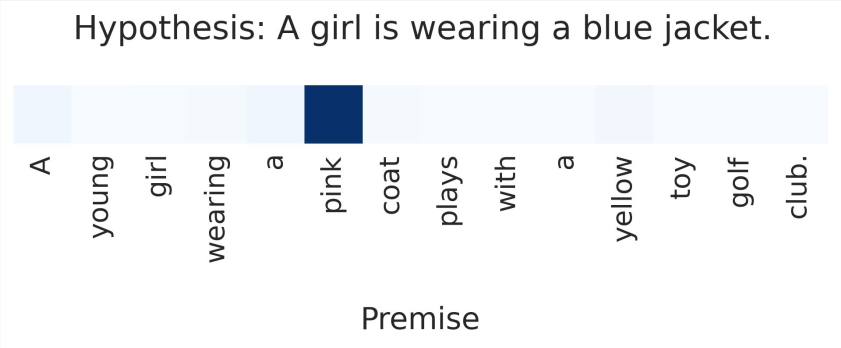
Word Matching

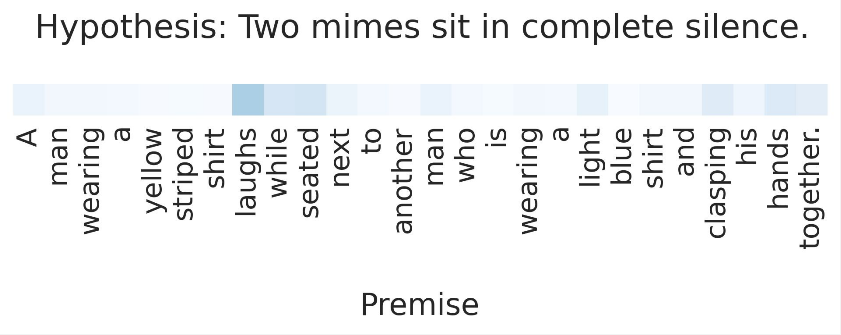




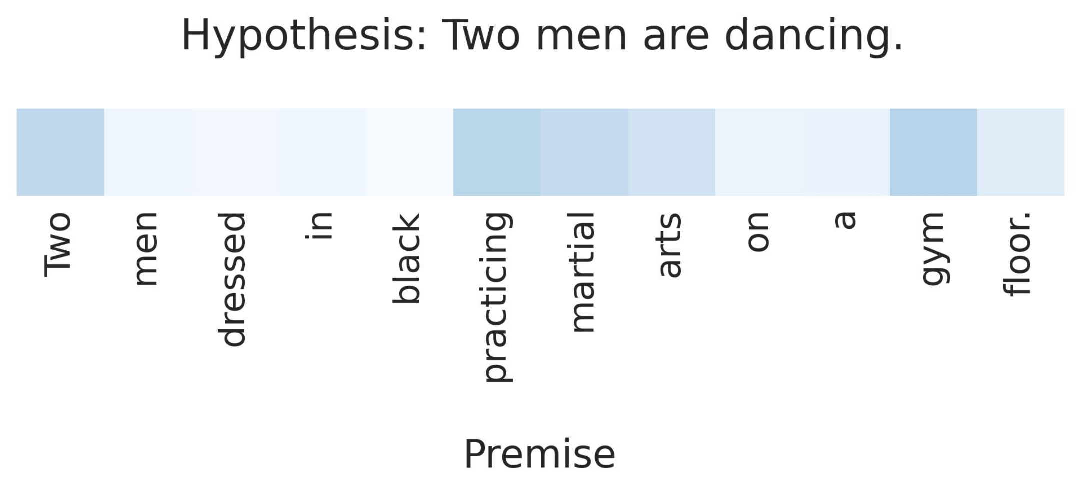
Spotting Contradictions



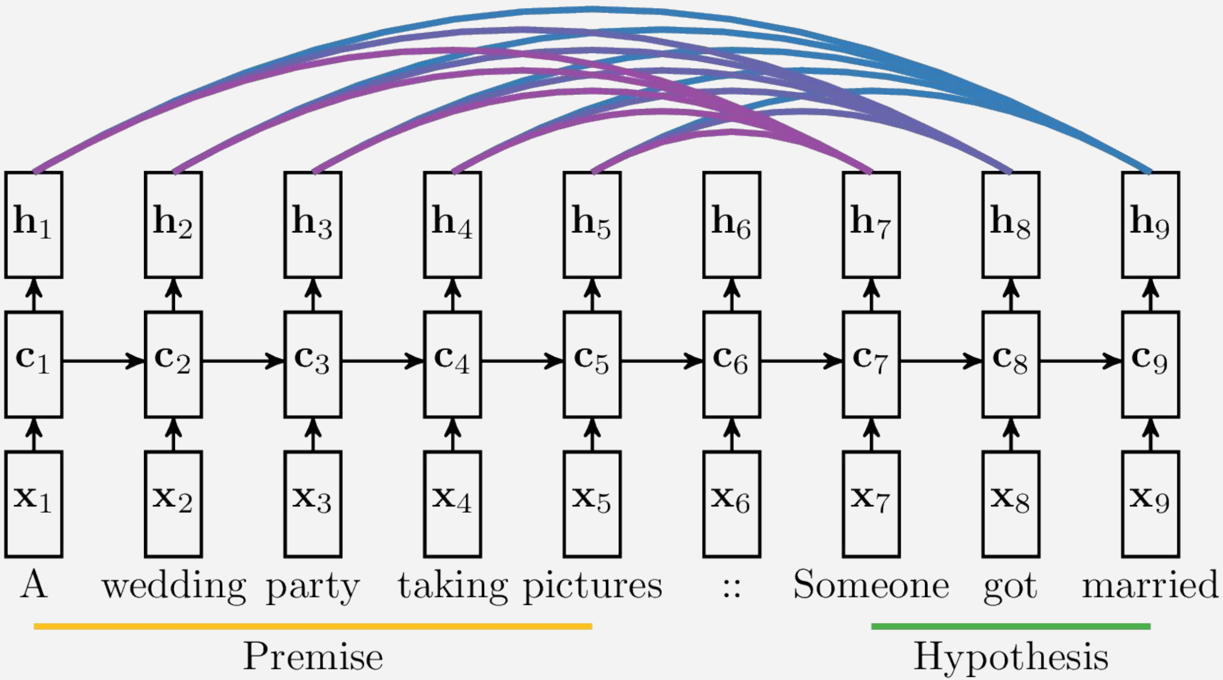




Fuzzy Attention

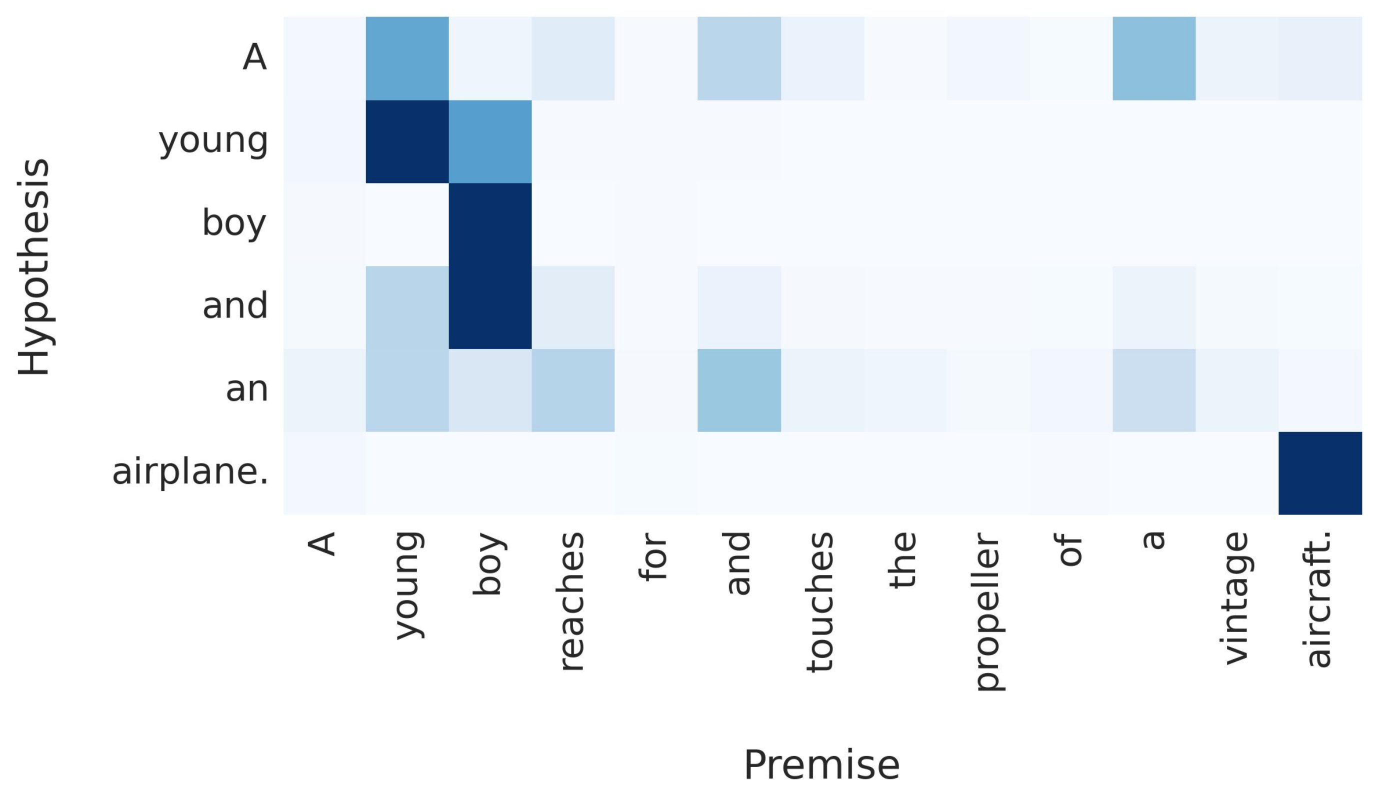


Word-by-Word Attention (Hermann et al. 2015)

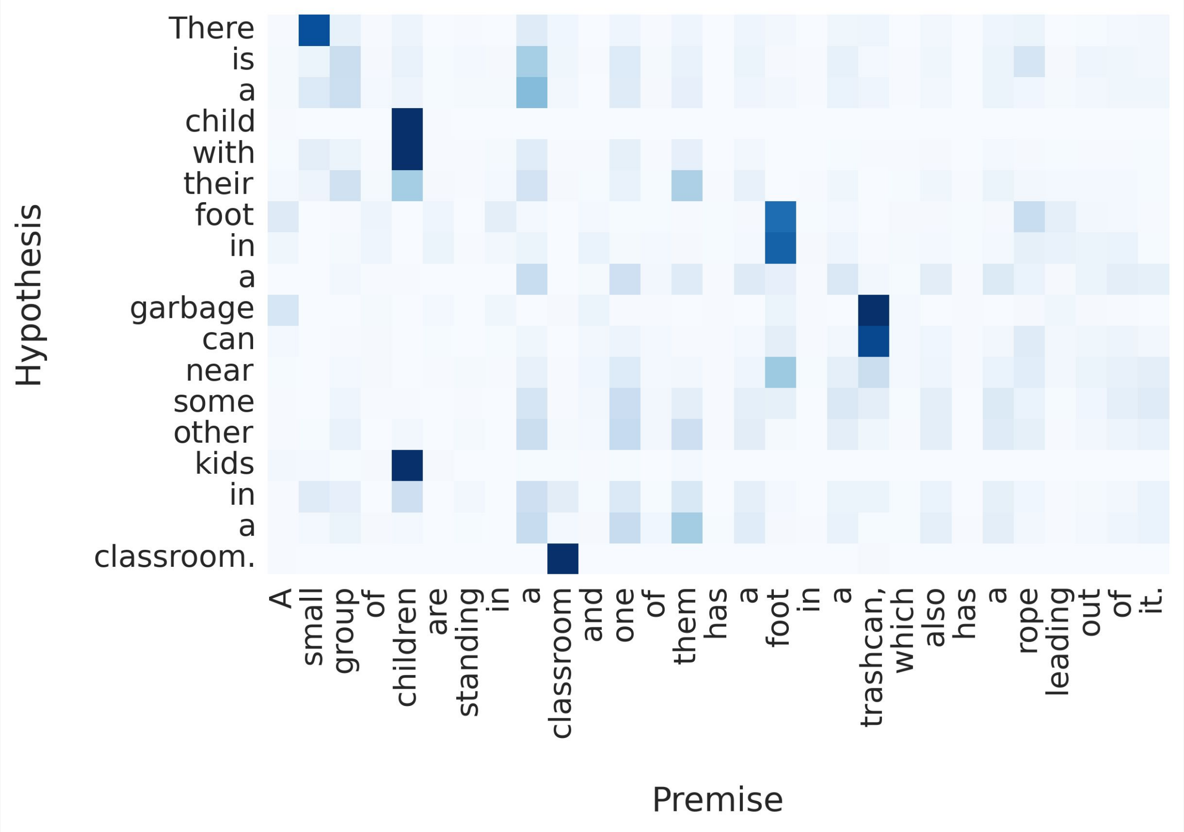


$$\mathbf{M}_t = \tanh(\mathbf{W}^y\mathbf{Y} + (\mathbf{W}^h\mathbf{h}_t + \mathbf{W}^r\mathbf{r}_{t-1}) \otimes \mathbf{e}_L)$$
$$\alpha_t = \mathrm{softmax}(\mathbf{w}^T\mathbf{M}_t)$$
$$\mathbf{r}_t = \mathbf{Y}\alpha_t + \tanh(\mathbf{W}^t\mathbf{r}_{t-1}).$$

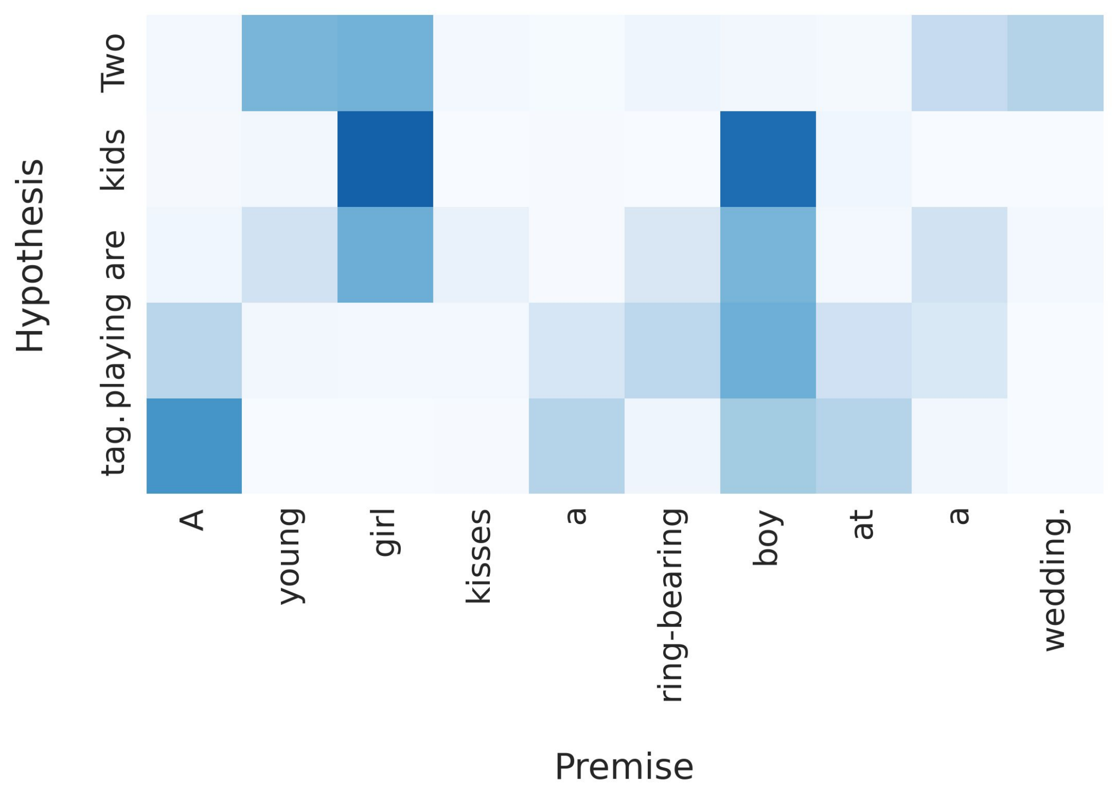
Word Matching and Synonyms



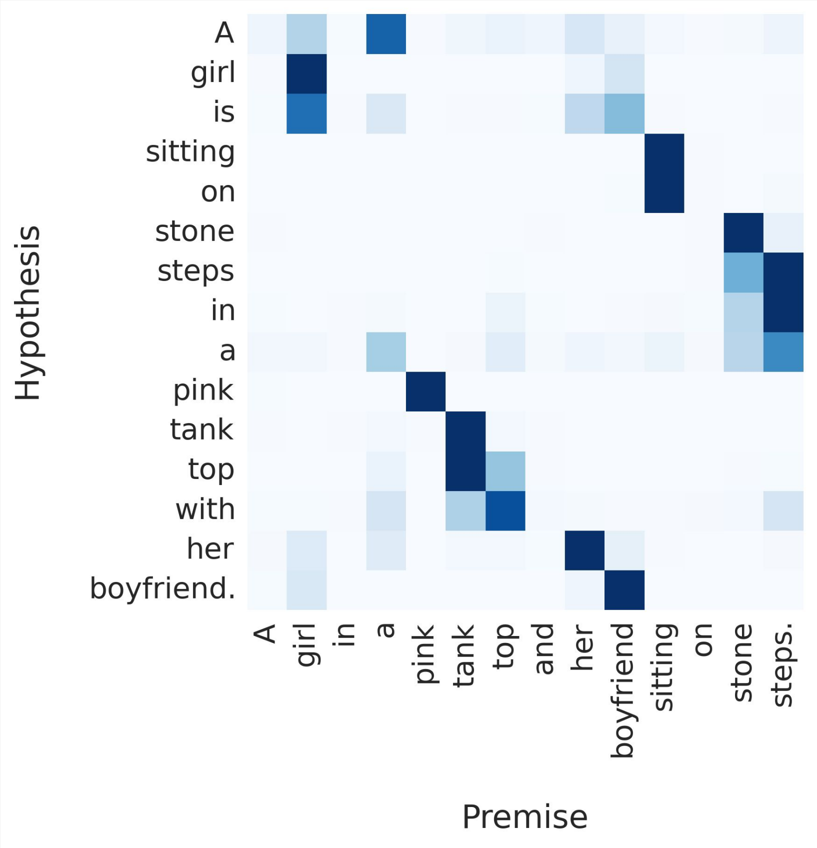
Words and Phrases



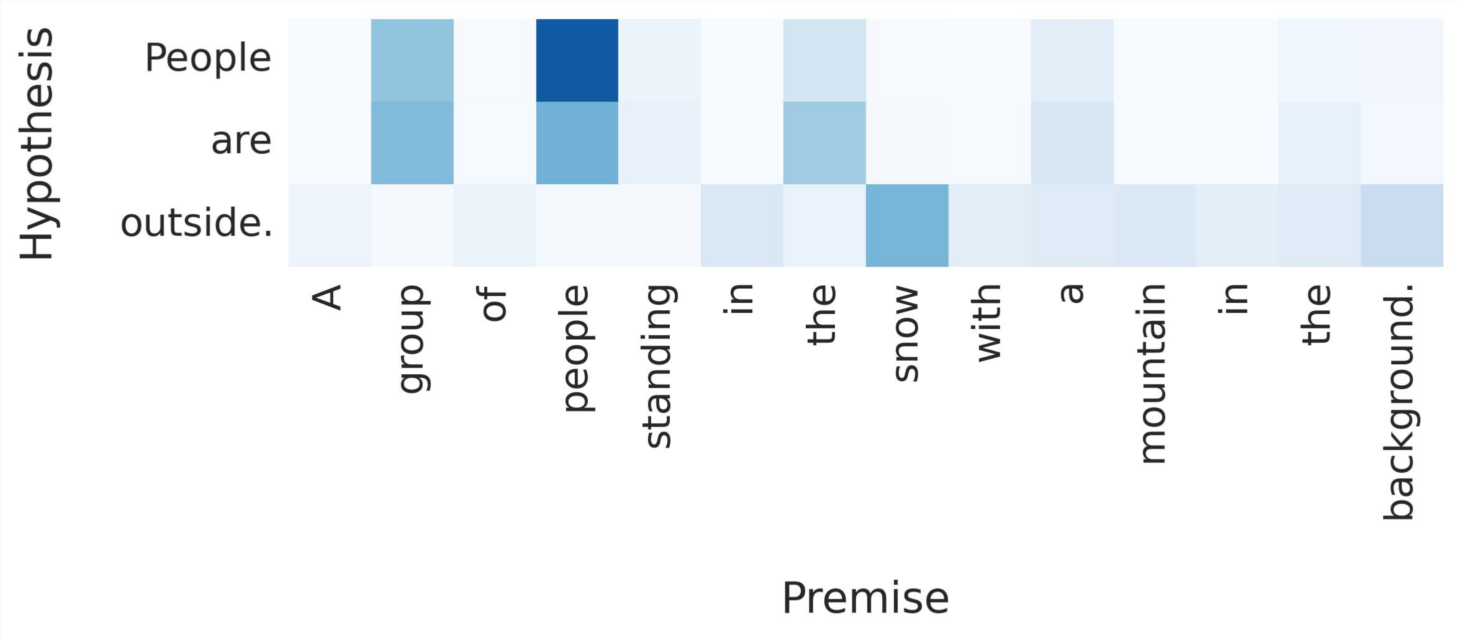
Girl + Boy = Kids



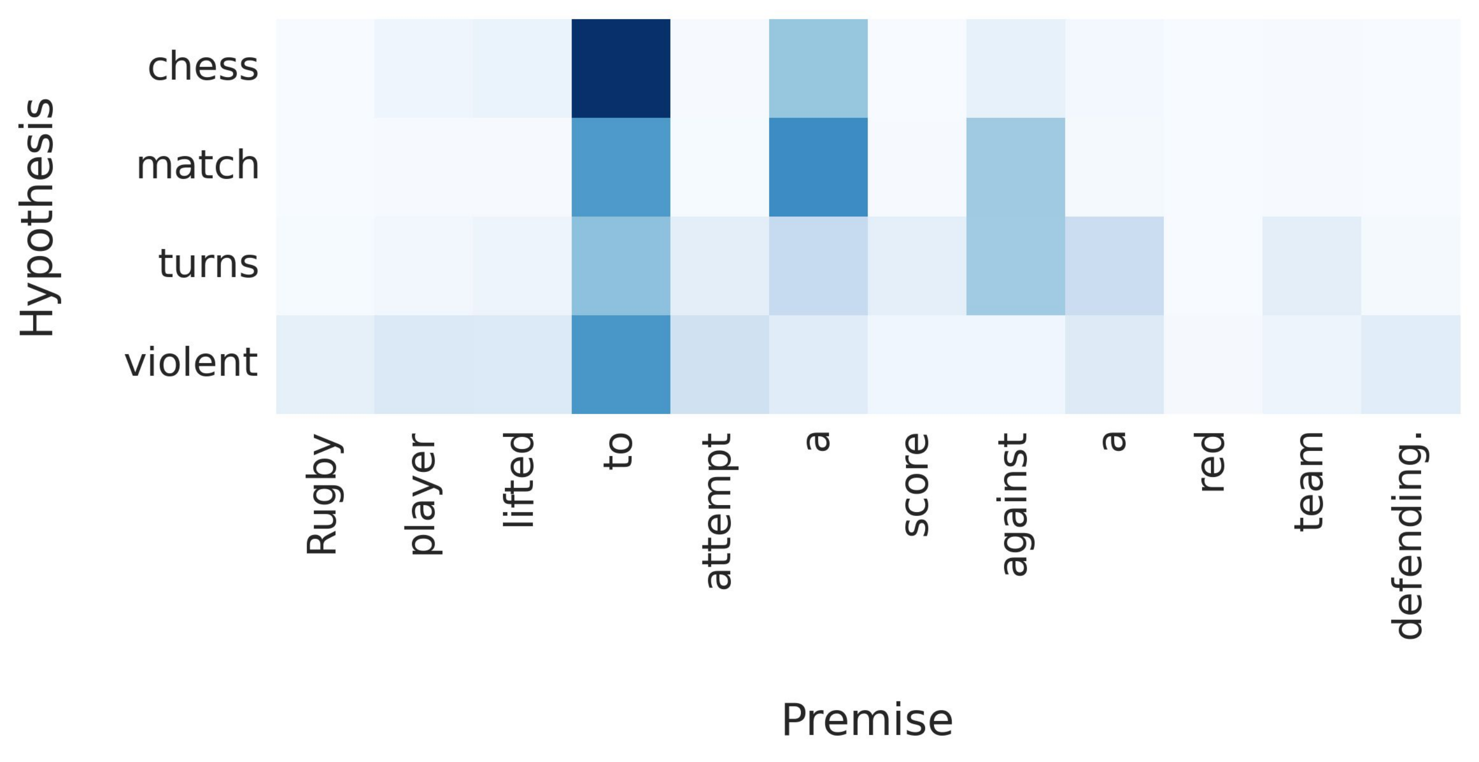
Reordering


Hypothesis
A
girl
is
sitting
on
stone
steps
in
a
pink
tank
top
with
her
boyfriend.
A
girl
in
a
pink
tank
top
and
her
boyfriend
sitting
on
stone
steps.
Premise

Snow is outside



It can get confused



Results

Model	k	$\|\theta\|_{\mathrm{W+M}}$	$\|\theta\|_{\mathrm{M}}$	Train	Dev	Test
LSTM [Bowman et al., 2015]	100	$\approx$ 10M	221k	84.4	-	77.6
Classifier [Bowman et al., 2015]	-	-	-	99.7	-	78.2
LSTM shared	100	3.8M	111k	83.7	81.9	80.9
LSTM shared	159	3.9M	252k	84.4	83.0	81.4
LSTMs	116	3.9M	252k	83.5	82.1	80.9
Attention	100	3.9M	242k	85.4	83.2	82.3
Attention two-way	100	3.9M	242k	86.5	83.0	82.4
Word-by-word attention	100	3.9M	252k	85.3	**83.7**	**83.5**
Word-by-word attention two-way	100	3.9M	252k	86.6	83.6	83.2

Thanks for listening!

Learning to Transduce with Unbounded Memory (NIPS 2015)
Grefenstette *et al.* 2015, arXiv:1506.02516 [cs.NE]

Teaching Machines to Read and Comprehend (NIPS 2015)
Hermann *et al.* 2015, arXiv:1506.03340 [cs.CL]

Reasoning about Entailment with Neural Attention (upcoming)
Rocktäschel *et al.* 2015, arXiv:1509.06664 [cs.CL]

joinus@deepmind.com