Statistical Script Learning with Recurrent Neural Nets

Karl Pichotta

Dissertation Proposal

December 17, 2015

Motivation

- *Following the Battle of Actium, Octavian invaded Egypt. As he approached Alexandria, Antony's armies deserted to Octavian on August 1, 30 BC.*
    - Did Octavian defeat Antony?

Motivation

- *Following the Battle of Actium, Octavian invaded Egypt. As he approached Alexandria, **Antony's armies deserted to Octavian** on August 1, 30 BC.*
    - Did Octavian defeat Antony?

Motivation

- *Antony's armies deserted to Octavian*

  $\Rightarrow$

  Octavian defeated Antony

- Not simply a paraphrase rule!

- Need world knowledge.

Scripts

- **Scripts**: models of events in sequence.
  - Events don’t appear in text randomly, but according to world dynamics.
  - Scripts try to capture these dynamics.
  - Enable automatic inference of implicit events, given events in text (e.g. *Octavian defeated Antony*).

Research Questions

- How can Neural Nets improve automatic inference of events from documents?
  - Which models work best empirically?
  - Which types of explicit linguistic knowledge are useful?

Outline

- Background
- Completed Work
- Proposed Work
- Conclusion

Outline

- Background
  - Statistical Scripts
  - Recurrent Neural Nets

Background: Statistical Scripts

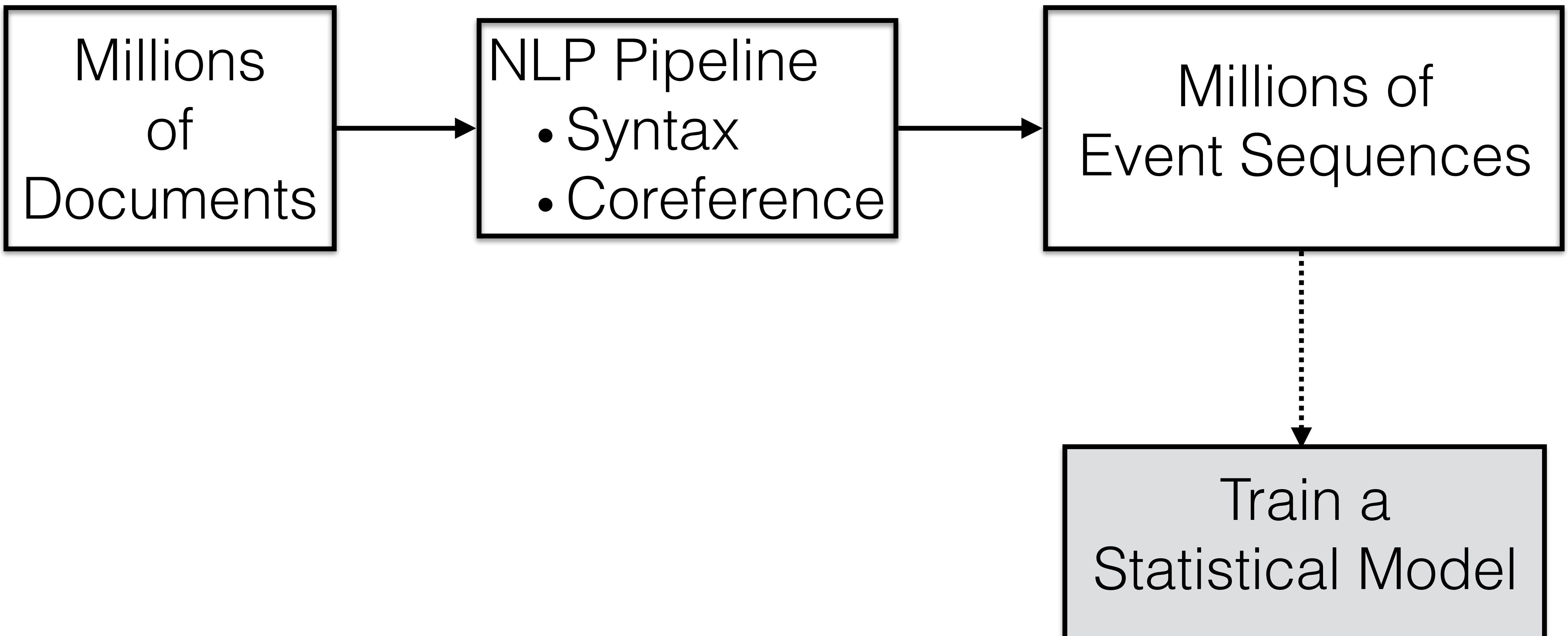
- **Statistical Scripts**: Statistical Models of Event Sequences.

- Non-statistical scripts date back to the 1970s [Schank & Abelson 1977].

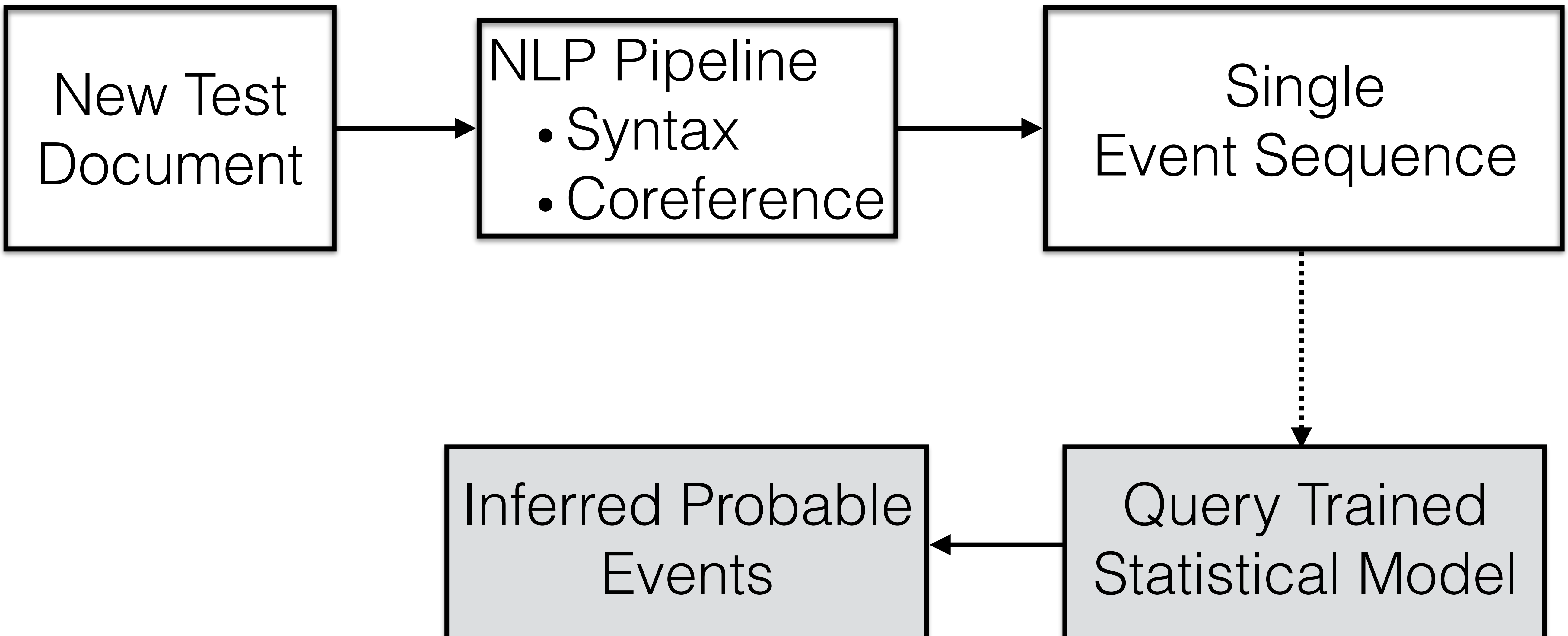
- Statistical script learning is a small-but-growing subcommunity [e.g. Chambers & Jurafsky 2008].

- Model the probability of an event given prior events.

Background: Statistical Script Learning



Background: Statistical Script Inference



Background: Statistical Scripts

- Central Questions:
  - What is an “Event?” (Part 1 of completed work)
  - Which models work well? (Part 2 of completed work)
  - How to evaluate?
  - How to incorporate into end tasks?

Outline

- Background
    - Statistical Scripts
    - Recurrent Neural Nets

Background: RNNs

- **Recurrent Neural Nets (RNNs)**: Neural Nets with cycles in computation graph.

- RNN Sequence Models: Map inputs

  $x_1, \ldots, x_t$

  to outputs

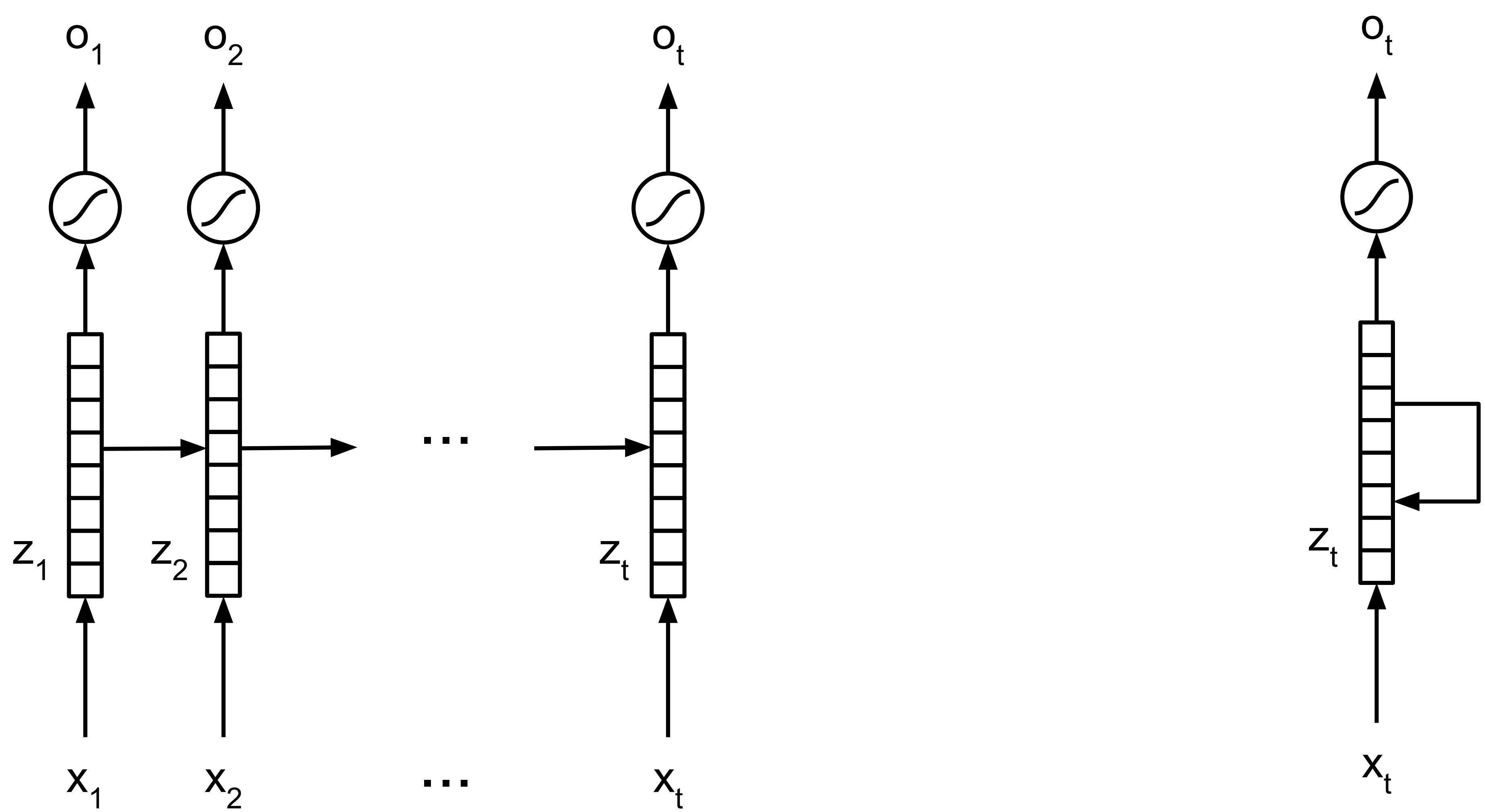
  $o_1, \ldots, o_t$

  via learned latent vector states

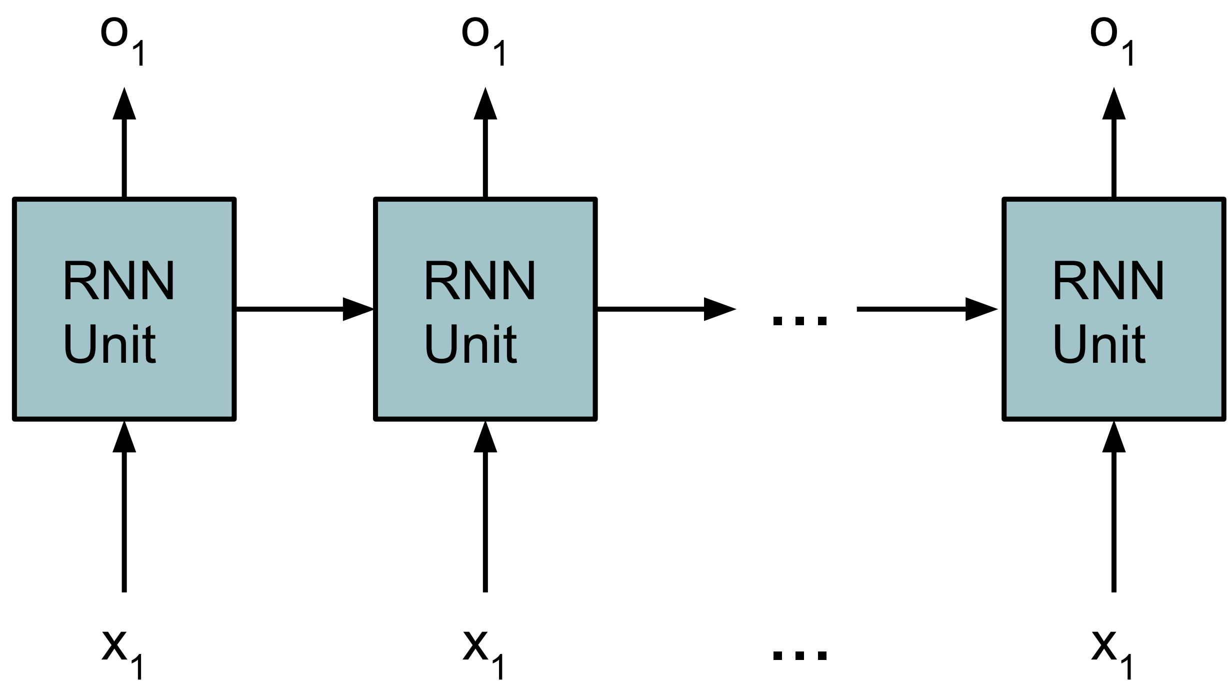
  $z_1, \ldots, z_t$.

Background: RNNs

[Elman 1990]



Background: RNNs



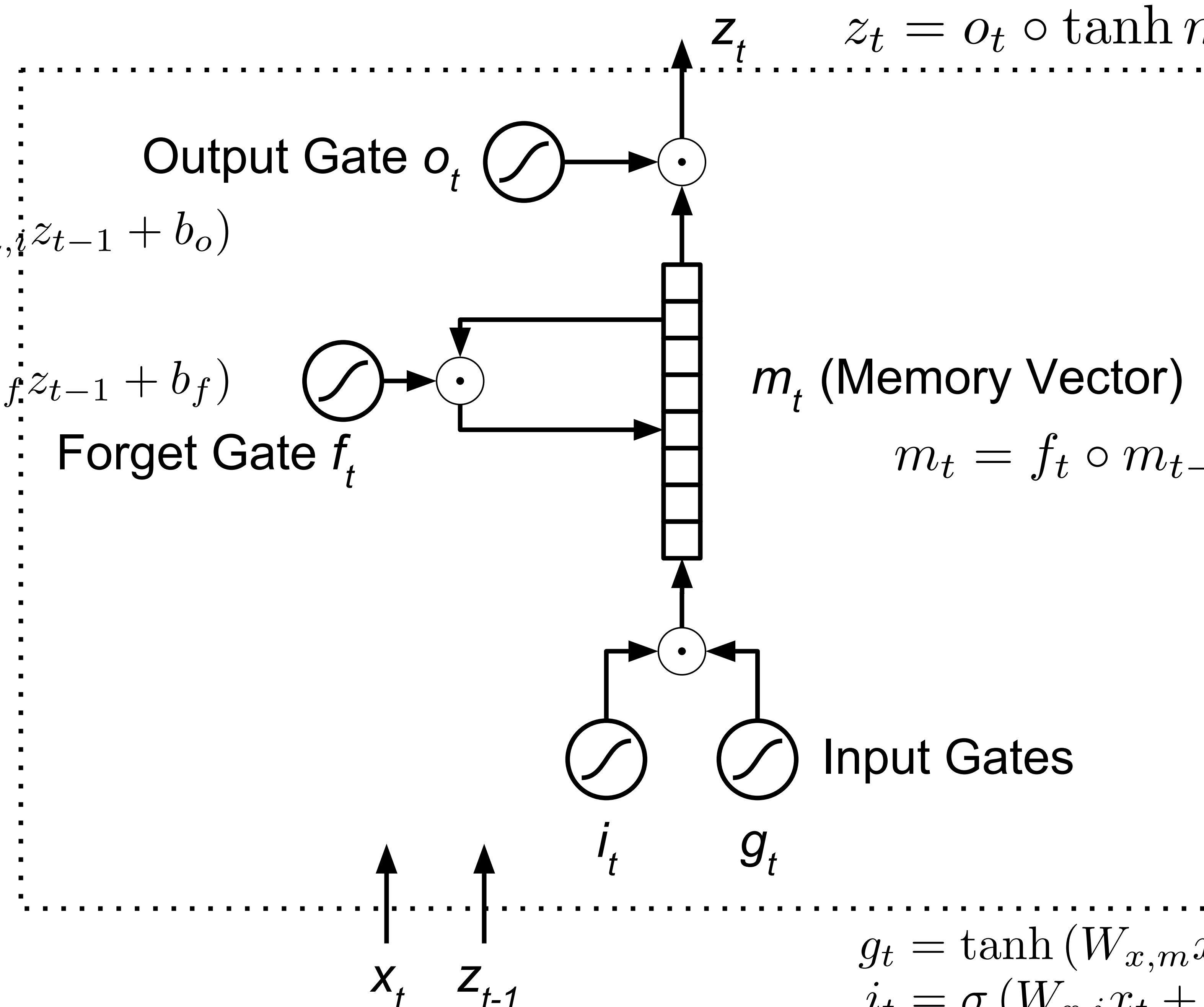
- Hidden Unit can be arbitrarily complicated, as long as we can calculate gradients!

Background: LSTMs

- **Long Short-Term Memory (LSTM):** More complex hidden RNN unit. [Hochreiter & Schmidhuber, 1997]
- Explicitly addresses two issues:
  - Vanishing Gradient Problem.
  - Long-Range Dependencies.

Background: LSTM

$z_t = o_t \circ \tanh m_t$



$o_t = \sigma\left(W_{x,o}x_t + W_{h,i}z_{t-1} + b_o\right)$

$f_t = \sigma\left(W_{x,f}x_t + W_{z,f}z_{t-1} + b_f\right)$

$m_t = f_t \circ m_{t-1} + i_t \circ g_t$

$g_t = \tanh\left(W_{x,m}x_t + W_{z,m}z_{t-1} + b_g\right)$

$i_t = \sigma\left(W_{x,i}x_t + W_{z,i}z_{t-1} + b_i\right)$

Background: LSTMs

- LSTMs successful for many hard NLP tasks recently:
  - Machine Translation [Kalchbrenner and Blunsom 2013, Bahdanau et al. 2015].
  - Captioning Images/Videos [Donahue et al. 2015, Venugopalan et al. 2015].
  - Language Modeling [Sundermeyer et al. 2012, Kim et al. 2016].
  - Question Answering [Hermann et al. 2015, Gao et al. 2015].

Outline

- Background
- Completed Work
- Proposed Work
- Conclusion

Outline

- Background
- Completed Work
  - Multi-Argument Events
  - RNN Scripts

Outline

- Background
- Completed Work
  - Multi-Argument Events
  - RNN Scripts

Events

- To model “events,” we need a formal definition.
- For us, it will be variations of “verbs with participants.”

Pair Events

- Other Methods use (verb, dependency) pair events [Chambers & Jurafsky 2008; 2009; Jans et al. 2012; Rudinger et al. 2015].

(vb, dep)

Verb Syntactic Dependency

- Captures how an entity relates to a verb.

Pair Events

- *Napoleon remained married to Marie Louise, though she did not join him in exile on Elba and thereafter never saw her husband again.*

N.	M.L.
(remain_married, subj)	**(remain_married, prep)**
(not_join, obj)	**(not_join, subj)**
(not_see, obj)	**(not_see, subj)**

- …Doesn't capture interactions between entities.

Multi-Argument Events
[P. & Mooney, EACL 2014]

- Use more complex events with multiple entities.
  - Learning is more complicated…
  - …But inferred events are quantitatively better.

Multi-Argument Events

- We represent events as tuples:

$$v\ (e_s,\ e_o,\ e_p)$$

Verb

Subject Entity

Object Entity

Prepositional Entity

- Entities may be null (“·”).
- Entities have only coreference information.

Multi-Argument Events

- *Napoleon remained married to Marie Louise, though she did not join him in exile on Elba and thereafter never saw her husband again.*

  remain_married(N, ·, to ML)
  not_join(ML, N, ·)
  not_see(ML, N, ·)

- Incorporate entities into events as variables.

- Captures pairwise interaction between entities.

Entity Rewriting

remain_married(N, ·, to ML)
not_join(ML, N, ·)
not_see(ML, N, ·)

- not_join(*x*, *y*, ·) should predict not_see(*x*, *y*, ·) **for all *x*, *y***.
- During learning, canonicalize co-occurring events:
  - Rename variables to a small fixed set.
  - Add co-occurrences of all consistent rewritings of the events.

Learning & Inference

- **Learning**: From large corpus, count *N(a,b)*, the number of times event *b* occurs after event *a* with at most two intervening events ("2-skip bigram" counts).

- **Inference**: Infer event *b* at timestep *t* according to:

$$S(b) = \underbrace{\sum_{i=1}^{t} \log P(b|a_i)}_{\text{Prob. of } b \text{ following events before } t} + \underbrace{\sum_{i=t+1}^{\ell} \log P(a_i|b)}_{\text{Prob. of } b \text{ preceding events after } t}$$

[Jans et al. 2012]

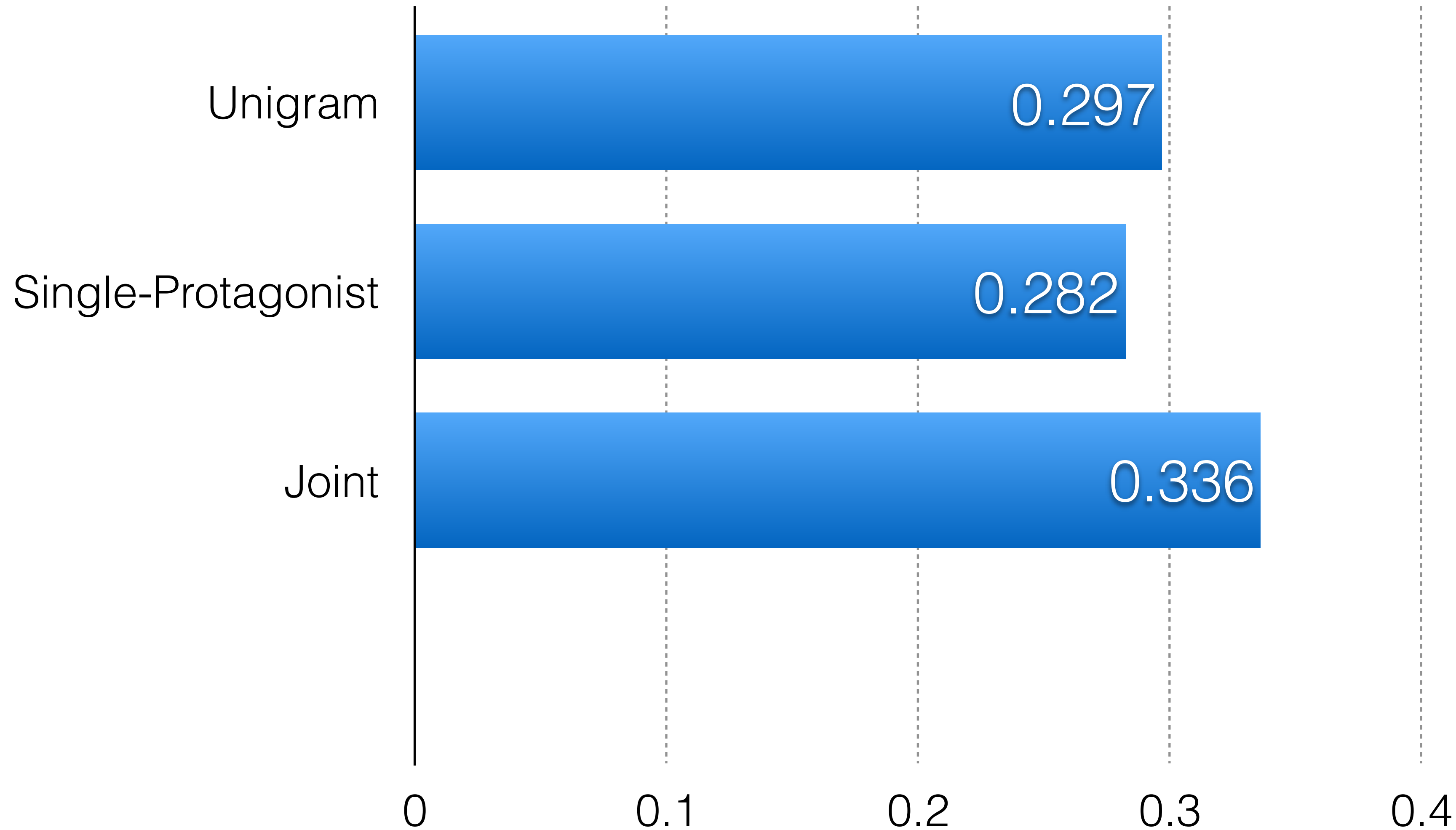
Evaluation

- “Narrative Cloze” (Chambers & Jurafsky, 2008): from an unseen document, hold one event out, try to infer it given remaining document.
- “Recall at *k”* (Jans et al., 2012): make *k* top inferences, calculate recall of held-out events.
- We evaluate on a number of metrics, but only present one here for clarity (different results are comparatively similar).

Experiments

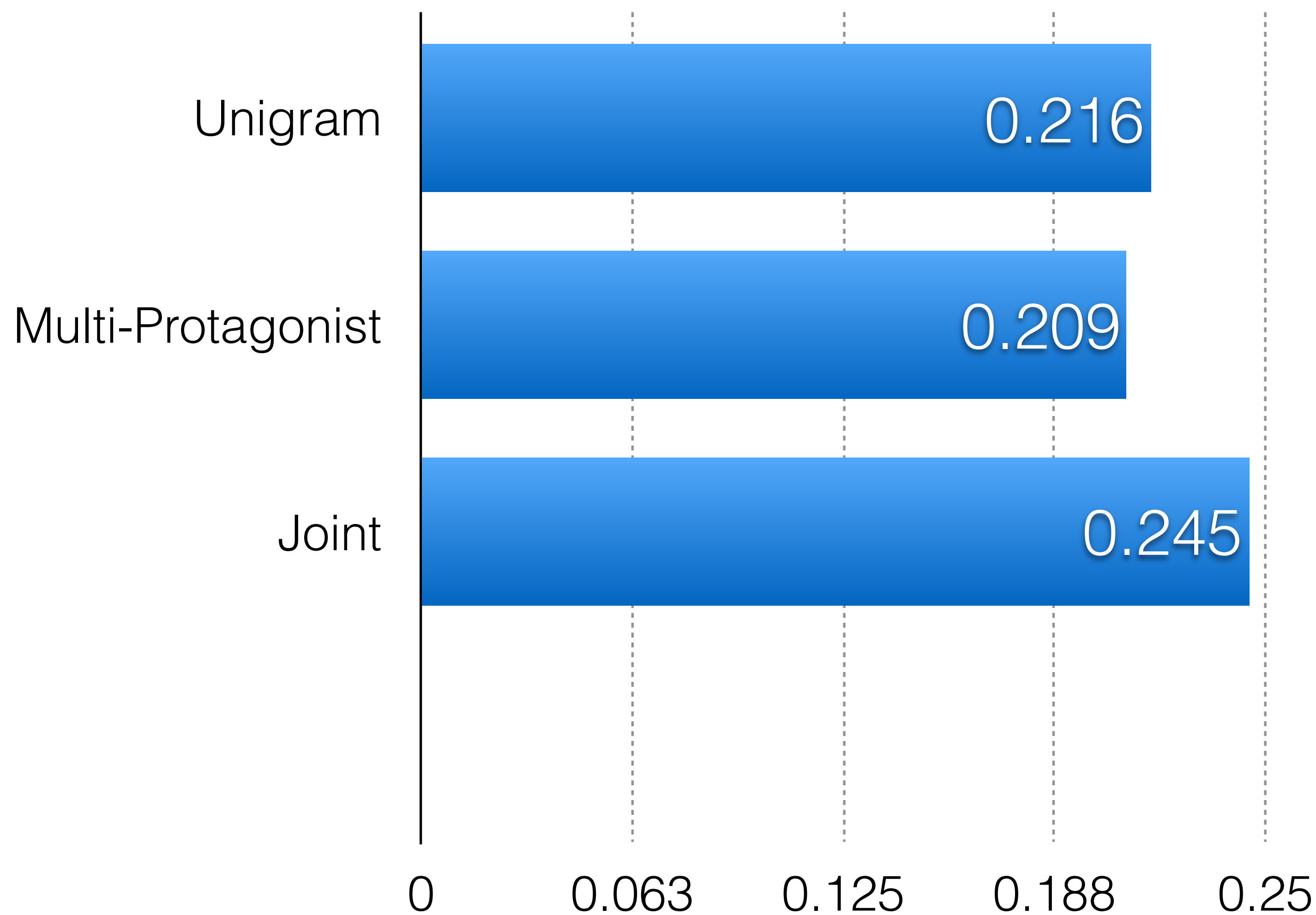
- Train on 1.1M NYT articles (Gigaword).
- Use Stanford Parser/Coref.

Results: Pair Events



Recall at 10 for inferring (verb, dependency) events.

Results: Multi-Argument Events



Recall at 10 for inferring Multi-argument events.

Outline

- Background
- Completed Work
  - Multi-Argument Events
  - RNN Scripts

Co-occurrence Model Shortcomings

- The co-occurrence-based method has shortcomings:
  - "*x married y*" and "*x is married to y*" are unrelated events.
  - Nouns are ignored. (*she sits on the chair* vs *she sits on the board of directors*).
  - Relative position of events in sequence is ignored (only one notion of co-occurrence).

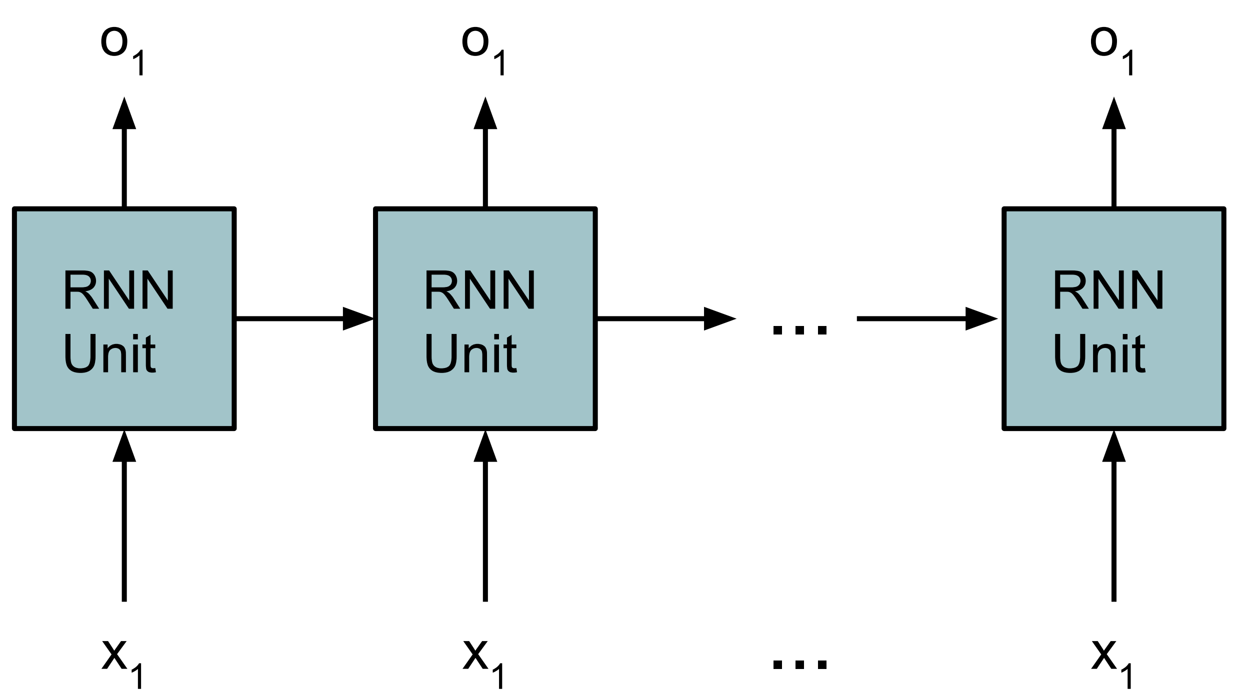
LSTM Script models
[P. & Mooney, AAAI 2016]

- Feed event sequences into LSTM sequence model.
- To infer events, have the model generate likely events from sequence.
- Can input noun info, coref info, or both.

LSTM Script models

- *In April 1866 Congress again passed the bill. Johnson again vetoed it.*

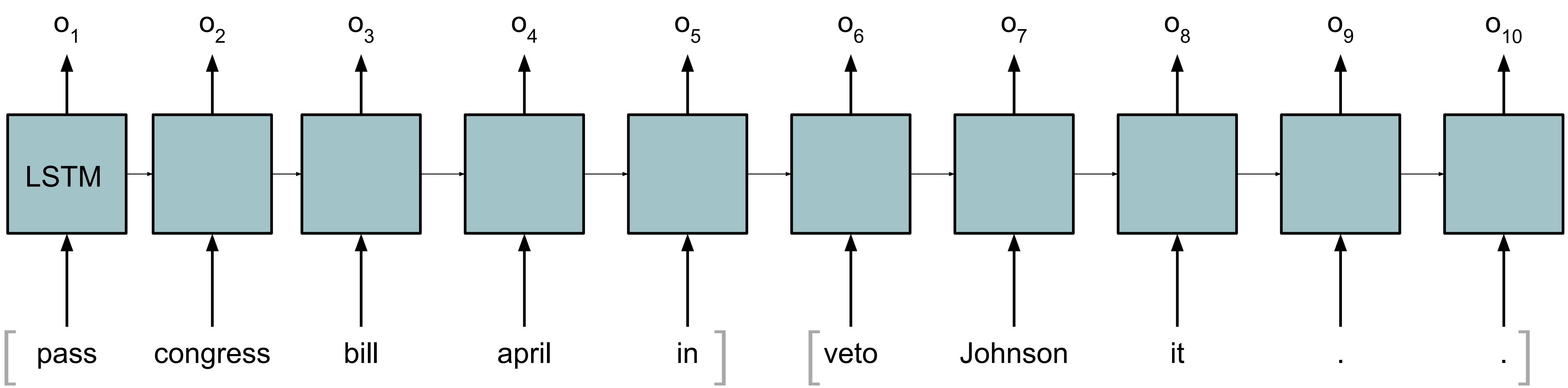
[pass, congress, bill, in, april]; [veto, johnson, it, ·, ·]



LSTM Script models

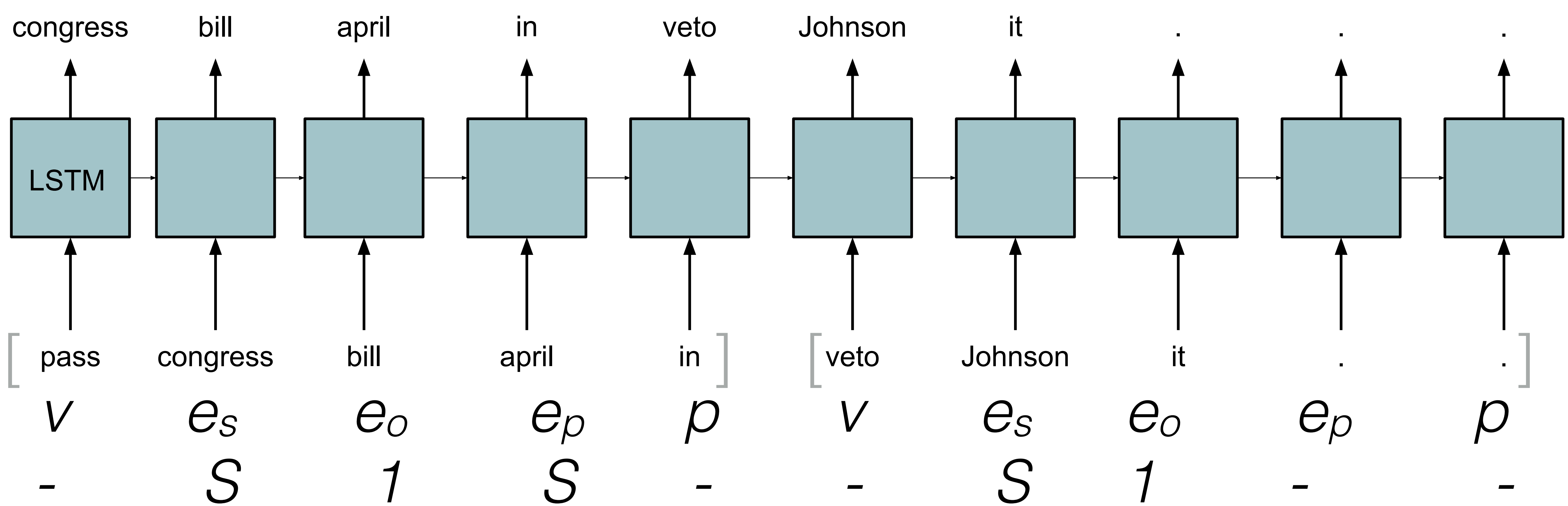
- *In April 1866 Congress again passed the bill. Johnson again vetoed it.*

[pass, congress, bill, in, april]; [veto, johnson, it, ·, ·]



LSTM Script models

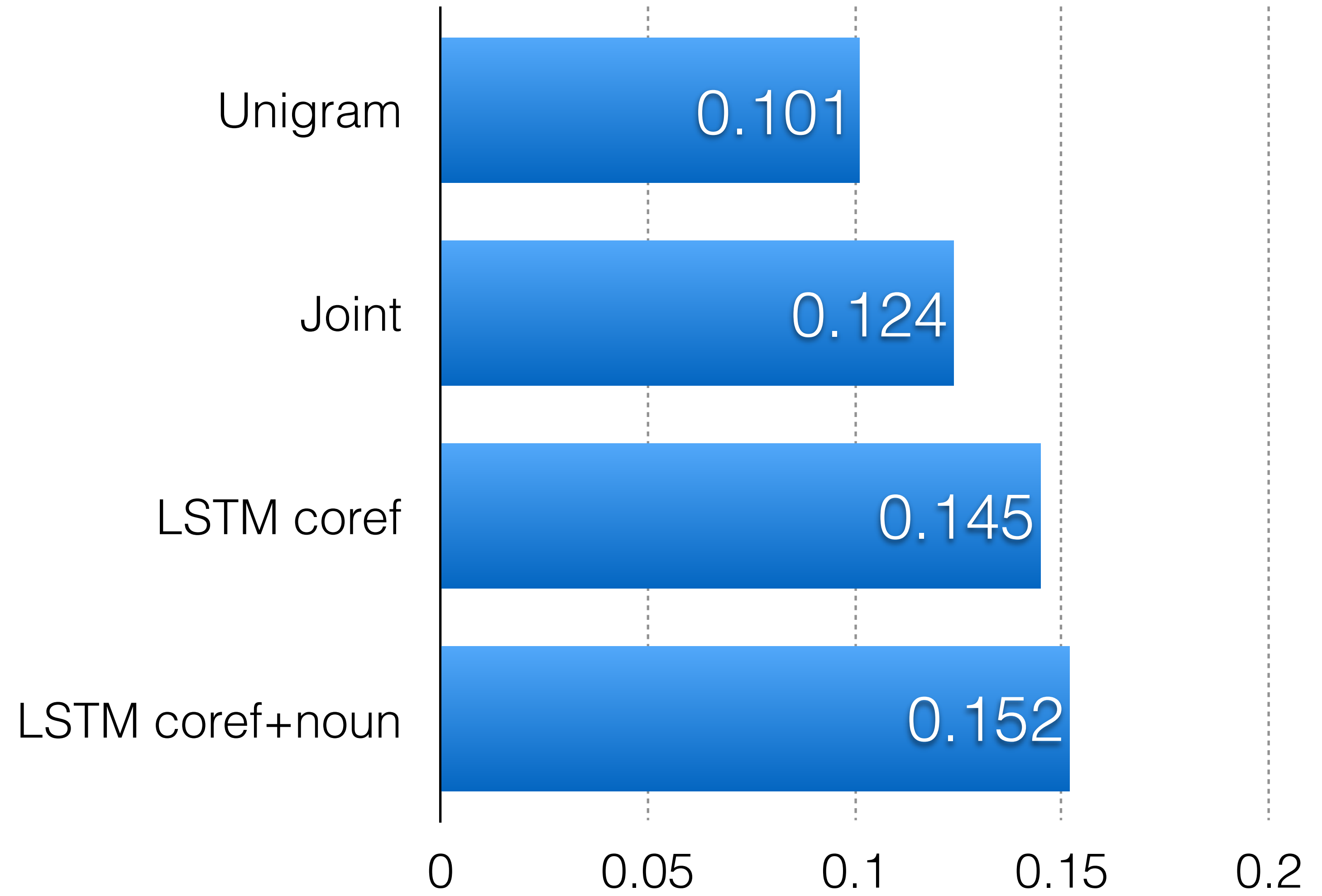
- *In April 1866 Congress again passed the bill. Johnson again vetoed it.*



LSTM Script models

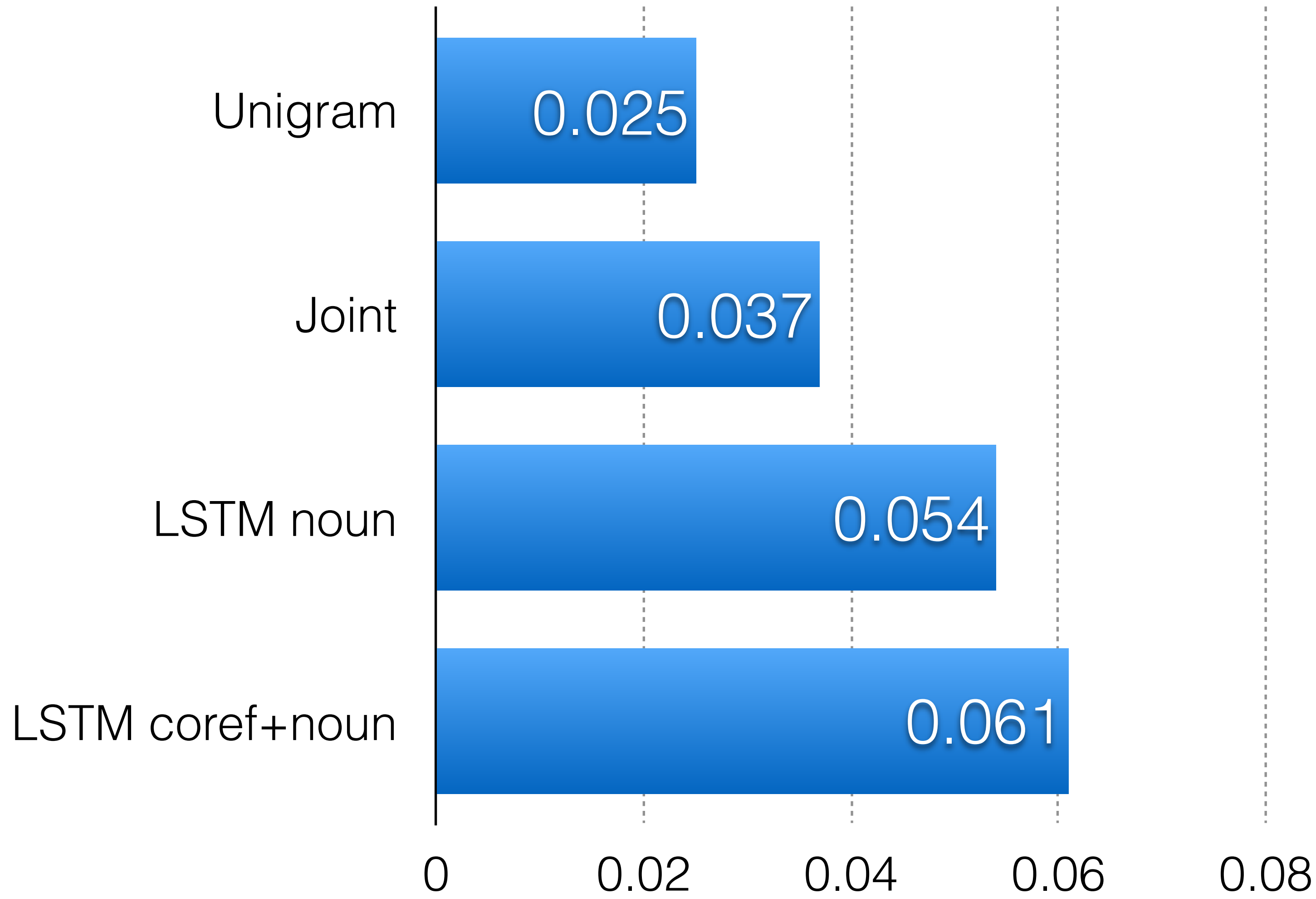
- Train on English Wikipedia.
- Run Stanford parser, coref; extract sequences of events.
- Train LSTM using Batch Stochastic Gradient Descent with Momentum.
    - Minimize cross-entropy loss of predictions.
    - Backpropagate error through layers and through time.
- To infer new events, just have the LSTM generate the next five outputs with highest probability (using beam search).

Results: Predicting Verbs & Coreference Info



Recall at 25 for inferring Verbs & Coref info

Results: Predicting Verbs & Nouns

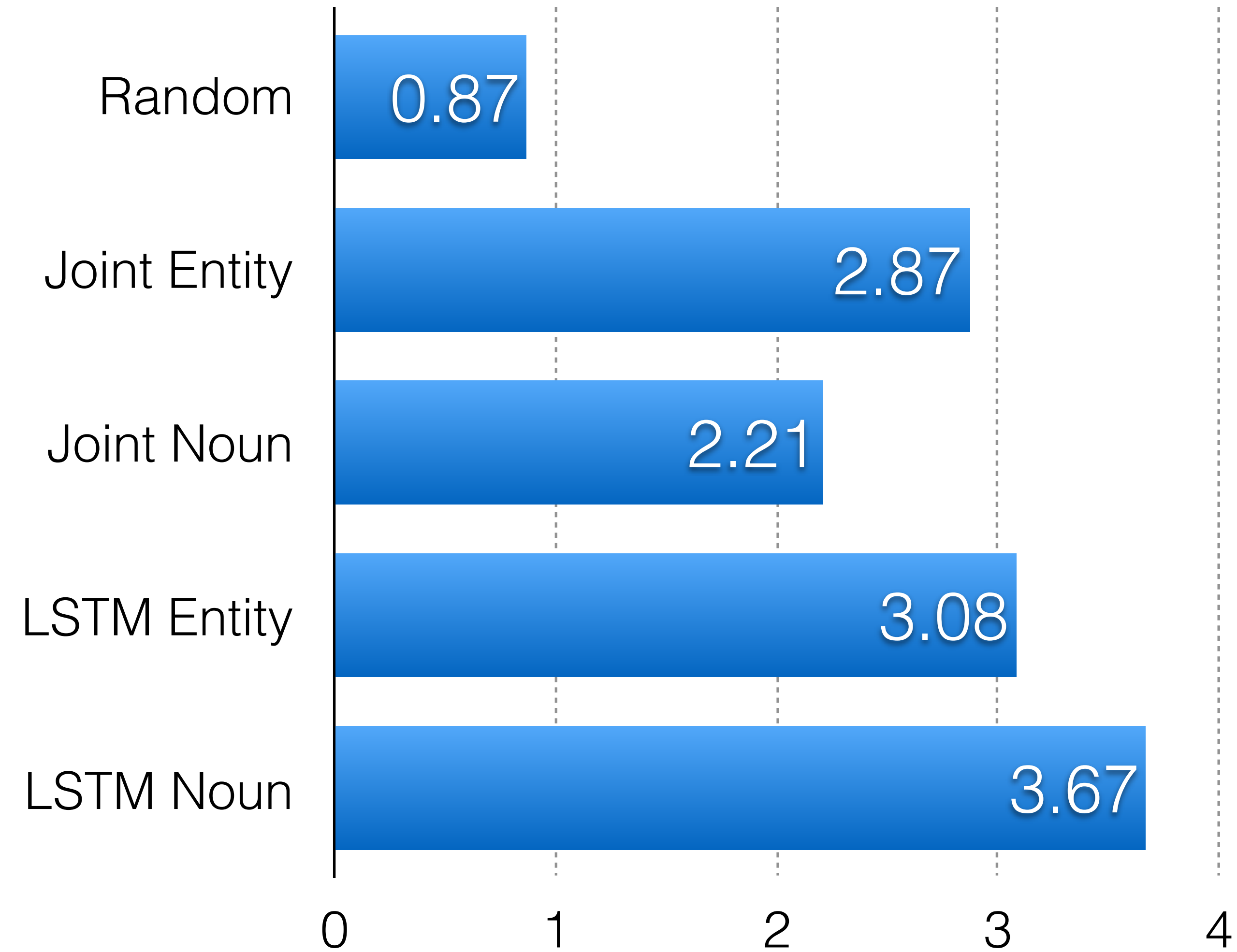


Recall at 25 for inferring Verbs & Nouns

Human Evaluations

- Solicit judgments on individual inferences on Amazon Mechanical Turk.
  - Have annotators rank inferences from 1-5 (or mark “Nonsense,” scored 0).
  - More interpretable.

Results: Crowdsourced Eval



Filtered Human judgments of top inferences (5 Max)

Annotator Examples

As a result , during the October municipal election , serious violence broke out on polling day , with shots exchanged by competing mobs .

Random	*appeal to X*	2.7
Joint Ent	*X has a X*	0.3
Joint Noun	*known as X*	3.3
LSTM Ent	*X has a X*	0.3
LSTM Noun	*X was arrested*	4.3

Annotator Examples

Today the remaining community has shrunk to about 50 mostly elderly people . The Kehila Kedosha Yashan Synagogue remains locked , only opened for visitors on request . Emigrant Romaniotes return every summer and open the old synagogue .

Random	*all of the X 's men were lost*	2.0
Joint Ent	*X found*	2.0
Joint Noun	*X wrote*	2.0
LSTM Ent	*build a X*	1.7
LSTM Noun	*synagogue was closed*	3.0

Generating “Stories”

- Can generate “stories” by starting with <S> beginning-of-sequence pseudo-event.
- Sample from distribution of initial event components (first verb).
- Take sample as first-step input, sample distribution of next components.
- Repeat until sampling </S> end-of-sequence token.

Generated “Stories”

Generated event tuples	**English Descriptions**
(bear, ., ., kingdom, into)	Born into a kingdom,…
(attend, she, brown, graduation, after)	…she attended Brown after graduation
(earn, she, master, university, from)	She earned her Masters from the University
(admit, ., she, university, to)	She was admitted to a University
(receive,she,bachelor,university,from)	She had received a bachelors from a University
(involve, ., she, production, in)	She was involved in the production
(represent, she, company, ., .)	She represented the company.

Outline

- Background
- Completed Work
- Proposed Work
- Conclusion

Research Questions

- How can Neural Nets improve automatic inference of events from documents?
  - Which models work best empirically?
  - Which types of explicit linguistic knowledge are useful?

Outline

- Background
- Completed Work
- Proposed Work
- Conclusion

Outline

- Background
- Completed Work
- Proposed Work
  - Better Models
  - Better Events
  - Discourse Relations
  - Bonus
    - Coreference
    - Question-Answering

Outline

- Background
- Completed Work
- Proposed Work
    - Better Models
    - Better Events
    - Discourse Relations
    - Bonus
        - Coreference
        - Question-Answering

Better Models

- Other Neural Approaches may work better than LSTM.

Better Models (1/3)

- Different kinds of RNN:
  - Gated Recurrent Units (GRUs) [Cho et al. 2014]
  - Grid LSTM [Kalchbrenner et al. 2015]
  - Gated Feedback Recurrent Units [Chung et al. 2015]
- Replacing one black box with another.

Better Models (2/3)

- Convolutional Neural Networks (CNNs):
  - Learn 1D convolution operators to apply to event sequences.
  - Arrive ultimately at vector predicting next event(s).
  - Recent success with NLP classification tasks. [Kalchbrenner, Grefenstette, & Blunsom 2014; Kim 2014; Zhang, Zhao, & LeCun 2015]

Better Models (3/3)

- **Attention-based Models**: contain explicit notion of where in input is most predictively useful (where to "pay attention").
- Recently shown to be useful in NLP tasks (Bahdanau et al. 2015, Hermann et al. 2015).

Better Models (3/3)

- Churchill had suffered a mild stroke while on holiday in the south of France in the summer of 1949. The strain of carrying the Premiership and Foreign Office contributed to his stroke at 10 Downing Street after dinner on the evening of 23 June 1953. Despite being partially paralysed down one side, he presided over a Cabinet meeting the next morning without anybody noticing his incapacity. Thereafter his condition deteriorated, and it was thought that he might not survive the weekend. Had Eden been fit, Churchill's premiership would most likely have been over. News of this was kept from the public and from Parliament, who were told that Churchill was suffering from exhaustion. He went to his country home, Chartwell, to recuperate, and by the end of June he astonished his doctors by being able, dripping with perspiration, to lift himself upright from his chair. He joked that news of his illness had chased the trial of the serial killer John Christie off the front pages. Churchill was still keen to pursue a meeting with the Soviets and was open to the idea of a reunified Germany. He refused to condemn the Soviet crushing of East Germany, commenting on 10 July 1953 that "The Russians were surprisingly patient about the disturbances in East Germany". He thought this might have been the reason for the removal of Beria. Churchill returned to public life in October 1953 to make a speech at the Conservative Party conference at Margate.

Better Models (3/3)

- Churchill had suffered a mild stroke while on holiday in the south of France in the summer of 1949. The strain of carrying the Premiership and Foreign Office contributed to his stroke at 10 Downing Street after dinner on the evening of 23 June 1953. Despite being partially paralysed down one side, he presided over a Cabinet meeting the next morning without anybody noticing his incapacity. Thereafter his condition deteriorated, and it was thought that he might not survive the weekend. Had Eden been fit, Churchill's premiership would most likely have been over. News of this was kept from the public and from Parliament, who were told that Churchill was suffering from exhaustion. He went to his country home, Chartwell, to recuperate, and by the end of June he astonished his doctors by being able, dripping with perspiration, to lift himself upright from his chair. He joked that news of his illness had chased the trial of the serial killer John Christie off the front pages. Churchill was still keen to pursue a meeting with the Soviets and was open to the idea of a reunified Germany. He refused to condemn the Soviet crushing of East Germany, commenting on 10 July 1953 that "The Russians were surprisingly patient about the disturbances in East Germany". He thought this might have been the reason for the removal of Beria. **Churchill returned to public life** in October 1953 to make a speech at the Conservative Party conference at Margate.

Better Models (3/3)

- **Churchill had suffered a mild stroke** while on holiday in the south of France in the summer of 1949. The strain of carrying the Premiership and Foreign Office contributed to his stroke at 10 Downing Street after dinner on the evening of 23 June 1953. Despite being partially paralysed down one side, he presided over a Cabinet meeting the next morning without anybody noticing his incapacity. Thereafter his condition deteriorated, and it was thought that he might not survive the weekend. Had Eden been fit, Churchill's premiership would most likely have been over. News of this was kept from the public and from Parliament, who were told that Churchill was suffering from exhaustion. **He went to his country home, Chartwell, to recuperate**, and by the end of June he astonished his doctors by being able, dripping with perspiration, to lift himself upright from his chair. He joked that news of his illness had chased the trial of the serial killer John Christie off the front pages. Churchill was still keen to pursue a meeting with the Soviets and was open to the idea of a reunified Germany. He refused to condemn the Soviet crushing of East Germany, commenting on 10 July 1953 that "The Russians were surprisingly patient about the disturbances in East Germany". He thought this might have been the reason for the removal of Beria. **Churchill returned to public life** in October 1953 to make a speech at the Conservative Party conference at Margate.

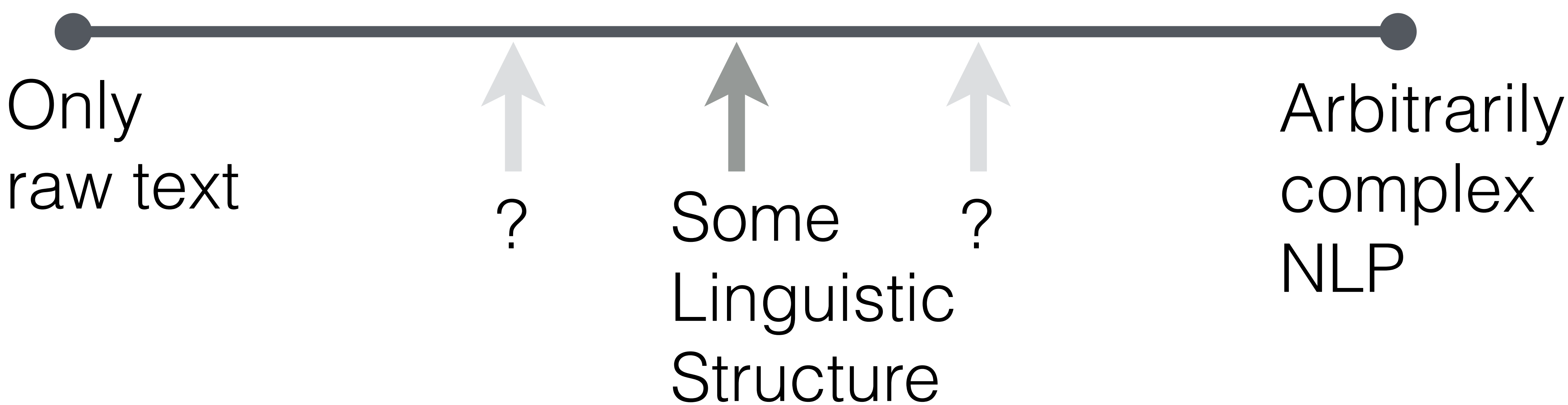
Better Models (3/3)

- An attention-based script system would add an explicit distribution of predictive utility over observed events.

Outline

- Background
- Completed Work
- Proposed Work
    - Better Models
    - Better Events
    - Discourse Relations
    - Bonus
        - Coreference
        - Question-Answering

Raw Text v. Linguistics



Better Events

- Events in P & Mooney (2016) are $v(e_s, e_o, e_p, p)$ 5-tuples.

- This throws away a lot of important information. Importance is an empirical question, but should be investigated.

- We will investigate a number of ways to add information to events (enumerated next…).

Better Events (1/4)

- Fixed-arity events throw away multiple Prepositional Phrases.

  - *In 1697, Peter the Great traveled incognito to Europe on an 18-month journey with a large Russian delegation to seek the aid of the European monarchs.*

- Is presently:
  (travel, peter, ·, (in 1697)).

- Could be something like
  (travel, Peter, ·, (in 1697), (to Europe), (on journey), (with delegation)).

Better Events (2/4)

- Many important modifiers that aren't grammatically prepositions.
  - *King Frederick William I nearly executed his son for desertion.*
  - Without "nearly" we make drastically wrong inferences!

Better Events (3/4)

- Head Nouns of Arguments are Insufficient:
  - *Martin Luther wrote to his bishop protesting the sale of indulgences.*
  - If “Sale of Indulgences” is just “sale”…
  - …we can’t conclude “Luther disapproved of indulgences.”

Better Events (4/4)

- Nominal (noun) events are very common:
  - *In the years following his **death**, a series of **civil wars** tore Alexander's empire apart.*
  - Noun events are crucial for inferring events.

Discourse Relations

- Relations *between* events are important.
  - *The Roman cavalry won an early victory **by** swiftly routing the Carthaginian horses.*
  - ***Because** the local authorities had forbidden students from forming organizations, Princip and other members of Young Bosnia met in secret.*
  - Connectives express relations between events, are likely useful for event inference.

Discourse Parsers

- Off-the-shelf **Discourse Parsers** have been trained to annotate discursively important relations between spans of text.

    - Trained on one of two Discourse treebanks [RST Treebank and Penn Discourse Treebank].

    - Label spans of text as being, e.g., causally or temporally related .

Using Discourse Parsers

- Could incorporate shallow discourse structure into events (i.e. "these two events are related by this discourse relation").
- Could also hypothetically incorporate discourse structure directly into structure of Neural Net.
    - RST parses are trees; could use recently-introduced Tree-LSTMs [Tai et al. 2015] on the topology.

Induced Connectives

- Could also induce a closed class of connectives (e.g. “before,” “because of,” …) in an unsupervised manner.
- A number of ways to integrate into RNN sequence model.

Outline

- Background
- Completed Work
- Proposed Work
    - Better Models
    - Better Events
    - Discourse Relations
    - Bonus
        - Coreference
        - Question-Answering

Scripts for Coreference

- ***Voltaire***, *pretending to work in Paris as an assistant to a notary, spent much of* ***his*** *time writing poetry. When his father found out, he sent* ***Voltaire*** *to study law. Nevertheless,* ***he*** *continued to write…*

- Unifying “he” and “Voltaire” could be done with script knowledge.

- Script information can improve coreference.

Scripts for Coreference

- A number of conceivable ways to incorporate scripts into ML-based coreference engines:
    - Add script probability, assuming a coreference decision is made, as feature to coref system.
    - Add probability of sequence of all events involving an entity as a feature.

Scripts for Question-Answering

- Scripts are intuitively useful for Question-Answering Systems.
  - Add confident script inferences to Knowledge Base about document.
  - Would allow inferences about implicit events.

Conclusion

- LSTMs do much better than Markov-like statistical script systems.
- We propose:
  - Using better neural models.
  - Better events.
  - More discourse-awareness.
  - Trying to improve coref.
  - Improving question-answering.

Thanks!