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

\[ \beta \]

\[ w_1 \quad w_2 \quad w_3 \]
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

\[ S = s_1 s_2 \ldots s_m \]

into some target sequence

\[ T = t_1 t_2 \ldots t_n \]
Transduction and RNNs

Approach:

1. Model $P(t_{i+1}|t_1...t_n; S)$ 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=8504
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

\[
\begin{align*}
\text{row 3} & : t = 1 \quad u_1 = 0 \quad d_1 = 0.8 \\
\text{row 2} & : t = 2 \quad u_2 = 0.1 \quad d_2 = 0.5 \\
\text{row 1} & : t = 3 \quad u_3 = 0.9 \quad d_3 = 0.9
\end{align*}
\]

\[r_1 = 0.8 \cdot v_1\]
\[r_2 = 0.5 \cdot v_2 + 0.5 \cdot v_1\]
\[r_3 = 0.9 \cdot v_3 + 0.1 \cdot v_1\]

\(v_3\) removed from stack
Example: A Continuous Stack
Controlling a Neural Stack
Synthetic Transduction Tasks

Copy

\[ a_1 a_2 a_3 \ldots a_n \rightarrow a_1 a_2 a_3 \ldots a_n \]

Reversal

\[ a_1 a_2 a_3 \ldots a_n \rightarrow a_n \ldots a_3 a_2 a_1 \]

Bigram Flipping

\[ a_1 a_2 a_3 a_4 \ldots a_{n-1} a_n \rightarrow a_2 a_1 a_4 a_3 \ldots a_n a_{n-1} \]
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

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Stack</th>
<th>Queue</th>
<th>DeQue</th>
<th>Deep LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy</td>
<td>Poor</td>
<td>Solved</td>
<td>Solved</td>
<td>Poor</td>
</tr>
<tr>
<td>Reversal</td>
<td>Solved</td>
<td>Poor</td>
<td>Solved</td>
<td>Poor</td>
</tr>
<tr>
<td>Bigram Flip</td>
<td>Converges</td>
<td>Best Results</td>
<td>Best Results</td>
<td>Converges</td>
</tr>
<tr>
<td>SVO-SOV</td>
<td>Solved</td>
<td>Solved</td>
<td>Solved</td>
<td>Converges</td>
</tr>
<tr>
<td>Conjugation</td>
<td>Converges</td>
<td>Solved</td>
<td>Solved</td>
<td>Converges</td>
</tr>
</tbody>
</table>

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
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. ...
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
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 ($X = \text{cancer}$)
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 World Cup final?
Answer: New Zealand

... our way

(enter23) enter7 are on course for a first ever enter15 title after a thrilling semifinal victory over enter34, 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 enter12, enter17 hit a six right at the death to confirm victory and send the enter83 crowd into raptures. It is the first time they have ever reached a enter15 final.

Question: ____ reach enter3 enter15 final?
Answer: enter7
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

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN valid</th>
<th>CNN test</th>
<th>Daily Mail valid</th>
<th>Daily Mail test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum frequency</td>
<td>26.3</td>
<td>27.9</td>
<td>22.5</td>
<td>22.7</td>
</tr>
<tr>
<td>Exclusive frequency</td>
<td>30.8</td>
<td>32.6</td>
<td>27.3</td>
<td>27.7</td>
</tr>
<tr>
<td>Frame-semantic model</td>
<td>32.2</td>
<td>33.0</td>
<td>30.7</td>
<td>31.1</td>
</tr>
<tr>
<td>Word distance model</td>
<td>46.2</td>
<td>46.9</td>
<td>55.6</td>
<td>54.8</td>
</tr>
</tbody>
</table>
We estimate the probability of word type $a$ from document $d$ answering query $q$:

$$p(a|d, q) \propto \exp(W(a)g(d, q)),$$

s.t. $a \in d$.

where $W(a)$ indexes row $a$ of $W$ and $g(d,q)$ embeds of a document and query pair.
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
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
## Impatience is a virtue - Results

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>valid</td>
<td>test</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td>26.3</td>
<td>27.9</td>
</tr>
<tr>
<td>Exclusive frequency</td>
<td>30.8</td>
<td>32.6</td>
</tr>
<tr>
<td>Frame-semantic model</td>
<td>32.2</td>
<td>33.0</td>
</tr>
<tr>
<td>Word distance model</td>
<td>46.2</td>
<td>46.9</td>
</tr>
<tr>
<td>Deep LSTM Reader</td>
<td>49.0</td>
<td>49.9</td>
</tr>
<tr>
<td>Uniform attention</td>
<td>31.1</td>
<td>33.6</td>
</tr>
<tr>
<td>Attentive Reader</td>
<td>56.5</td>
<td>58.9</td>
</tr>
<tr>
<td>Impatient Reader</td>
<td><strong>57.0</strong></td>
<td><strong>60.6</strong></td>
</tr>
</tbody>
</table>
The Attentive Reader - Correct Example

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

A yellow __________ 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 described 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 Morningstar station reported.

Lamborghini 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.
A driver was caught in the HOV lane 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)

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

10k sentence pairs, partly synthetic

570k sentence pairs from Mechanical Turkers

EMNLP 2015 “best data set or resource” award!
Model

General Artificial Intelligence

H = \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}

i_t = \sigma(W^i H + b^i)

f_t = \sigma(W^f H + b^f)

o_t = \sigma(W^o H + b^o)

c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W^e H + b^e)

h_t = o_t \odot \tanh(c_t).
Attention (Bahdanau et al., 2014; Mnih et al., 2014)

\[ M = \tanh(W^yY + W^h h_N \otimes e_L) \]
\[ \alpha = \text{softmax}(w^T M) \]
\[ r = Y\alpha. \]
Word Matching

Hypothesis: A boy is riding an animal.

Premise

Hypothesis: A woman with a hat holding a poster.

Premise

A boy rides on a camel in a crowded area while talking on his cellphone.

A woman wearing a white baseball cap holding a poster.
Spotting Contradictions

Hypothesis: Two dogs swim in the lake.

Premise:

Hypothesis: A girl is wearing a blue jacket.

Premise:

Hypothesis: Two mimes sit in complete silence.

Premise:
Fuzzy Attention

Hypothesis: Two men are dancing.

Premise

Two | men | dressed | in | black | practicing | martial | arts | on | a | gym | floor.
Word-by-Word Attention (Hermann et al. 2015)

\[ M_t = \tanh(W^vY + (W^h h_t + W^r r_{t-1}) \otimes e_t) \]
\[ \alpha_t = \text{softmax}(w^T M_t) \]
\[ r_t = Y \alpha_t + \tanh(W^r r_{t-1}) \]

A wedding party

Premise

taking pictures

:: Someone got married

Hypothesis
Word Matching and Synonyms

Hypothesis:
- young
- boy
- and
- an
- airplane.

Premise:
- A
- young
- boy
- reaches
- for
- and
- touches
- the
- propeller
- of
- a
- vintage
- aircraft.
There is a child with their foot in a garbage can near some other kids in a classroom.

A small group of children are standing in a classroom and one of them has a foot in a trashcan, which also has a rope leading out of it.
Girl + Boy = Kids
Reordering
Snow is outside
It can get confused
## Results

| Model                               | $k$ | $|\theta|_{w+M}$ | $|\theta|_M$ | 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  | $\textbf{83.7}$ | $\textbf{83.5}$ |
| 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