

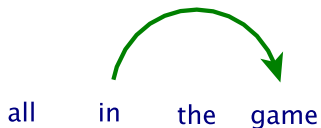
# Teaching Machines to Read and Comprehend

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette,  
Lasse Espeholt, Will Kay, Mustafa Suleyman, Lei Yu,  
and **Phil Blunsom**

[pblunsom@google.com](mailto:pblunsom@google.com)



DEPARTMENT OF  
**COMPUTER  
SCIENCE**



$$p(\text{game}|\text{in}) \propto \exp(\mathbf{w}^T \Phi(\text{game}, \text{in}))$$

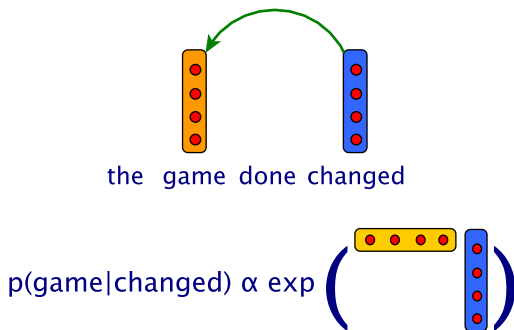
$$\Phi_1(x, y) = \begin{cases} 1, & \text{if PoS}(x)=\text{Noun} \ \& \ y=\text{in} \\ 0, & \text{otherwise} \end{cases}$$

$$\Phi_2(x, y) = \begin{cases} 1, & \text{if } x=\text{game} \ \& \ \text{PoS}(y)=\text{Prep} \\ 0, & \text{otherwise} \end{cases}$$

etc.

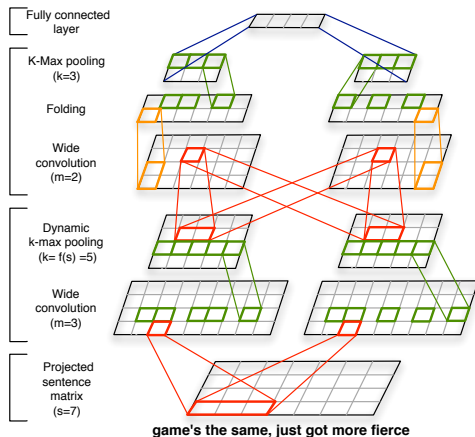
Twenty years ago log-linear models allowed greater freedom to model correlations than simple multinomial parametrisations, but imposed the need for feature engineering.

# Features and NLP



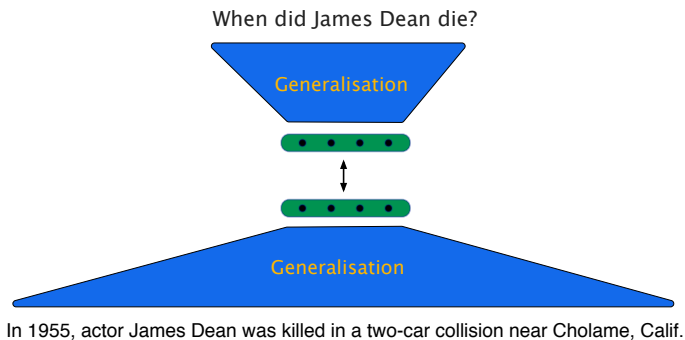
Distributed/neural models allow us to learn shallow features for our classifiers, capturing simple correlations between inputs.

# Deep Learning and NLP



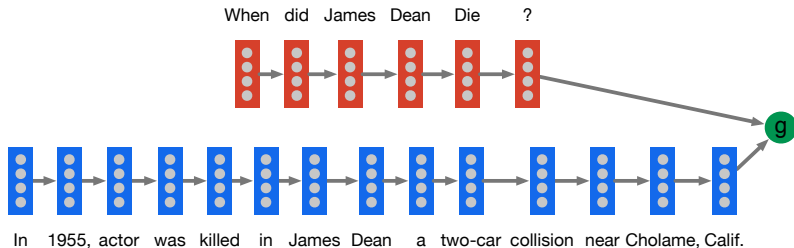
Deep learning should allow us to learn hierarchical generalisations.

# Deep Learning and NLP: Question Answer Selection



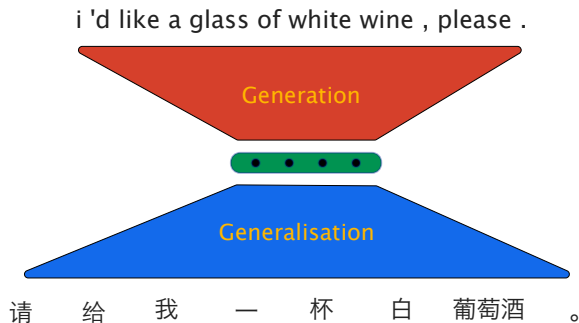
Beyond classification, deep models for embedding sentences have seen increasing success.

# Deep Learning and NLP: Question Answer Selection



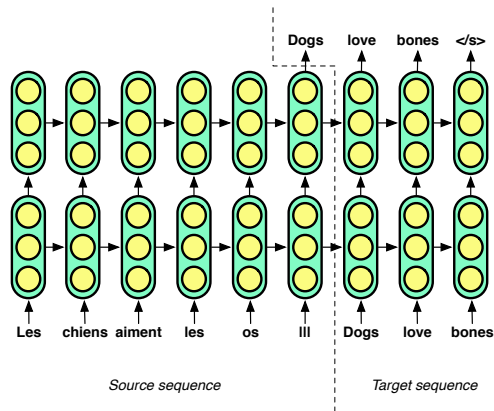
Recurrent neural networks provide a very practical tool for sentence embedding.

# Deep Learning for NLP: Machine Translation



We can even view translation as encoding and decoding sentences.

# Deep Learning for NLP: Machine Translation



Recurrent neural networks again perform surprisingly well.





## Small steps towards NLU:

- reading and understanding text,
- connecting natural language, action, and inference in real environments.

# Supervised Reading Comprehension



To achieve our aim of training supervised machine learning models for machine reading and comprehension, we must first find data.

# Supervised Reading Comprehension: MCTest

## Document

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.

...

## Question

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

## Synthetic example from the FaceBook data set

John picked up the apple.

John went to the office.

John went to the kitchen.

John dropped the apple.

Where was the apple before the kitchen? **A:office**

An alternative to real language is to generate scripts from a synthetic grammar.

# Supervised Reading Comprehension

Cookie Policy / Facebook [Like](#) [+1](#) [Follow](#) [@DailyMail](#) [@DailyMail](#) Sunday, Jun 7th 2015 10PM EDT 8°C 1AM EDT 5-Day Forecast

## MailOnline

Home | News | U.S. | Sport | TV & Shows | Australia | Femail | Fashion Finder

Latest Headlines | News | Arts | Headlines | Pictures | Most Popular

Kate Winslet | Royal Navy | Pack | Do | You Might Like

Why it's hell living next to the REAL interns: £4,000-a-month Googlers 'ter residents at San Francisco apartment complex with their constant partying'

- "I thought it was summer camp." Around 400 interns receive Crestmont Village Apartments in North San Jose last month; residents weren't warned.
- The covered poolside pay \$6,000/month with free food, transportation, and leisure activities.

By DAVID H. REPPORTER  
PUBLISHED: 18:24, 9 July 2015 | UPDATED: 07:55, 9 July 2015

Facebook | Twitter | LinkedIn | YouTube | RSS | Email

Hundreds of Google interns have flooded a San Francisco Bay Area complex and the fabric residents say their partying and late hours gotten out of control.

The month-long summer internship at the luxury college student apartments of \$5,500 per month and, apparently, also afford the best view in Silicon Valley.

Meanwhile, the children and families who already called Crestmont apartments in North San Jose home are wishing these twenty-somethings go back to where they came from.



Google reveals it is developing a computer so smart it can program ITSELF

Neural Turing Machine being developed by DeepMind, which Google bought in January

Project mimics properties of the human brain's short-term working memory

Test grant is also working on quantum chips based on the human brain

By ANNA FRISCH and VICTORIA WOOLLASTON FOR MAILONLINE  
PUBLISHED: 20:28, 28 October 2014 | UPDATED: 00:30, 26 October 2015

Facebook | Twitter | LinkedIn | YouTube | RSS | Email

Google's secretive artificial intelligence researchers have revealed a computer that first tests will one day be able to program itself.

Developers at Google's secretive DeepMind start-up, which it bought for \$12 million earlier this year, are attempting to mimic some of the properties of human brain's short-term working memory.

By combining the way ordinary computers work with the way the human brain works, the researchers hope the machine will learn to program itself.



Don't worry, I can't Google that. Max, the smart computer from Stanley Kubrick's 2001 moved a step closer today as developers at Google's secretive DeepMind start-up revealed they are attempting a mimic some of the properties of the human brain's short-term working memory.

Google's DeepMind division is using Daily Mail and CNN articles to artificial intelligence programs to read.

Using the unique style of articles on the sites, with concise bullet summarizing a story at the top of a page, artificial intelligence we key facts about articles to answer queries.

Ultimately, scientists hope that the study could lead to complex of that can read entire documents and respond to questions just as if human.



11 Benefits to a Toned Body  
10 Top/Best Regions List  
10 Top/Best Areas/You Don't Know...  
14 C Who Got Rich

Unica Method  
Regions List  
Hot Land  
8 Top/Best  
14 C Who Got Rich

## MailOnline

Home | News | U.S. | Sport | TV & Shows | Australia | Femail | Health | Science | Money | Video | Travel

Latest Headlines | Science | Pictures

You Might Like

Sponsored Links by Taboola

### Happy 75th birthday, Chuck Norris!

By Todd Leopold, CNN | Updated 2:09 GMT (05:09 EDT) March 10, 2015



More Top Stories

- NCAA: We're not responsible for education
- Californians get mandatory water restrictions
- Senator indicted on bribery, other charges
- Research charges date from Frank, dad
- A real 'Fault in Our Stars' couple
- Inside Mike Tyson's abandoned mansion

More from CNN

- The great artificial intelligence debate
- NASA to test flying tractor

Story Highlights

Tuesday is Chuck Norris' 75th birthday.

Chuck Norris cracked to infinity. Twice.

Death once had a near-Chuck Norris experience.

Chuck Norris is celebrating his 75th birthday — but the calendar is only allowed to turn 36.

The actor, martial-arts star and world's favorite tough-guy (see subject box) was born March 10, 1940, which makes him 75 today.

Or perhaps he IS 36. Because maybe YOU can't beat him, but Chuck Norris can beat anything.

Happy birthday!

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

# Supervised Reading Comprehension

## CNN article:

**Document** 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.” . . .

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

**Answer** Oisin Tymon

We formulate *Cloze* style queries from the story paraphrases.

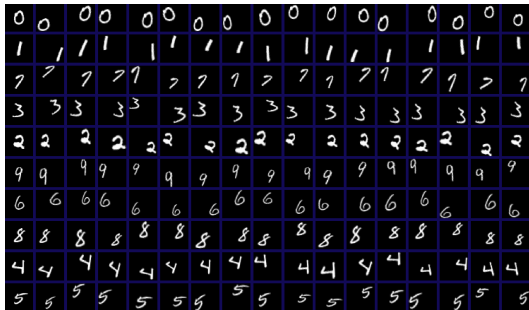
## From the Daily Mail:

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

An ngram language model would correctly predict (**X** = *cancer*), regardless of the document, simply because this is a frequently cured entity in the Daily Mail corpus.

# Supervised Reading Comprehension

MNIST example generation:



We generate quasi-synthetic examples from the original document-query pairs, obtaining exponentially more training examples by anonymising and permuting the mentioned entities.



# Supervised Reading Comprehension

Original Version	Anonymised Version
<b>Context</b> 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." ...	the <i>ent381</i> producer allegedly struck by <i>ent212</i> will not press charges against the " <i>ent153</i> " host, his lawyer said Friday. <i>ent212</i> , who hosted one of the most-watched television shows in the world, was dropped by the <i>ent381</i> Wednesday after an internal investigation by the <i>ent180</i> broadcaster found he had subjected producer <i>ent193</i> "to an unprovoked physical and verbal attack." ...
<b>Query</b> Producer <b>X</b> will not press charges against Jeremy Clarkson, his lawyer says.	producer <b>X</b> will not press charges against <i>ent212</i> , his lawyer says.
<b>Answer</b> Oisin Tymon	<i>ent193</i>

Original and anonymised version of a data point from the Daily Mail validation set. The anonymised entity markers are constantly permuted during training and testing.

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	108k	1k	1k	195k	12k	11k
# queries	438k	4k	3k	838k	61k	55k
Max # entities	456	190	398	424	247	250
Avg # entities	30	32	30	41	45	45
Avg tokens/doc	780	809	773	1044	1061	1066
Vocab size		125k			275k	

Articles were collected from April 2007 for CNN and June 2010 for the Daily Mail, until the end of April 2015. Validation data is from March, test data from April 2015.

Category	Sentences		
	1	2	$\geq 3$
Simple	12	2	0
Lexical	14	0	0
Coref	0	8	2
Coref/Lex	10	8	4
Complex	8	8	14
Unanswerable		10	

Distribution (in percent) of queries over category and number of context sentences required to answer them based on a subset of the CNN validation data.

## Frequency baselines (Accuracy)

	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

A simple baseline is to always predict the entity appearing most often in the document. A refinement of this is to exclude entities in the query.

# Frame semantic matching

A stronger benchmark using a state-of-the-art frame semantic parser and rules with an increasing recall/precision trade-off:

	Strategy	Pattern $\in Q$	Pattern $\in D$	Example (Cloze / Context)
1	Exact match	$(p, V, y)$	$(x, V, y)$	X loves Suse / <b>Kim</b> loves Suse
2	be.01.V match	$(p, be.01.V, y)$	$(x, be.01.V, y)$	X is president / <b>Mike</b> is president
3	Correct frame	$(p, V, y)$	$(x, V, z)$	X won Oscar / <b>Tom</b> won Academy Award
4	Permuted frame	$(p, V, y)$	$(y, V, x)$	X met Suse / Suse met <b>Tom</b>
5	Matching entity	$(p, V, y)$	$(x, Z, y)$	X likes candy / <b>Tom</b> loves candy
6	Back-off strategy	<i>Pick the most frequent entity from the context that doesn't appear in the query</i>		

$x$  denotes the entity proposed as answer,  $V$  is a fully qualified PropBank frame (e.g. *give.01.V*). Strategies are ordered by precedence and answers determined accordingly.

# Frame semantic matching

	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

Failure modes:

- The Propbank parser has poor coverage with many relations not picked up as they do not adhere to the default predicate-argument structure.
- The frame-semantic approach does not trivially scale to situations where several frames are required to answer a query.

# Word distance benchmark

Consider the query *“Tom Hanks is friends with X’s manager, Scooter Brown”* where the document states *“... turns out he is good friends with Scooter Brown, manager for Carly Rae Jepsen.”*

The frame-semantic parser fails to pickup the friendship or management relations when parsing the query.

# Word distance benchmark

Word distance benchmark:

- align the placeholder of the *Cloze* form question with each possible entity in the context document,
- calculate a distance measure between the question and the context around the aligned entity,
- sum the distances of every word in  $Q$  to its nearest aligned word in  $D$ .

Alignment is defined by matching words either directly or as aligned by the coreference system.



# Word distance benchmark

	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

This benchmark is robust to small mismatches between the query and answer, correctly solving most instances where the query is generated from a highlight which in turn closely matches a sentence in the context document.

Use neural encoding models for estimating the probability of word type  $a$  from document  $d$  answering query  $q$ :

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

where  $W(a)$  indexes row  $a$  of weight matrix  $W$  and function  $g(d, q)$  returns a vector embedding of a document and query pair.

# Deep LSTM Reader

We employ a Deep LSTM cell with skip connections,

$$x'(t, k) = x(t) || y'(t, k - 1),$$

$$i(t, k) = \sigma (W_{kxi}x'(t, k) + W_{khi}h(t - 1, k) + W_{kci}c(t - 1, k) + b_{ki}),$$

$$f(t, k) = \sigma (W_{kxf}x(t) + W_{khf}h(t - 1, k) + W_{kcf}c(t - 1, k) + b_{kf}),$$

$$c(t, k) = f(t, k)c(t - 1, k) + i(t, k) \tanh (W_{kxc}x'(t, k) + W_{khc}h(t - 1, k) + b_{kc}),$$

$$o(t, k) = \sigma (W_{kxo}x'(t, k) + W_{kho}h(t - 1, k) + W_{kco}c(t, k) + b_{ko}),$$

$$h(t, k) = o(t, k) \tanh (c(t, k)),$$

$$y'(t, k) = W_{ky}h(t, k) + b_{ky},$$

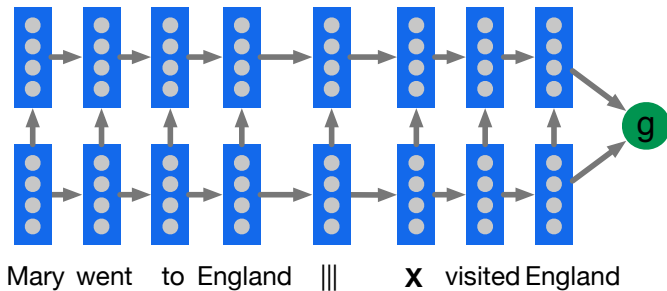
$$y(t) = y'(t, 1) || \dots || y'(t, K),$$

where  $||$  indicates vector concatenation  $h(t, k)$  is the hidden state for layer  $k$  at time  $t$ , and  $i, f, o$  are the input, forget, and output gates respectively.

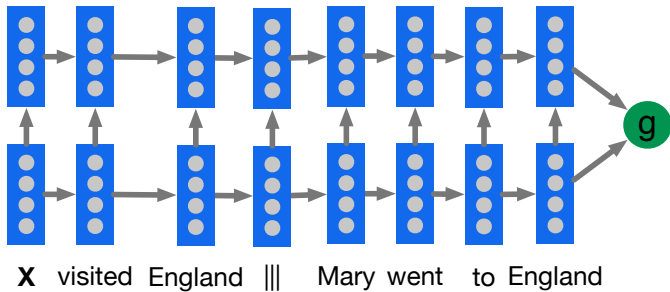
$$g^{\text{LSTM}}(d, q) = y(|d| + |q|)$$

with input  $x(t)$  the concatenation of  $d$  and  $q$  separated by the delimiter  $|||$ .

# Deep LSTM Reader



# Deep LSTM Reader



	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

Given the difficulty of its task, the Deep LSTM Reader performs very strongly.

# The Attentive Reader

Denote the outputs of a bidirectional LSTM as  $\vec{y}(t)$  and  $\overleftarrow{y}(t)$ . Form two encodings, one for the query and one for each token in the document,

$$u = \vec{y}_q(|q|) \parallel \overleftarrow{y}_q(1), \quad y_d(t) = \vec{y}_d(t) \parallel \overleftarrow{y}_d(t).$$

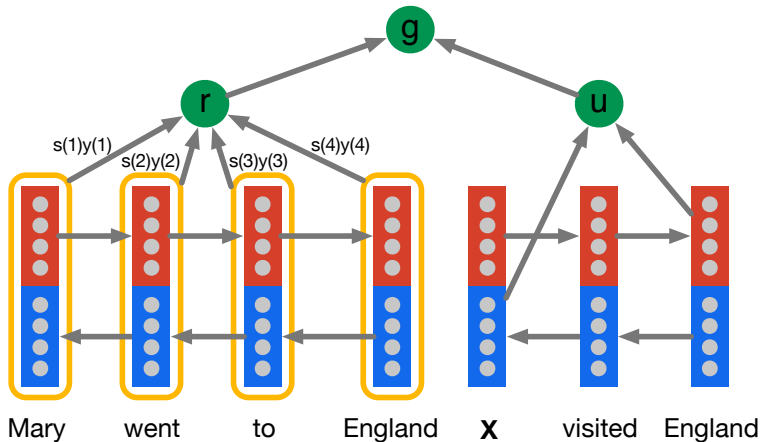
The representation  $r$  of the document  $d$  is formed by a weighted sum of the token vectors. The weights are interpreted as the model's attention,

$$\begin{aligned} m(t) &= \tanh(W_{ym}y_d(t) + W_{um}u), \\ s(t) &\propto \exp(w_{ms}^T m(t)), \\ r &= y_d s. \end{aligned}$$

Define the joint document and query embedding via a non-linear combination:

$$g^{\text{AR}}(d, q) = \tanh(W_{rg}r + W_{ug}u).$$

# The Attentive Reader





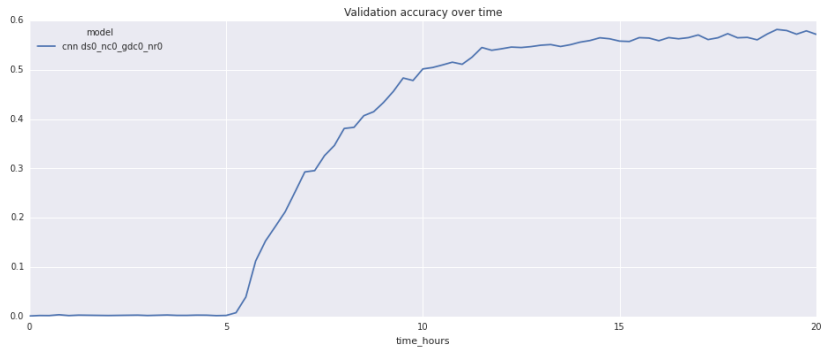
# The Attentive Reader

	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 <sup>1</sup>	31.1	33.6	31.0	31.7
Attentive Reader	56.5	58.9	64.5	63.7

The attention variables effectively address the Deep LSTM Reader's inability to focus on part of the document.

<sup>1</sup>The Uniform attention baseline sets all  $m(t)$  parameters to be equal.

# Attentive Reader Training



Models were trained using asynchronous minibatch stochastic gradient descent (RMSProp) on approximately 25 GPUs.

# The Attentive Reader: Predicted: *ent49*, Correct: *ent49*

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

# The Attentive Reader: Predicted: *ent27*, Correct: *ent27*

by *ent82*, *ent38* updated 9:35 am et, mon march 2, 2015 (*ent38*) *ent27* went familial for fall at its fashion show in *ent23* on sunday, dedicating its collection to ``mamma'' with nary a pair of ``mom jeans'' in sight. *ent57* and *ent78*, who are behind the *ent72* brand, sent models down the runway in decidedly feminine dresses and skirts adorned with roses, lace and even embroidered doodles by the designers' own nieces and nephews. many of the looks featured saccharine needlework phrases like ``i love you, mamma'' and ``*ent46*'' (for the most beautiful mother in the world) as a tableau vivant of moms and daughters stood and posed as a backdrop for the runway. our little munchkins backstage *ent44* babies # friends # \_UNK\_ a photo posted by *ent58* (@\_UNK\_) on mar 1, 2015 at \_UNK\_ *ent17* even the usually stoic - faced front row could n't help but applaud and smile as a few models carried their own high-fashion progeny down the runway. almost ready for the show : watch the *ent87* live today at *ent8* (*ent65*) on *ent87* website. #\_UNK\_ #\_UNK\_ #\_UNK\_ #\_UNK\_ #\_UNK\_ #\_UNK\_ a photo posted by *ent27* (@\_UNK\_) on mar 1, 2015 at \_UNK\_ *ent17*

X dedicated their fall fashion show to moms

# The Attentive Reader: Predicted: *ent85*, Correct: *ent37*

by *ent52* and *ent22* ,*ent43* updated 7:12 am et ,fri march 20 ,2015 *ent74* ,*ent37* (*ent43*) a passenger train overshot a stop and jumped its tracks in northern *ent37* on friday ,killing at least 30 people and injuring more than 50 others ,a railway spokesman said .the train was headed from *ent85* to the *ent27* holy city of *ent13* when it overshot an intended stop more than halfway along the route ,about 35 kilometers ( 22 miles ) east of *ent11* in the northern state of *ent56* , railway spokesman *ent20* said .two coaches and the locomotive derailed .video from the site ,shown by *ent43* affiliate *ent33* ,showed emergency workers pulling passengers from the train as a crowd looked on .the cause of the incident will be investigated ,*ent20* said .*ent43* 's *ent52* reported from *ent74* .*ent43* 's *ent22* wrote in *ent15* .

a passenger train derails about 35 kilometers ( 22 miles ) east of *ent11* in northern X

# The Attentive Reader: Predicted: *ent24*, Correct: *ent2*

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

# The Impatient Reader

At each token  $i$  of the query  $q$  compute a representation vector  $r(i)$  using the bidirectional embedding  $y_q(i) = \vec{y}_q(i) \parallel \overleftarrow{y}_q(i)$ :

$$m(i, t) = \tanh(W_{dm}y_d(t) + W_{rm}r(i-1) + W_{qm}y_q(i)), 1 \leq i \leq |q|,$$

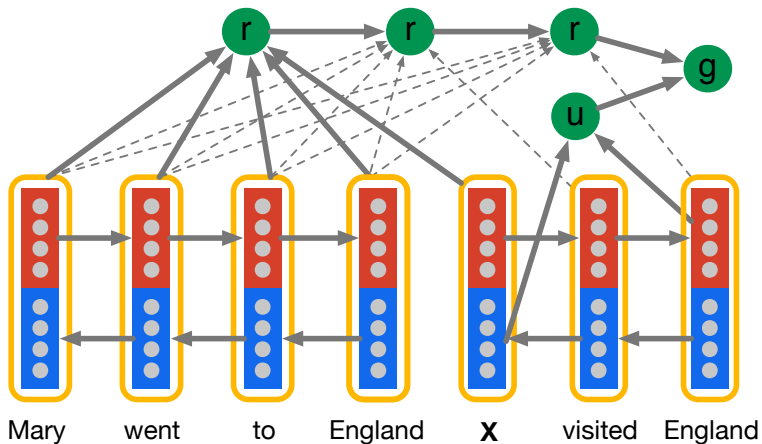
$$s(i, t) \propto \exp(w_{ms}^T m(i, t)),$$

$$r(0) = \mathbf{r}_0, \quad r(i) = y_d^T s(i), \quad 1 \leq i \leq |q|.$$

The joint document query representation for prediction is,

$$g^{\text{IR}}(d, q) = \tanh(W_{rg}r(|q|) + W_{qg}u).$$

# The Impatient Reader



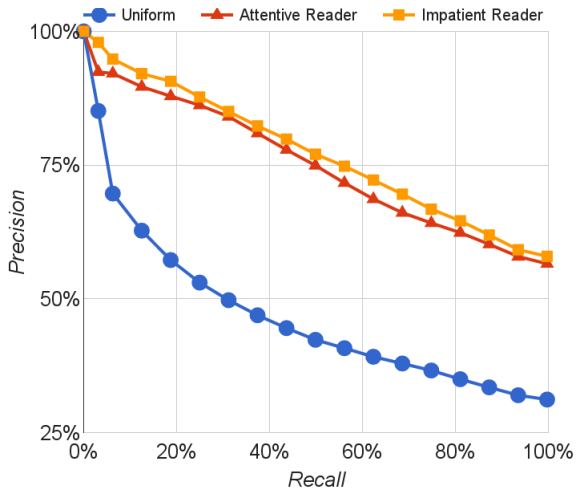


# The Impatient Reader

	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	<b>57.0</b>	<b>60.6</b>	<b>64.8</b>	<b>63.9</b>

The Impatient Reader comes out on top, but only marginally.

# Attention Models Precision@Recall



Precision@Recall for the attention models on the CNN validation data.

## Summary

- supervised machine reading is a viable research direction with the available data,
- LSTM based recurrent networks constantly surprise with their ability to encode dependencies in sequences,
- attention is a very effective and flexible modelling technique.

## Future directions

- more and better data, corpus querying, and cross document queries,
- recurrent networks incorporating long term and working memory are well suited to NLU task.



Google DeepMind



DEPARTMENT OF  
**COMPUTER  
SCIENCE**