FEEL: Featured Event Embedding Learning

I-Ta Lee, Dan Goldwasser

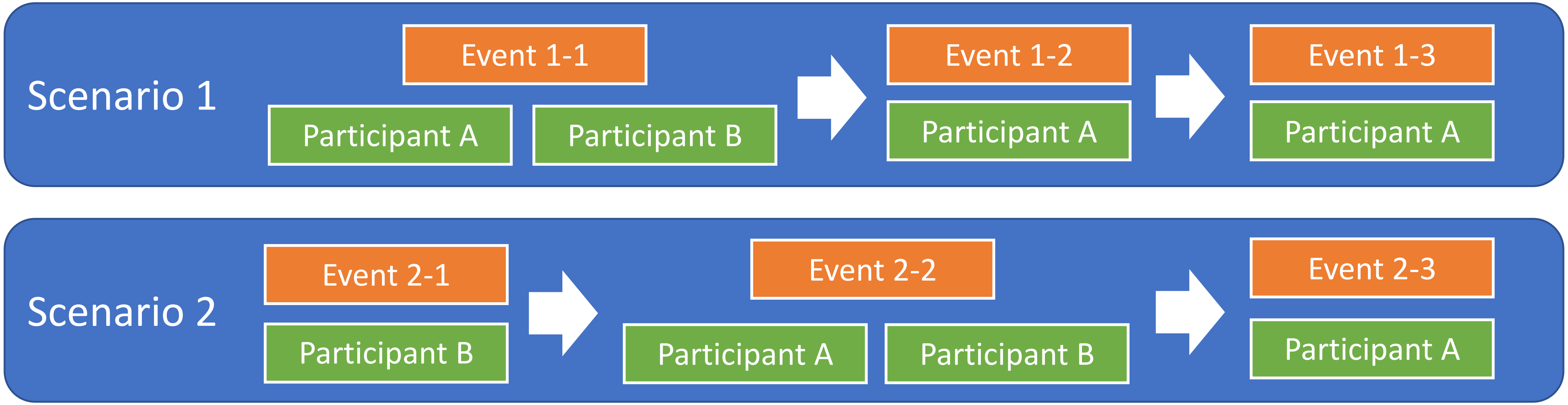
Purdue University

Introduction

- *Statistical Script Learning (SSL)* is one efficient way to acquire **world knowledge**, conduct **common sense reasoning**, and **disambiguate texts**.
  - The learned models are helpful in many Natural Language Processing (NLP) tasks that need common sense inference, such as **question answering, machine reading, coreference resolution** and so on.

What is "Scripts"?

- *Scripts*, introduced by *Schank and Abelson (1977),* are **structured knowledge representations** capturing the relationships between **event sequences and their participants**.
  - Scenarios repeat themselves.



How to use SSL models?

e1 e2

Jenny went to a restaurant and ordered a lasagna plate. Jenny liked the food and felt satisfied.

e3 e4

Which of the following events could happen *next*?

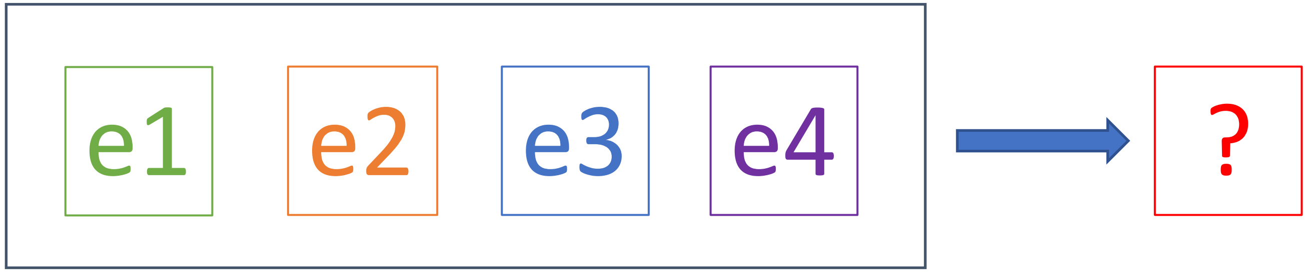
(a) She scolded the server.

(b) She fell asleep

(c) She left a big tip.

(d) She ran out of battery

(e) She was angry

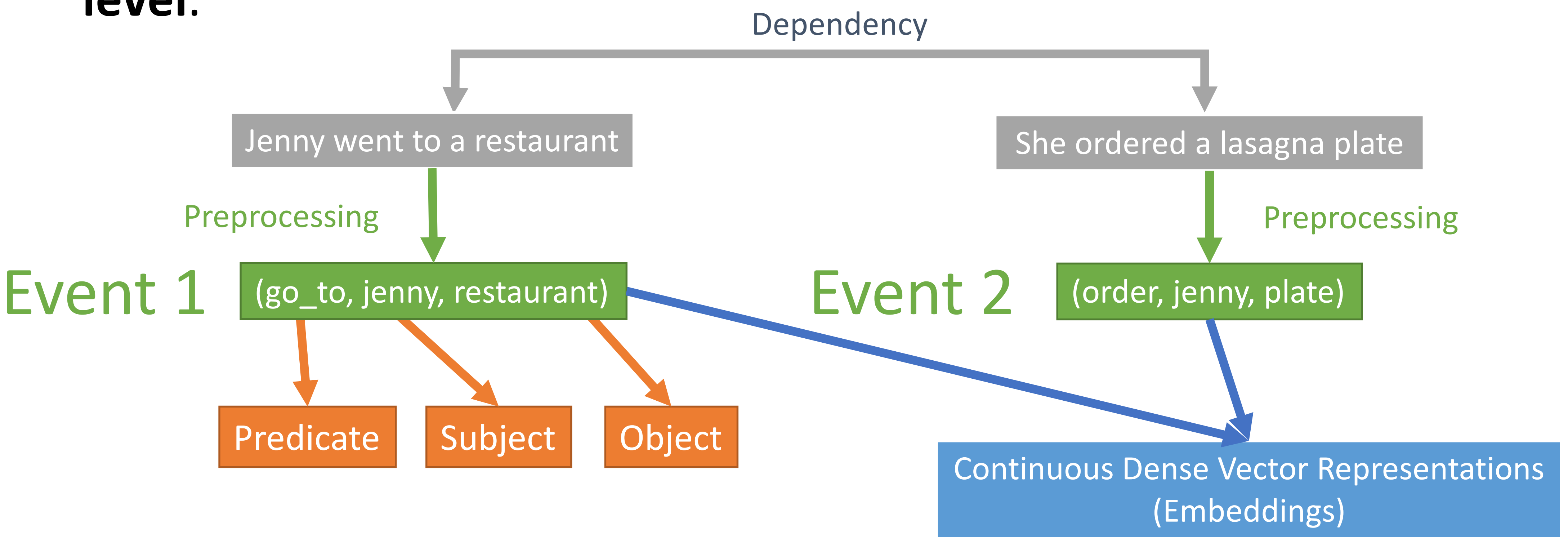


Narrative Cloze Test (Chamber and Jurafsky 08')

Multiple-Choice (Granroth-Wilding et al., 2016)

Previous Works

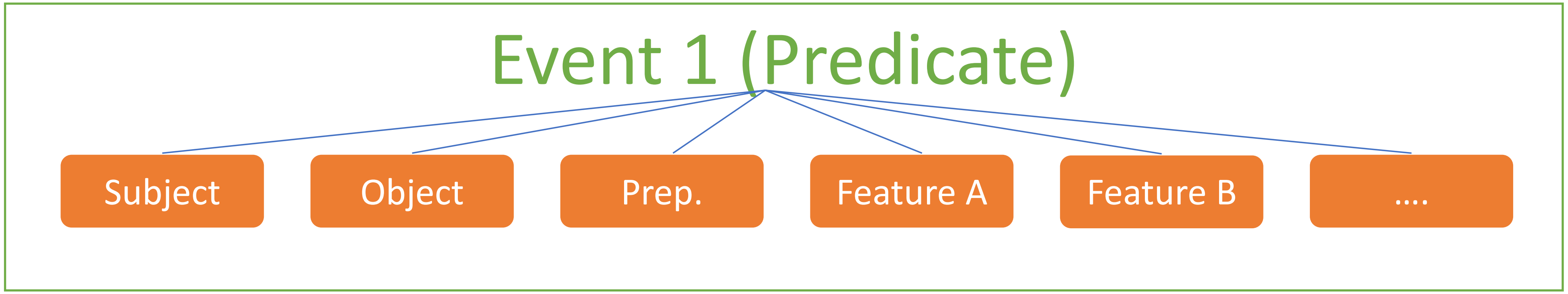
- Previous works focus on capturing patterns between events at **lexicon level**.



(K. Pichotta and R. J. Mooney, 2014)

FEEL: Featured Event Embedding Learning

- We propose *FEEL*, an SSL model, which is designed to capture **fine-grained event features** that can be exploited to reduce ambiguity when inferring future events.
  - We believe considering an event as **an predicate and a set of interested features** is a more flexible setting, as it allows plug-in features.

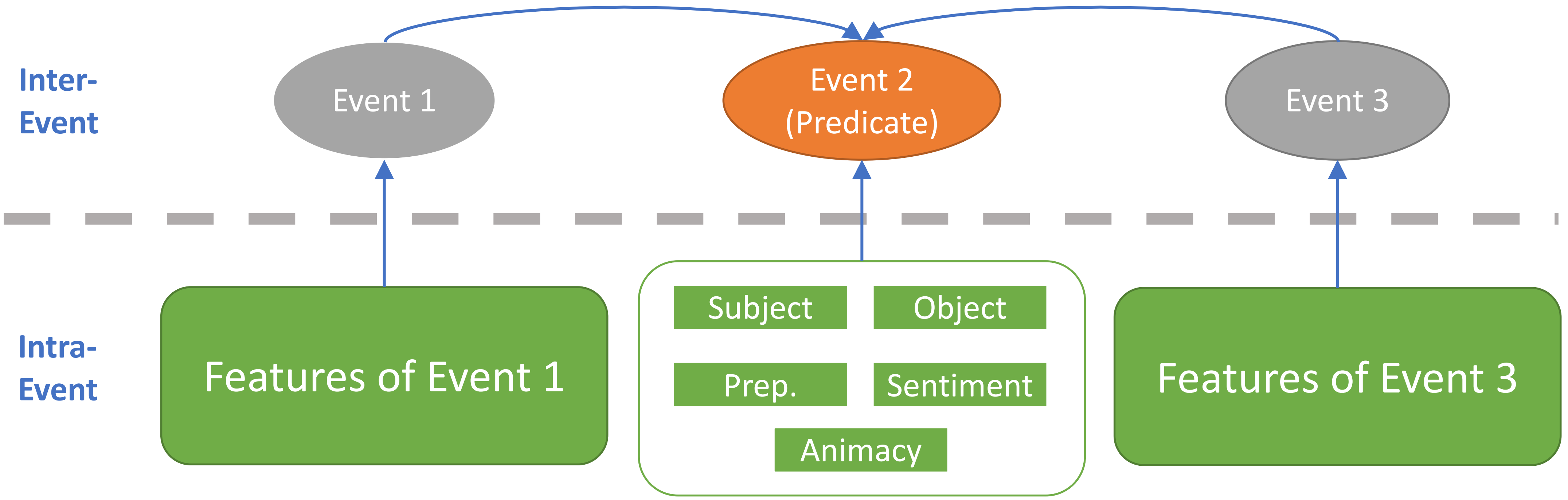


Example Features

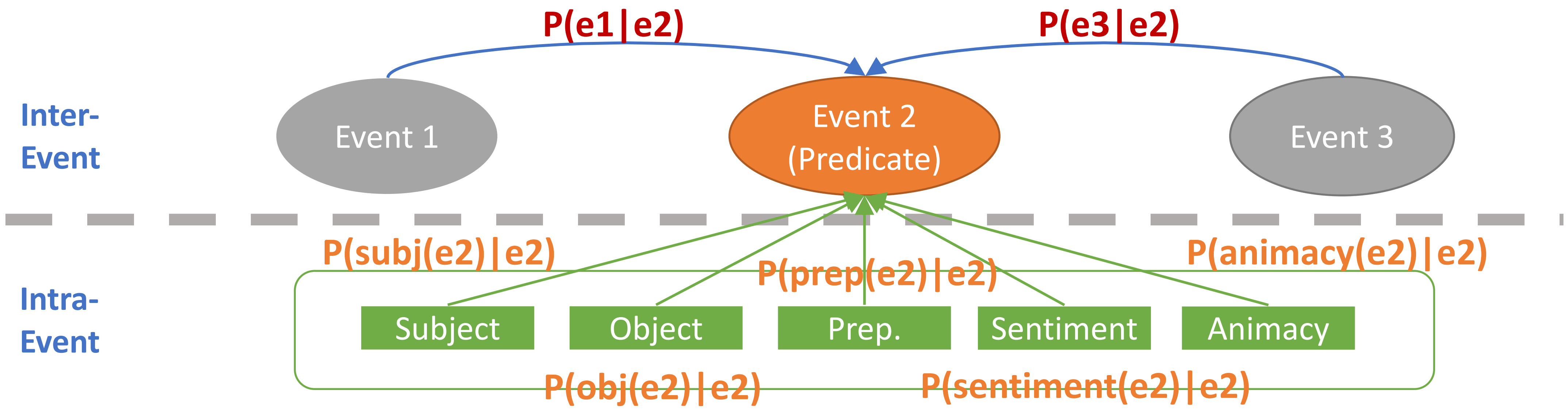
- *Sentiment Polarity* of a given event can impact the probability distribution of future events
  - Given *"Jenny liked the food"* **Positive Sentiment**
  - *"She left a big tip"* is more probable than *"She scolded the server" to happen*
    Positive **Negative**
- *Animacy* of the event arguments
  - *"This song is sick!"*
  - *"This person is sick!"*

Hierarchical Script Model

- Event sequences are hierarchical in nature if we consider their features



Model Objectives



$$p(C(e)|e) = \prod_{e' \in C(e)} p(e'|e) = \prod_{e' \in C(e)} \frac{\exp(v_{e'} \cdot v_e)}{\sum_{e^* \in E} \exp(v_{e*} \cdot v_e)}$$

Skip-Gram (T. Mikolov et al., 2013)

Multi-Task Learning

- We have
  - the inter-event objective:
    - *event-event (C)*
  - the intra-event objectives:
    - *event-subj (S), event-object (O), event-prep (P), event-sentiment (S), event-animacy (A)*
- Optimize them using Margin-based Ranking Loss

$$L_i(e) = \sum_{e' \in c(e)} \sum_{e^* \notin c(e)} \max(0, \delta - v_e \cdot v_{e'} + v_e \cdot v_{e^*})$$

$$i \in \{C, S, O, P, T, A\}$$

$$\mathcal{L}(e) = \lambda_r ||w|| + \sum_{i \in \{C,S,O,P,T,A\}} \lambda_I L_i$$

Basic Evaluation: Multi-Choice Narrative Cloze (MCNC)

(Granroth-Wilding et al., 2016)

Jenny went to a restaurant and ordered a lasagna plate. Jenny liked the food and felt satisfied.

Which of the following events could happen *next*?

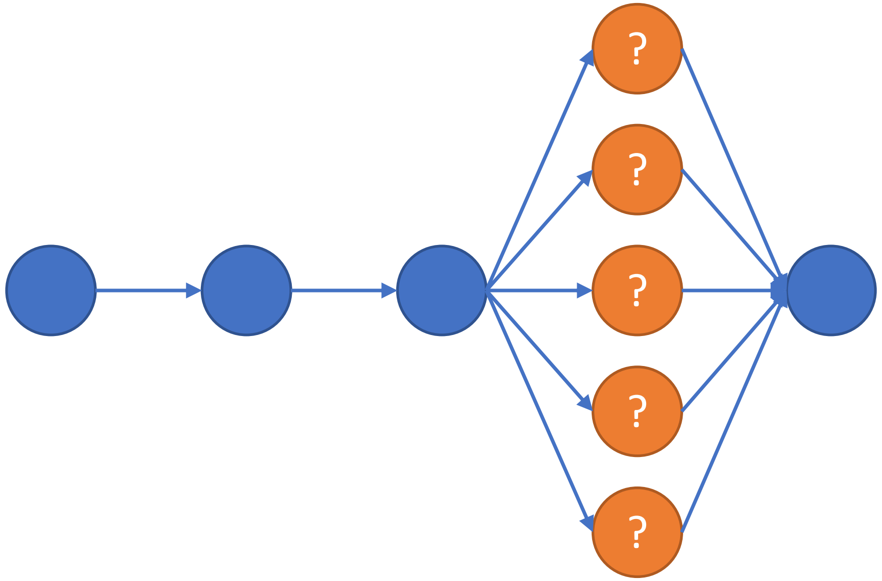
(a) She scolded the server.

(b) She fell asleep

(c) She left a big tip.

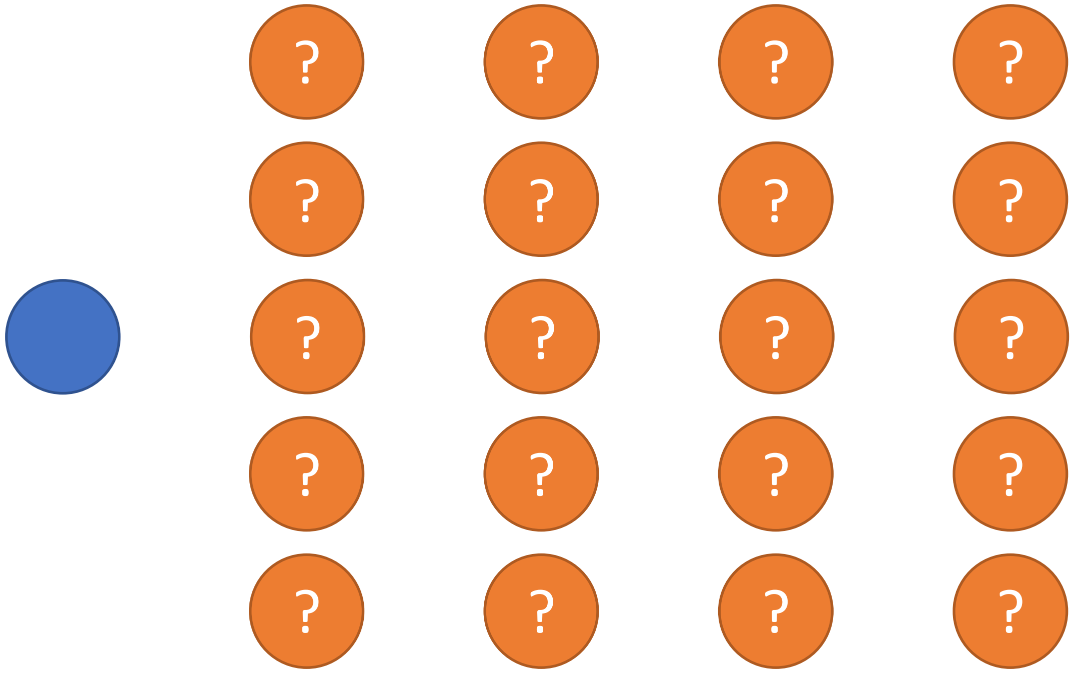
(d) She ran out of battery

(e) She was angry.



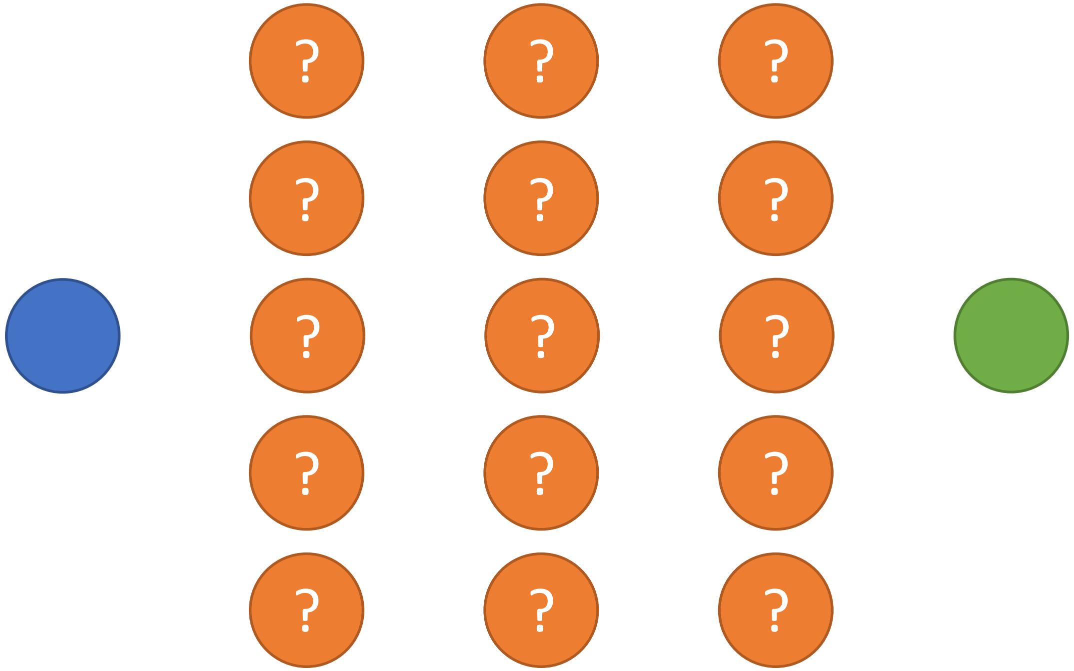
More interesting evaluation tasks (1)

- Multi-Choice Narrative Sequences (MCNS)
  - Evaluate Models' ability to do longer inference
  - Story Generation



More interesting evaluation tasks (2)

- Multi-Choice Narrative Explanation (MCNE)
  - Evaluate Models' ability to explain what happens in between
  - Story Explanation



Jenny went to a restaurant and left a big tip.

Which of the following event chains *explain* what happened?

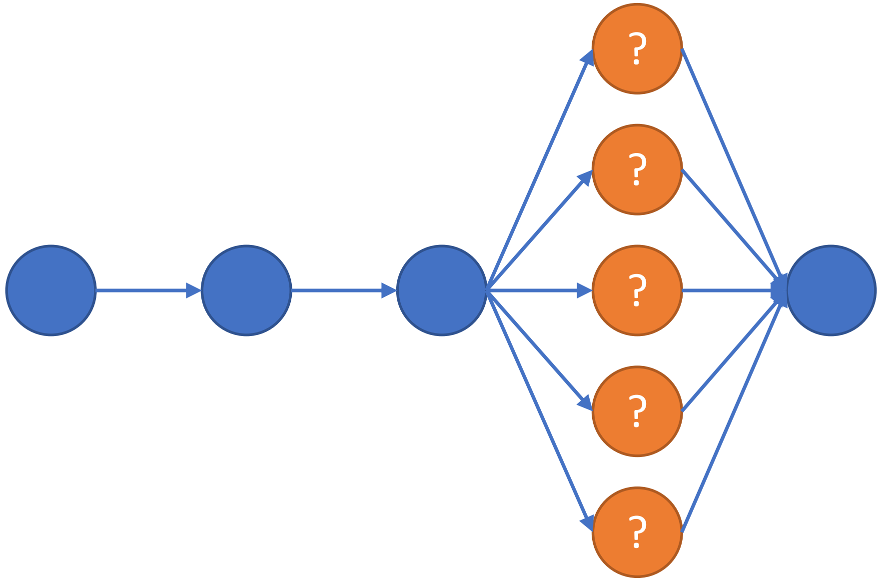
(a) She ordered her food and liked it.

(b) She hate her food and left angry.

(c) She walked to a bus station and got on a bus.

Results: Multi-Choice Narrative Cloze

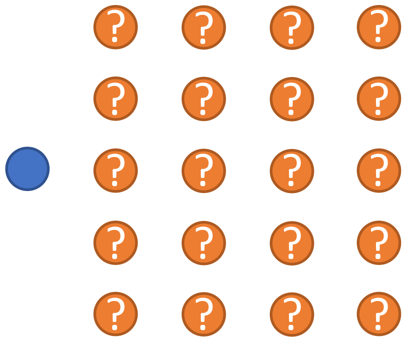
	Accuracy
Granroth-Wilding et al., 2016	0.4957
Wang et al., 2017	**0.5512**
Pred	0.4232
Pred+Args	0.5135
Pred+Args+S	0.5166
Pred+Args+A	**0.5503**
Pred+Args+S+A	0.5418

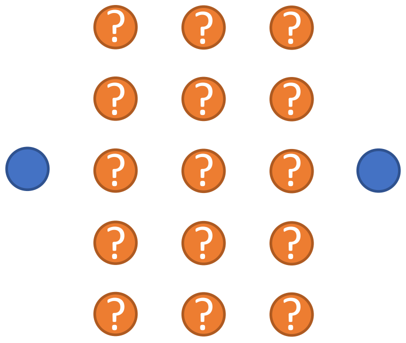


Results:
MCNS and MCNE

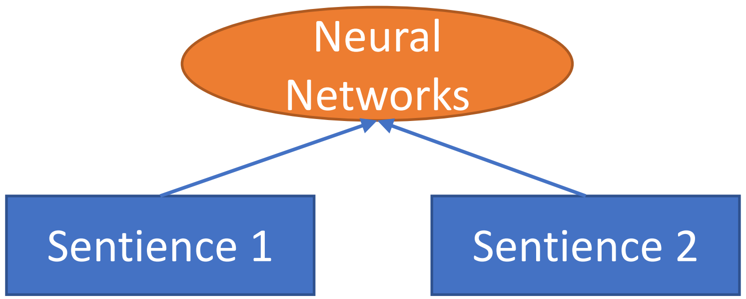
Accuracy	MCNS-Viterbi	MCNE-Viterbi
GloVe	0.353	0.385
GloVe+Pred	0.359	0.389
GloVe+Pred+Args	0.332	0.37
GloVe+Pred+Args+S	**0.416**	**0.448**
GloVe+Pred+Args+A	0.399	0.429
GloVe+Pred+Args+S+A	0.365	0.403

GloVe
(J. Pennington et al., 2014)





Results: Shared Tasks



- Semantic Relatedness (SICK)
  - SemEval 2014 Shared Task
  - **Regression** Task with Neural Networks
  - Pearson Scores

Pearson	Pred+Args+A	Pred+Args+S+A
GloVe	0.7102	
FEEL	0.6714	0.6714
GloVe+FEEL	**0.7676**	0.7604

- Implicit Discourse Sense (IDSC)
  - CONLL 2016 Shared Task
  - **Multi-class Classification** with Neural Networks
  - Micro-Average F1

Micro Average F1	Test	Blind Test
GloVe	0.2982	0.2815
GloVe+PredDep	0.2921	0.2886
GloVe+PredDep+Args	0.2983	0.2862
GloVe+PredDep+Args+S	0.2996	0.3102
GloVe+PredDep+Args+A	0.3063	**0.3111**
GloVe+PredDep+Args+S+A	**0.3174**	**0.3111**

Conclusions

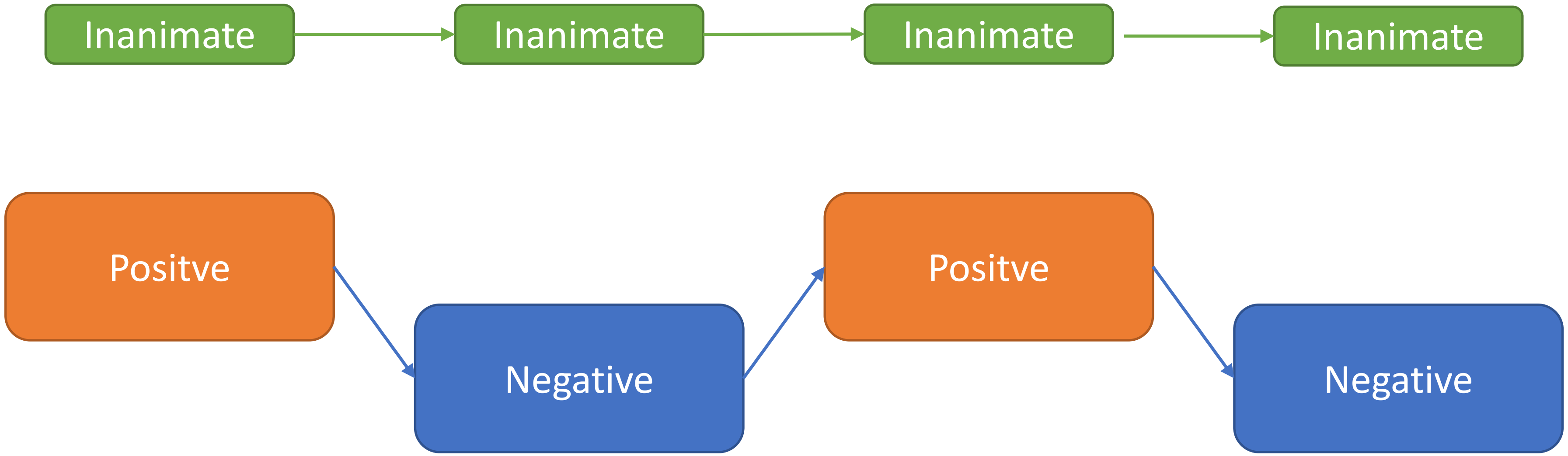
- We proposed FEEL: Featured Event Embedding Learning
  - Hierarchical multi-task representation learning
  - Feature-enriched event embeddings
- We also proposed two novel tasks for evaluating structured event sequence.
  - Story Generation and Story Explanation
- We showed that
  - FEEL is competitive with the most recent state-of-the-art in the narrative cloze test.
  - The resulting embedding can be used as a strong representation for advanced semantic tasks.

Thanks for Listening.
Any Questions?

Appendix

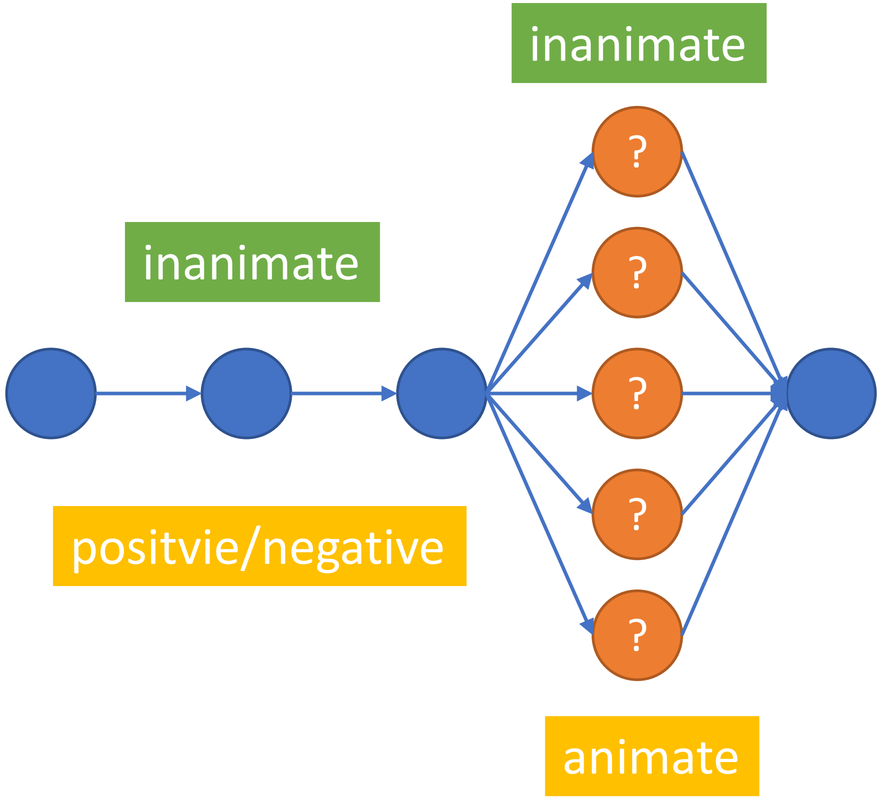
Cool Observations

- Animacy is more useful when making single prediction
- Sentiment is more useful when making multiple predictions
  - Sentimental Trajectory



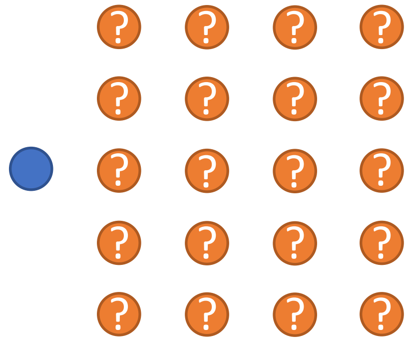
Evaluation:
Multi-Choice Narrative Cloze

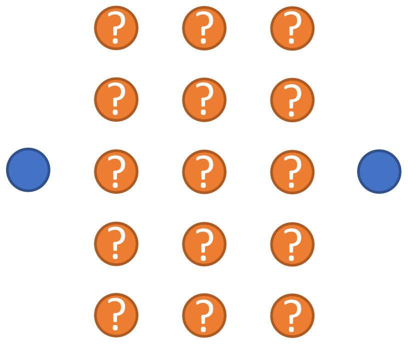
	Accuracy
Granroth-Wilding et al., 2016	0.4957
Wang et al., 2017	**0.5512**
Pred	0.4232
Pred+Args	0.5135
Pred+Args+S	0.5166
Pred+Args+A	**0.5503**
Pred+Args+S+A	0.5418



Results:
MCNS and MCNE

	MCNS-Viterbi	Baseline	Skyline	MCNE-Viterbi
GloVe	0.353	0.297	0.356	0.385
GloVe+PredDep	0.359	0.302	0.362	0.389
GloVe+PredDep+Args	0.332	0.366	0.434	0.37
GloVe+PredDep+Args+S	**0.416**	0.385	0.460	**0.448**
GloVe+PredDep+Args+A	0.399	0.396	0.465	0.429
GloVe+PredDep+Args+S+A	0.365	0.383	0.452	0.403

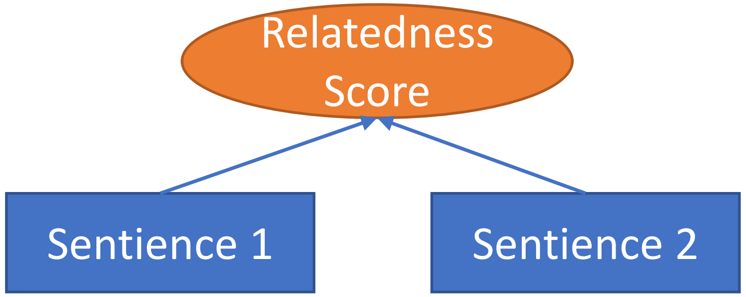




Evaluation: Shared Tasks

- Semantic Relatedness (SICK) - SemEval 2014 Shared Task
  - Regression Task with Neural Networks
  - Pearson Scores

Pearson	PredDep	PredDep+Args	PredDep+Args+S	PredDep+Args+A	PredDep+Args+S+A
GloVe	0.7102				
FEEL	0.4452	0.6574	0.6791	0.6714	0.6714
GloVe+FEEL	0.7382	0.7572	0.7518	**0.7676**	0.7604



Evaluation:
Shared Tasks

- Implicit Discourse Sense (IDSC)- CONLL 2016 Shared Task
  - Multi-class Classification with Neural Networks
  - Micro-Average F1

Micro Average F1	Test	Blind Test
GloVe	0.2982	0.2815
GloVe+PredDep	0.2921	0.2886
GloVe+PredDep+Args	0.2983	0.2862
GloVe+PredDep+Args+S	0.2996	0.3102
GloVe+PredDep+Args+A	0.3063	**0.3111**
GloVe+PredDep+Args+S+A	**0.3174**	**0.3111**