First half: NLP State-of-the-union
Second half: Where is NLP going?
Third half*: On-going work in the LUNR lab.

* Car Talk fans?
First half: NLP State-of-the-union
NLP = \{ \text{Text} \rightarrow \text{Label}, \text{Text} \rightarrow \text{Structure}, \text{Text} \rightarrow \text{Text} \}\}
NLP = \{ 
  \text{Text} \rightarrow \text{Label} \\
  \text{Text} \rightarrow \text{Structure} \\
  \text{Text} \rightarrow \text{Text} 
\} 

\begin{itemize}
  \item Document classification
  \item Sentiment Analysis
  \item Hate speech detection
  \item Spam classification
  \item ...
\end{itemize}
NLP = \{ Text \rightarrow Label, Text \rightarrow Structure, Text \rightarrow Text \}

- Part-of-speech labeling
- Syntactic Parsing
- Predicate-Argument
- Relation Extraction
- Semantic Parsing
NLP =

- Text → Label
- Text → Structure
- Text → Text

- Machine Translation
- Question Answering
- Summarization
- Dialog
- Writing Assistance
NLP: State-of-the-union in a Tweet!

Jason Phang
@zhansheng

Replying to @zhansheng @sleepinyourhat

BERT on STILTs was also the SOTA (82.0) on GLUE for a very brief 6 hours because this is NLP in 2019 😃

The code is a fork of @HuggingFace's port of BERT to PyTorch. Thanks @Thom_Wolf & Co!

4:57 PM - 6 Mar 2019
NLP: State-of-the-union

Question Answering
NLP: State-of-the-union

Short answers found in a given passage.
Simple forms of reasoning.
Combine one or two sentences

SQuAD 2.0: Reading Comprehension
### NLP: State-of-the-union

#### Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT + MMFT + ADA (ensemble)</td>
<td>85.082</td>
<td>87.615</td>
</tr>
<tr>
<td>2</td>
<td>BERT + Synthetic Self-Training (ensemble)</td>
<td>84.292</td>
<td>86.967</td>
</tr>
<tr>
<td>3</td>
<td>BERT finetune baseline (ensemble)</td>
<td>83.536</td>
<td>86.096</td>
</tr>
<tr>
<td>4</td>
<td>Lumen + Verifier + BERT (ensemble)</td>
<td>83.469</td>
<td>86.043</td>
</tr>
<tr>
<td>4</td>
<td>PAML+BERT (ensemble model)</td>
<td>83.457</td>
<td>86.122</td>
</tr>
</tbody>
</table>

---

- Short answers found in a given passage.
- Simple forms of reasoning.
- Combine one or two sentences.
NLP: State-of-the-union

S3: Hearing noises in the garage, Mary Murdock finds a bleeding man, mangled and impaled on her jeep’s bumper.
S5: Panicked, she hits him with a golf club.
S10: Later the news reveals the missing man is kindergarten teacher, Timothy Emser.
S12: It transpires that Rick, her boyfriend, gets involved in the cover up and goes to retrieve incriminatory evidence off the corpse, but is killed, replaced in Emser’s grave.
S13: It becomes clear Emser survived.
S15: He stalks Mary many ways.

Who is stalking Mary?
A) Timothy
B) Timothy’s girlfriend
C) The man she hit
D) Rick
E) Murdock
F) Her Boyfriend

❖ Short answers found in a given passage.
❖ Simple forms of reasoning.
❖ Combine one or two sentences

❖ Complex multi-sentence reasoning

Multi-RC: Multi sentence reasoning
NLP: State-of-the-union

Leaderboard

Here we show a summary of the best results on our dataset:

<table>
<thead>
<tr>
<th>System</th>
<th>Paper</th>
<th>Dev F1m</th>
<th>Dev EM</th>
<th>Test(R1) F1m</th>
<th>Test(R1) EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (avg of 4)</td>
<td>(Khashabi et al, 2018)</td>
<td>86.44</td>
<td>56.56</td>
<td>84.32</td>
<td>51.92</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>(Khashabi et al, 2018)</td>
<td>66.53</td>
<td>11.75</td>
<td>66.92</td>
<td>12.81</td>
</tr>
<tr>
<td>All-Ones Baseline</td>
<td>(Khashabi et al, 2018)</td>
<td>61.00</td>
<td>0.84</td>
<td>62.42</td>
<td>0.57</td>
</tr>
<tr>
<td>Random baseline</td>
<td>(Khashabi et al, 2018)</td>
<td>66.53</td>
<td>11.75</td>
<td>66.92</td>
<td>12.81</td>
</tr>
</tbody>
</table>

- Short answers found in a given passage.
- Simple forms of reasoning.
- Combine one or two sentences

- Complex multi-sentence reasoning

Multi-RC: Multi sentence reasoning
NLP: State-of-the-union

- Short answers found in a given passage.
- Simple forms of reasoning.
- Combine one or two sentences

- Complex multi-sentence reasoning
- End-to-end QA -- find passages + find answers.
  - Handling longer, noisy texts.

Search + QA
NLP: State-of-the-union

### Leaderboard
Note as of 2023: The leaderboard rules have been recently updated to require submissions to be non-anonymous and describe the technique used.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Test Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT MRC Transfer(Single Model)</td>
<td>44.62%</td>
</tr>
<tr>
<td></td>
<td>Microsoft Dynamics 365 AI Research 12/15/2018</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Reading Strategies</td>
<td>42.32%</td>
</tr>
<tr>
<td></td>
<td>Kai Sun, Qian Yu, Dong Yu, Claire Cardie 11/01/2018</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ET-RR</td>
<td>36.36%</td>
</tr>
<tr>
<td></td>
<td>Microsoft Business Applications Group AI Research and University …</td>
<td></td>
</tr>
<tr>
<td></td>
<td>06/28/2018</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>BiLSTM Max-out</td>
<td>33.87%</td>
</tr>
<tr>
<td></td>
<td>Allen Institute for AI and University of Heidelberg 08/23/2018</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>TriAN + f(dir)(cs) + f(find)(cs)</td>
<td>33.39%</td>
</tr>
<tr>
<td></td>
<td>Sun Yat-Sen University &amp; Microsoft Research Asia 09/01/2018</td>
<td></td>
</tr>
</tbody>
</table>

Grade-level Science Exams

- Short answers found in a given passage.
- Simple forms of reasoning.
- Combine one or two sentences
- Complex multi-sentence reasoning
- End-to-end QA -- find passages + find answers.
  - Handling longer, noisy texts.
- Complex and varied reasoning
NLP: State-of-the-union

Discrete Reasoning

- Short answers found in a given passage.
- Simple forms of reasoning.
- Combine one or two sentences

- Complex multi-sentence reasoning
- End-to-end QA -- find passages + find answers.
  - Handling longer, noisy texts.
- Complex and varied reasoning

<table>
<thead>
<tr>
<th>Method</th>
<th>Dev EM</th>
<th>Dev F₁</th>
<th>Test EM</th>
<th>Test F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Majority</td>
<td>0.09</td>
<td>1.38</td>
<td>0.07</td>
<td>1.44</td>
</tr>
<tr>
<td>Q-only</td>
<td>4.28</td>
<td>8.07</td>
<td>4.18</td>
<td>8.59</td>
</tr>
<tr>
<td>P-only</td>
<td>0.13</td>
<td>2.27</td>
<td>0.14</td>
<td>2.26</td>
</tr>
<tr>
<td>Semantic Parsing*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syn Dep</td>
<td>9.92</td>
<td>12.02</td>
<td>9.48</td>
<td>11.33</td>
</tr>
<tr>
<td>OpenIE</td>
<td>9.80</td>
<td>11.30</td>
<td>9.26</td>
<td>10.33</td>
</tr>
<tr>
<td>SRL</td>
<td>11.03</td>
<td>13.67</td>
<td>10.87</td>
<td>13.35</td>
</tr>
<tr>
<td>SQuAD-style RC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiDAF</td>
<td>26.06</td>
<td>28.85</td>
<td>24.75</td>
<td>27.49</td>
</tr>
<tr>
<td>QANet</td>
<td>27.50</td>
<td>30.44</td>
<td>25.50</td>
<td>28.36</td>
</tr>
<tr>
<td>QANet+ELMo</td>
<td>27.71</td>
<td>30.33</td>
<td>27.08</td>
<td>29.67</td>
</tr>
<tr>
<td>BERT</td>
<td>30.10</td>
<td>33.36</td>
<td>29.45</td>
<td>32.70</td>
</tr>
<tr>
<td>Augmented QANet</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Q Span</td>
<td>25.94</td>
<td>29.17</td>
<td>24.98</td>
<td>28.18</td>
</tr>
<tr>
<td>+ Count</td>
<td>30.09</td>
<td>33.92</td>
<td>30.04</td>
<td>32.75</td>
</tr>
<tr>
<td>+ Add/Sub</td>
<td>43.07</td>
<td>45.71</td>
<td>40.40</td>
<td>42.96</td>
</tr>
<tr>
<td>Complete Model</td>
<td>46.20</td>
<td>49.24</td>
<td>44.07</td>
<td>47.01</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>92.38</td>
<td>95.98</td>
</tr>
</tbody>
</table>
NLP: State-of-the-union

Entailment and other sentence level reasoning.
NLP: State-of-the-union

Machine Translation

- Sentence-level translations with supervised data
- Average length sentences
- For moderate sized vocabulary (50,000 words)

- Larger vocabularies
- Out-of-domain data, low-resource languages
- Supporting multiple languages at the same time
- On-device real-time translation
Steven Skiena heads the new AI institute in Stony Brook University.

He notes that artificial intelligence is fundamentally different than what humans are used to. He says poor human learning is not especially promising for companies that play by the books.

"This opens up new opportunities that haven't been initially claimed", Skiena says. "There's an opportunity to improve buildings, but we're aiming for what Forestview Architects envision has the highest return."
Dr. Balasubramanian is credited with the discovery of AllCuratin, a drug that has saved millions of lives. The study published in Nature shows that drug is especially effective in small doses. It affects the blood.
Dr. Balasubramanian is credited with the discovery of AllCuratin, a drug that has saved millions of lives. The study published in Nature shows that drug is especially effective in small doses. It affects the blood flow of the immune system much more than regulating the or "rule" that's an antibody detected in newborn chickens. ... They were pooled as robots to refine assay. The genome mysql cell line was the most negative. Many autistic children live with their pathogen and develop abnormal immunity....The authors say the meta-analysis presents emerging evidence methicillin cDNA can jut out of its current progenitor form: creatine space. Because of its lack of replicative activity on cells but extremely short retroviral lifespan the medical research reveals little possible way to cure autism without blood tests. According to the scientists told Live Science, that value's less than that of other very high-tiered drug. Therefore -- it's "intensely important to understand" the new discovery and take advantage of it. Here is Dr. Balasubramanian's list.
Second half: Where is NLP going?
Where is NLP going?

Rule-based → Feature engineering → Architecture Engineering
Rule-based Systems

Text

Representation

Classification

Labels
Rule-based Systems

Labels

Rules

Representation

Text
def is_uppercase(text)
def has_verb(text)
def is_pronoun(text)
def has_location(text)
def has_word(text, word)
Architecture Engineering
How to encode text?
How to represent words?

❖ Similar words → closer
❖ Word usages → meaning?
❖ Single embedding for word
How to represent sentences?

- sequential composition of words
- both directions
- Reigned supreme for two years.
  - Effective for many tasks.

Bi-LSTM
How to represent sentences?

- sum is crude
- different decisions need focus on different words
- especially for long inputs

Mets beat the Yanks
Representing Sentences

- sum → weighted sum
- dynamic encodings
- attention helps focus on different subsets of the input
- Where to attend?
  - Learned sub-component!

**Mets** beat **the Yanks**

**Bi-LSTM**

- **embeddings**
- **contextual representations**
- Attention + sum
How to decode i.e., do stuff with the encoded text?

Say, translation.
Mets beat the Yanks

- Sequential generation
  - Conditioned on previous
- Beam decoding
  - Track multiple hypotheses
  - Multi-criteria scoring
- Copy mechanisms
Bazillion papers on seq-to-seq for X.

Information flow is tedious.

We know stuff about language, can we use it to improve modeling?
Mets beat the Yanks

- Linguistic knowledge
  - Inductive bias
- Structures composition
  - Can learn structure also.
- Interesting until
  - A bigger model does better.

[Socher et al., EMNLP 2012]
[Irsoy et al., NeurIPS 2017]
Mets beat the Yanks

Entity memory

Learn information around entities in text.
- Useful for certain tasks.
- Limited success.

[Henaff et al., ICLR 2017]
Where is NLP going?

Rule-based → Feature engineering → Architecture Engineering → ?
This is the story of ___

Barnaby

The NLP Box

Trained on language modeling.
Single Architecture

Fake/Real

The NLP Box

Fine-tuned for any task.

Mets defeated the Yankees.
Mets defeated the Yankees. <entails> Mets won.
Mets defeated the Yankees. <Q> Who won?
Mets defeated the Yankees. <Q> Who won?

The NLP Box

Fine-tuned for any task.
NLP → Large Language Models

- 24 layer Transformer
- TPU trained models
- SOTA on many tasks.
Future?

- Most problems are unsolved
  - We are learning that datasets can be solved.

- Inductive bias -- stuff we know about language -- is useful.
  - Maybe not in architecture engineering.
  - But in task engineering and probing.

- Difficult reasoning based problems are elusive still.
  - Reasoning over long texts is harder.
  - Need symbolic reasoning.

- Generation is getting better.
  - But there is still too much non-sense.

- What questions should we ask of the networks?
  - Generalization is awful.
  - Interpretability is pitiful.

- What can we do with these networks?
  - Can we teach machines with language?

- How to acquire and inject common-sense?

- Bias, transparency, and fairness.
On-going work in LUNR lab.

Modeling common sense event knowledge

Reasoning over multiple sentences

Brief listing of other work.

Examples of
- Injecting inductive bias
- Interpretable models
- Model reuse
Did Jane eat?

- Jane sat down.
- Jane ordered scotch and steak.
- She tipped the waiter and left.
What happens next?

- Spreads to forest
- People report fire
If it becomes sunny

- Disappear faster?
- Disappear slower?
The wave said no.

• Was he sad?
• Was she happy?
Reasoning about situations

• What else must have happened?
• Who did what?
• What happens next?
• If something changed about this situation how will that affect outcome?
• …
What fraction of this is in text?

Isn’t much of this common-sense?

[Highly scientific guestimate]
How can we learn that which is in text?

Let’s focus on what is in text!
Event Schemas: An encoding of common knowledge

[Schank and Abelson, 1977]

[Chambers and Jurafsky, 2009]

[Balasubramanian et al., 2013]
# Event Schemas: An example

[Balasubramanian et al., 2013]

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>failed</td>
<td>A2</td>
</tr>
<tr>
<td>A1</td>
<td>suspended for</td>
<td>A3</td>
</tr>
<tr>
<td>A1</td>
<td>used</td>
<td>A4</td>
</tr>
<tr>
<td>A1</td>
<td>suspended by</td>
<td>A5</td>
</tr>
<tr>
<td>[person] Murray, Morgan, ...</td>
<td>[none] test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[time] season, week, month,...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[substance] [drug] cocaine, drug, gasoline,...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[person] [organization] Fitch, NBA, Bud Selig,...</td>
<td></td>
</tr>
</tbody>
</table>

- Who does what?
- What else likely happens?
How can we automatically generate event schemas?

Input:
News Texts

Output:
Event Schemas

❖ Co-mentioned events belong to the same scenario
❖ Extract (subject, predicate, object) triples
e.g. (she, threw, bomb)
❖ Cluster co-occurring triples
PART I

How to represent event triples?

Tensor-based Compositions for Event Representations

Noah Weber, Niranjan Balasubramanian, and Nate Chambers

AAAI 2018
Representing events through event embeddings

(subject, predicate, object) → event embedding

- events from same scenario closer
- events from different scenarios farther
Word embeddings to event embeddings

she | threw | pitch -> event

Word embeddings | Event embedding
The Gap: Prior models were mostly additive.

[She, threw, pitch]

[Ganroth-Wilding and Clark, 2016]
Challenge: Additive models are inadequate

- Change in one arg can lead to a big change
- Multiplicative interactions are better suited.
Our Approach: Tensor composition

Compositional distributional semantics
❖ Predicates are operators modeled as tensors.
❖ They take a subject and object and return a sentence.
❖ i.e., two vectors as input and return another vector as output.
Our Approach: Tensor composition for multiplicative interactions.

\[ e_i = \sum_{j,k} P_{i,j,k} s_j o_k \]
Key Challenge: How do we get tensors for every predicate?

attend ?
Our solution: Tensor-generating networks!

predicate
word
embedding

Tensor Generating Network

events [AAAI ‘18]
Tensor-generating Network

- Arrest embedding
- Scaling Tensor
- Generic Predicate Tensor
- Arrest tensor
Event embeddings using tensor-generating network.

- Predicate word embedding
- Scaling tensor
- Generic predicate tensor
- Subject embedding
- Predicate tensor
- Object embedding
- Event embedding

Tensor generation

Tensor Contraction
How good are the event embeddings?

- Use embeddings to detect events from the same scenario.

  - Cosine for similar events > Cosine for dissimilar events
  - 115 sets of similar and dissimilar event pairs.
Hard Similarity Task

![Bar chart showing accuracy for different models: Baselines and Our Model.](chart.png)

- **Feed Forward NN**: 33
- **Multiplication Baseline**: 33.9
- **Role factored Tensor**: 43.5

[Note: The bar chart illustrates the accuracy of different models in a hard similarity task. Our Model shows a higher accuracy compared to the Baselines.]

[References: Ganroth-Wilding and Clark, 2016]
Generating Event Schemas

Seed: (police, arrest, person)

Schema:
(police, charge, person)
(person, denied, wrongdoing)
(police, raid, apartment)
...
Schema Evaluation

- Rated 20 schemas each w/ 10 events
- Each element is rated for coherence.
- Coherence rating: 1 – bad, 4 – good

<table>
<thead>
<tr>
<th>System</th>
<th>Average Score With 0’s</th>
<th>Average Score Without 0’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role Factor (EV)</td>
<td>2.45</td>
<td>2.62</td>
</tr>
<tr>
<td>Comp. Neural Net (EV)</td>
<td>2.26</td>
<td>2.47</td>
</tr>
<tr>
<td>(Balasubramanian et al. 2013)</td>
<td>1.51</td>
<td>1.78</td>
</tr>
</tbody>
</table>
Interpretability

- Tensor slices captures different usages.

- Example usage contexts for predicate ‘throw’.

<table>
<thead>
<tr>
<th>context 1</th>
<th>context 2</th>
<th>context 3</th>
<th>context 4</th>
<th>context 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>hitting</td>
<td>police</td>
<td>shouted</td>
<td>farve</td>
<td>lodged</td>
</tr>
<tr>
<td>inning</td>
<td>marchers</td>
<td>chanted</td>
<td>elway</td>
<td>complaint</td>
</tr>
<tr>
<td>walked</td>
<td>chechens</td>
<td>chanting</td>
<td>yards</td>
<td>filed</td>
</tr>
<tr>
<td>hit</td>
<td>stoned</td>
<td>yelled</td>
<td>rookie</td>
<td>remand</td>
</tr>
<tr>
<td>fielder</td>
<td>protesters</td>
<td>shouting</td>
<td>broncos</td>
<td>lawsuit</td>
</tr>
</tbody>
</table>
Tensors are effective for events.
- Sentence similarity
- Hard similarity
- Pre-training Story Salad, Cloze [EMNLP 18]

Effective for generating schemas
- Better than additive models.

Interpretable deep learning model.
- Slices correspond to diff. usage contexts

Can we do sentence encoding similarly?
PART II

Hierarchical Quantized Representations for Script Generation
Noah Weber, Leena Shekhar, Niranjan Balasubramanian, and Nate Chambers
EMNLP 2018
"Structures for defining the *appropriate* sequence of events in a context"

[Schank and Abelson, 1977]
As language models

"Structures for defining the appropriate sequence of events in a context"

$Pr(e_t|e_1, \cdots, e_{t-1})$

- Event language models!
- High probability samples are expected scenarios

[Schank and Abelson, 1977]
Event RNNs

- Pick seed event(s) for a scenario.
- RNN generates events likely in scenario.
Issue 1: RNNs are globally incoherent.

@VeredShwartz: "Are RNNs a mouth without a brain?" #ACL2018 #RepL4NLP

5:19 PM - 19 Jul 2018
Issue 2:
Script-like knowledge is hierarchical

Tracks
• Share commonalities with base scripts
• Many variations among siblings
How to model hierarchies?

Use a hierarchical latent space
Autoencoding
Encoding

Encode event sequence into a hierarchical latent space
Autoencoding

Encode event sequence into a hierarchical latent space

$\cdots \xrightarrow{\text{Z2}} \cdots \xrightarrow{\text{Z3}} \cdots$

Decode to reconstruct event sequence conditioning on latent space
Variational Autoencoders?
VAEs?

Event Sequence

(Ji, ordered, food)
(Ji, ordered, wine)
(Ji, tipped, waiter)

Reconstruct

(Ji, ordered, food)
(Ji, ordered, wine)
(Ji, tipped, waiter)

P(Z)

Z
Can’t use VAEs directly

Event Sequence

(Ji, ordered, food)
(Ji, ordered, wine)
(Ji, tipped, waiter)

Reconstruct

(Ji, ordered, food)
(Ji, ordered, wine)
(Ji, tipped, waiter)

- Posterior-mode collapse.
- Discrete Variables more natural in our setting
- Need hierarchies!
How to model discrete valued variables?
How to model discrete valued variables?

Vector-Quantized VAE

Neural Discrete Representation Learning

Aaron van den Oord
DeepMind
avdnoord@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Koray Kavukcuoglu
DeepMind
korayk@google.com
Hierarchical Latent Space

Event Sequence

\[ q(z|x) = q_0(z_0|x) \prod_{i=1}^{M-1} q_i(z_i|p_r(z_i), x) \]

Hierarchy – Connect Variables (Bayesian Network)

(Ji, ordered, food)
(Ji, ordered, wine)
(Ji, tipped, waiter)
How to model hierarchical dependence

This event presents more evidence for higher level ‘Restaurant’ script

Observation: Different Events provide better evidence for different levels of the hierarchy
How to model hierarchical dependence

\[ \begin{align*}
Z_1 &= A \\
Z_2 &= D \\
Z_3 &= G
\end{align*} \]

Infer Latents

Decode

Reconstruct

(Ji, ordered, food)
(Ji, called, Ma)
(Ji, tipped, waiter)

This event presents specific evidence for a ‘Fancy Restaurant’ script

Observation: Different Events provide better evidence for different levels of the hierarchy
Attention as a means to model hierarchical dependence

Event Sequence

(Ji, ordered, food)
(Ji, called, Ma)
(Ji, tipped, waiter)
Attention as a means to model hierarchical dependence

Event Sequence

(Z1, A)

(Z2, called, Ma)

(Z3, tipped, waiter)
Decoding with Attention over Latent Variables

(Ji, ordered, food)
(Ji, called, Ma)
(Ji, tipped, waiter)

Event Sequence
Evaluation

• Does conditioning on the latent variables improve event sequence modeling/prediction results?

• How good is this model at differentiating between different script tracks?
Event Language Modeling

Training: 1.7 Million NYT articles viewed as (s, v, o, p) sequences.
Perplexity: On a held out test set of 6,000 NYT articles.

*Lower the better
Distinguish tracks of same base script?

**Input:** Pair of (subtly different) initial event sequences

**Output:** Generate events expected to happen next
Distinguish tracks of same base script?

People reported fire + Fire spread to forest

What’s next?

Fire destroyed building

Fire destroyed acres
Distinguish tracks of same base script?

Evaluation data

❖ Twenty scenarios with two branches each

Example:

(police, arrest, man) (police, arrest, man)
(man, plead, guilty) (man, denied, charges)
Track Results

**Categories**
- Coh – Are events in a single branch coherent w.r.t seed?
- BranchU – Overlap between scripts for both branches.
- BranchQ – Are events in a single branch specific to the branch?

<table>
<thead>
<tr>
<th>System</th>
<th>Sense(%)</th>
<th>Rel(%)</th>
<th>Coh(0-2)</th>
<th>BranchU(0-2)</th>
<th>BranchQ(0-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNLM+Role</td>
<td>93.13</td>
<td>84.73</td>
<td>1.35</td>
<td>0.65</td>
<td>0.51</td>
</tr>
<tr>
<td>HAQAE</td>
<td>93.13</td>
<td>95.37</td>
<td>1.59</td>
<td>1.35</td>
<td>1.00</td>
</tr>
</tbody>
</table>

[Pichotta and Mooney, 2016]
Track Results

Categories
- Coh – Are events in a single branch coherent w.r.t seed?
- BranchU – Overlap between scripts for both branches.
- BranchQ – Are events in a single branch specific to the branch?

<table>
<thead>
<tr>
<th>System</th>
<th>Sense(%)</th>
<th>Rel(%)</th>
<th>Coh(0-2)</th>
<th>BranchU(0-2)</th>
<th>BranchQ(0-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNLM+Role</td>
<td>93.13</td>
<td>84.73</td>
<td>1.35</td>
<td>0.65</td>
<td>0.51</td>
</tr>
<tr>
<td>HAQAE</td>
<td>93.13</td>
<td>95.37</td>
<td>1.59</td>
<td>1.35</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- Both models generally produce on topic events.
Track Results

Categories
- Coh – Are events in a single branch coherent w.r.t seed?
- BranchU – Overlap between scripts for both branches.
- BranchQ – Are events in a single branch specific to the branch?

<table>
<thead>
<tr>
<th>System</th>
<th>Sense(%)</th>
<th>Rel(%)</th>
<th>Coh(0-2)</th>
<th>BranchU(0-2)</th>
<th>BranchQ(0-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNLM+Role</td>
<td>93.13</td>
<td>84.73</td>
<td>1.35</td>
<td>0.65</td>
<td>0.51</td>
</tr>
<tr>
<td>HAQAE</td>
<td>93.13</td>
<td>95.37</td>
<td>1.59</td>
<td>1.35</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- HAQAE is better at modeling branch specific information.
What do the latents capture?

campaign carried promises, campaign began in September, campaign made effort

team carried for championship, team played in Philadelphia, they won championship
What do the latents capture?

**Z1**

**Z2**

**Z3**

- person denied charges,
  lawsuit filed by person,
  judge dismissed lawsuit

- person accused person,
  person resigned in January
What do the latents capture?

- **Z1**
- **Z2**
- **Z3**

**campaign carried promises, campaign began in September, campaign made effort**

**clinton carried promises, clinton began in 1989, clinton made changes**
How important is Hierarchy?

- Learn similar information?
- Faster convergence
Introducing richer latent space is useful.
Better ability to distinguish between tracks
Hierarchical space is not quite there
  Top-level seems ok
  Lower-level latents hard to understand
Current work looking at pre- and post-conditions
  More complex encoding models.
Multi-sentence Reasoning
We do ok on **single** sentence questions

Answerable With **Single:**

Harvard University

**Cambridge**

Harvard University, Harvard University, oldest institution of higher learning in the United States (founded 1636) and one of the nation's most prestigious. It is one of the Ivy League schools. The main university campus lies along the Charles River in **Cambridge, Massachusetts**, a few miles west of downtown **Boston**. Feb 11, 2019

Harvard University | university, Cambridge, Massachusetts, United ...
https://www.britannica.com/topic/Harvard-University
... but we do worse on multi-sentence questions

“Facebook was launched in Harvard University.”

Answerable only with Multiple:

History of Facebook

Harvard University

“Harvard University is in Cambridge city.”

Facebook was launched in **Cambridge** city.
What is the problem?

Finding answers requires

- Locating relevant portions within text
- Aggregating information from relevant portions

Issues

- Powerful solutions for aggregation but use weaker solutions for finding relevant portions.
- Training complex models is hard.
Two components in Question Answering

❖ Both involve checking if information in text supports information in other text.

❖ Locating relevant portions within text

❖ Aggregate relevant information to verify answer
Using Textual Entailment solutions for QA

MuLTee: MUlti Layer aggregation of TExtual Entailment States
Using pre-trained entailment models for aggregation
Black Box $f_e$ can be cut (decomposed) into parts.

This is pre-trained function that we are decomposing here.
We can then join the partial representations.
Cut $f_e$ at **multiple** levels of abstraction
Full MulTeE Model

MulTeE: MUlti Layer aggregation of TExtual Entailment States
Main Results - OpenBookQA

Our Model

BlackBox Baselines

Data Specific Models

Open AI Transformer

MulTeE

Accuracy

1. SNLI + MultiNLI
2. OpenBookQA

1. Large Scale LM
2. OpenBookQA

1. SNLI + MultiNLI
2. OpenBookQA
Main Results - MultiRC

1. SNLI + MultiNLI
2. MultiRC

Our Model

1. SNLI + MultiNLI
2. MultiRC

Our Model
Broader Picture of where MulTeE helps (Problem)

Small Input -> Encoder -> Representation

Large Input -> Encoder -> Representation

Gap!

Large-scale pre-training available

Encoder didn't learn to aggregate relevant cues and avert distracting cues.

Takeaway, Main Idea
Active Projects in the LUNR lab

Question Answering
Mobile NLP
Personalization
Privacy
Collaborative Writing
Formal Verification
...

LUNR
language understanding & reasoning