CIS-700

Spring 2020

Reasoning for Natural Language Understanding

Dan Roth

Computer and Information Science

University of Pennsylvania

Introduction Part II

This class

- Understand early and current work on Reasoning
  - (Learn to) read critically, present, and discuss papers
- Understand some of the difficulties in NLU from the perspective of reasoning
  - Conceptual and technical
- Try some new ideas

- How:
  - Presenting/discussing papers
    - Probably: 2 presentations each; 4 discussants
  - Writing a few critical reviews
  - "Small" individual project (reproducing);
  - Large project (pairs)
  - Tentative details are on the web site.

- Today: discuss first project
  - Content + Timetable
- Tomorrow: release list of papers
  - Timetable

- Machine Learning
  - 519/419
  - 520
  - Other?
- NLP
  - Yoav Goldberg's book
  - Jurafsky and Martin
  - Jacob Eisenstein

- Attendance is mandatory
- Participation is mandatory

- **Time of class?**
- **Expectations?**

This Class (2)

- Most of this class will be technical – reading, discussing, trying to understand current literature in this area (and some earlier literature).
- Before we get there, I'd like to spend the next two meetings on:

1. What are the questions we should ask in this area (and why)?

2. What were the earlier foci of the relevant research communities (and why)?

Questions

- What is Reasoning?
  - Do we do reasoning? Yes, we do.
  - How can we formulate it?
- Reasoning requires knowledge
  - How do we represent it?
  - What types of knowledge
  - What types of representations?
- We want to think about these in the context of natural language understanding
  - In what ways does it change the game?
  - Is Reasoning for/in NLU different than “Reasoning”?

John McCarthy on Natural Language Understanding

AN EXAMPLE FOR NATURAL LANGUAGE UNDERSTANDING AND THE AI PROBLEMS IT RAISES

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1976

A New York Times Story

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday in his downtown Brooklyn store by two robbers while a third attempted to crush him with the elevator car because they were dissatisfied with the $1,200 they had forced him to give them.

The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.

A New York Times Story (Cont.)

Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter. The store at 340 Livingston Street is part of the Seaman's Quality Furniture chain.

Mr. Hug was removed by members of the Police Emergency Squad and taken to Long Island College Hospital. He was badly shaken, but after being treated for scrapes of his left arm and for a spinal injury was released and went home. He lives at 62-01 69th Lane, Maspeth, Queens.

He has worked for seven years at the store, on the corner of Nevins Street, and this was the fourth time he had been held up in the store. The last time was about one year ago, when his right arm was slashed by a knife-wielding robber."

New York Times Story: Questions

- An intelligent person or program should be able to answer the following questions based on the information in the story:
  - The article proceeds with 22 questions:

1. Who was in the store when the events began?
   - Probably Mr. Hug alone, although the robbers might have been waiting for him, but if so, this would have been stated.

2. What did the porter say to the robbers?
   - Nothing, because the robbers left before he came.

20. Why did Mr. Hug yell from the bottom of the elevator shaft?
   - So as to attract the attention of someone who would rescue him.

- "The above list of questions is rather random. I doubt it covers all facets of understanding the story."

McCarthy's Challenges

The QA module is not being trained
Once the program knows English, and has the relevant background knowledge, it should answer the questions

- A formalism capable of expressing the assertion of the sentences free from dependence on the grammar of the English language. ("Artificial Natural Language", ANL)
  - Semantic Parser
- An "understander" that constructs the "facts" from the text.
  - Information Extraction: Entities, Relations, Temporal, Quantities,…
- Expression of the "general information" about the world that could allow getting the answers to the questions from the "facts" and the "general information"
  - Background Knowledge
- A "problem solver" that could answer the above questions on the basis of the "facts".
  - Question Answering Engine

Lessons

- What can we learn from this example?
  - Difficulties of NLU
  - Importance of reasoning
    - Part of Reasoning here seems to be “providing the reasons”, not only the “answers”
  - Decoupling learning from reasoning
    - McCarthy thinks that there is a need for some level of abstraction – an abstract representation of the text and the relevant knowledge so that a generic module can work on it and “do the reasoning”.
- Is this important/Essential?

Class Discussion

- You will spend the next 10 minutes on:

  - ☐ Suggest a reasoning problem.
    - Describe it
    - Suggest a way to formulate it so that you can write a program that solves it

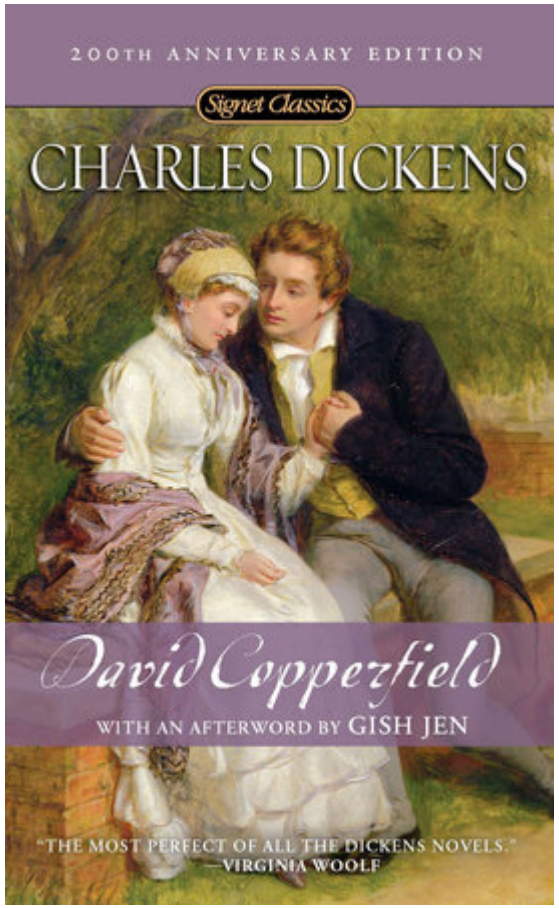
  - ☐ Think about knowledge needed
    - Describe the type of knowledge you think is needed and why/when
    - Suggests ways to formulate it: represent it and use it

More Examples

- Skip

(1) I Have Some Questions About….

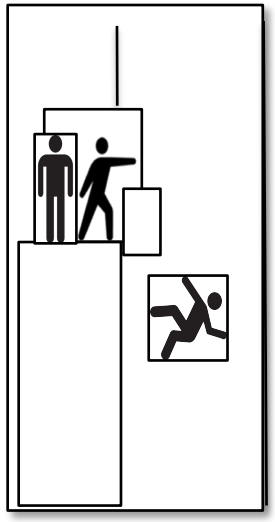
- Scenario:
  - You are reading the book, but left it for a couple of weeks.
  - You need a refresher: some of the events, entities, the current relationships between David and James Steerforth.
- Conversing about it is challenging:
  - Many chapters, multiple voices, long periods of time,…
  - The novel features the character David Copperfield, his journey of change and growth from infancy to maturity, as many people enter and leave his life and he passes through the stages of his development. (**Fiction, and you know it**)
  - London and England in the 19-th century; socio-economic state, child exploitation; schools, prisons, emigration to Australia (**true historical facts**)
- What computational tasks should we think about?



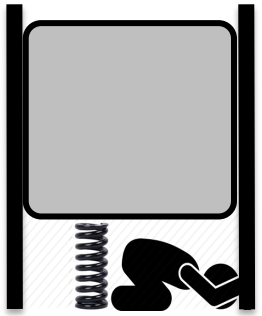
(2) I Want to Talk about this News Story

Who was in the store when the events began?

- ☐ The story doesn't say.
- ☐ Probably Mr. Hug alone, although the robbers might have been waiting for him, but if so, this would have been stated.



- Why was he crying?
  - ☐ Maybe he was scared.
  - ☐ Maybe he was injured.
  - ☐ Maybe he called for help.



- Understanding the **story** and **conversing** about it require **Situated Reasoning**: Model-based Reasoning

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday in his downtown Brooklyn store by two robbers while a third attempted to crush him with the elevator car because they were dissatisfied with the $1,200 they had forced him to give them.

The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.

Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter. The store at 340 Livingston Street is part of the Seaman's Quality Furniture chain.

Mr. Hug was removed by members of the Police Emergency Squad and taken to Long Island College Hospital. He was badly shaken, but after being treated for scrapes of his left arm and for a spinal injury was released and went home. He lives at 62-01 69th Lane, Maspeth, Queens.

He has worked for seven years at the store, on the corner of Nevins Street, and this was the fourth time he had been held up in the store. The last time was about one year ago, when his right arm was slashed by a knife-wielding robber."

[Example due to John McCarthy]

(3) Some questions to my Sports' Assistant

Modified version of a question for AI2's DROP dataset

Coming off a road win over the Cowboys, the Redskins traveled to Lincoln Financial Field for a Week 5 NFC East duel with the Philadelphia Eagles. In the first quarter, the Redskins trailed early **as RB Brian Westbrook scored on a 9-yard TD run** and **the Eagles DeSean Jackson returned a punt 68 yards for a touchdown.** Washington still trailed at half time 14:9, **with field goals from Shaun Suisham**. In the third quarter, the Redskins took the lead on a trick play as WR Antwaan Randle El threw an 18-yard **TD pass to TE Chris Cooley**. In the fourth quarter, the Redskins increased their lead when **Clinton Portis scored on a 4-yard TD run**. The Eagles managed one more score in the final quarter for a **final score of 17:23**.

Football Scoring Rule Book:

- TD could be 6, 7, or 8 points.
  - Kick….
- Field goal is worth 3 points
- …..

General rules that are to be inst

aned from the recap.
ograms on the text
g team))
n is necessary.

What computational tasks should we think about?

(4) Let's Talk about Dinner

- → Let's talk about dinner.
  - **A:** Where do you want to go?
- → I had a big lunch
  - [This is not an answer; can the Assistant figure it out?]
  - It's probably just a hint that we should go for a light dinner
- → I don't like crowds
  - [This is not an answer; can the Assistant figure it out?]
  - Perhaps a preference for small venues?
- → I had a lot of pizza the last few weeks
  - [Again; not a direct; how do we understand it? How do we represent it?

But, there is another big difference between these two scenarios.

- The first applies only today.
- The second is a general rule that I'd like the Assistant to remember.

Formulation

- What is the role of **formal theories** of reasoning and representation?
  - They assume that we can map text and world knowledge to a "symbolic" representation; given that, the problem is solved (so people think).
    - Note that this is true even when people use neural networks for all/part of the computation
  - If this is wrong – where is it wrong?
    - Is it the infeasibility of the mapping?
    - Is it that our formal theories of reasoning are missing something?

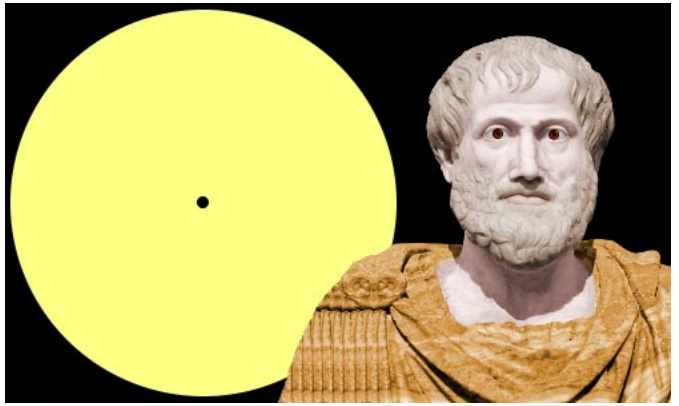
- Think also about the statements I expressed last time
  - Reasoning is about giving reasons
- What are the implications of this (whether you agree with it or not) on the need to have "symbols"?

Formulation (cont.)

- Eventually, we may want to think about a neural implementation of Reasoning
  - Is it necessary? Is this where the challenge is?
    - Note that neuro-symbolic AI goes back many years
  - Is it ok to first think about formal theories, and then encode them in neural architectures?
  - Or, there is an advantage to directly thinking about neural representations.
    - This latter view means that there is **no other representation** of what neural architectures are doing.
      - E.g., is an embedding of a sentence different than other representations of it in some principled way?
      - Or is it just a more compact representation?
- Next, we will describe things from the perspective of Learning to Reason.
  - The presentation will mostly focus on the logical approach, but similar ideas can be extended to other formalisms.
  - This will hopefully serve as introduction to formalisms,

It's Time for Reasoning: **Outline**

- Reasoning about **events** and **time** in natural language
  - Temporal ordering of events
    - Learning & Inference paradigms to support reasoning
  - Temporal common sense
  - Reasoning & Supervision paradigms



Did Aristotle have a laptop?

- Reflects an important move in NLU from **sentence level** to **situation level**
- Addresses issues in combining learning and reasoning, and **supervision.**

- More about Temporal Common Sense

- Initial thoughts on additional Reasoning paradigms
  - Decomposition, and computing functions over sets of variables

Constrained Conditional Models [Abductive Reasoning; Chang et al.'12]

Supporting structured, knowledge intensive, NLP decisions

Variables are models

$$y = \text{argmax}_y \sum \mathbf{1}_{\emptyset(x,y)}\, \mathbf{w}_{x,y} \quad \text{subject to Constraints } C(x,y)$$

Penalty for violating the constraints.

Formulation goes back to (Roth & Yih 2004). Also related to PR (Ganchev et al. 2010)

...nowledge component: (Soft) constraints

A linear function over models – can be used to model any logical function

Features, Models, NN (non-linearity comes here)

How far are the decisions (y) is from a "legal/expected" assignment

- Is it needed/useful in the NN era? Yes [Deutsch et al. CoNLL'19] ; **later: a neural implementation of CCMs**
  - But, it's not sufficient to support all types or reasoning we care about.
- It has been used extensively and successfully used in many NLU tasks from IE to discourse to summarization.
- A good starting point for thinking about further progress in natural language understanding.

- Training: Learning the objective function (**w, u**)? Learning all the intermediate functions **Ø(x, y)?**
  - Joint? Decoupled? Learned/provided constraints? Hard/soft?
  - There is some understanding for when to do what [IJCAI'05]
- Reasoning: A way to push a **function over learned models** to satisfy output expectations
  - (can also think about expectations from a latent representation)

Two Events

People were angry

Police used tear gas

Time

People **were angry** at something (which ended in violent conflicts with the police)…The police finally **used tear gas** (to restore order).

Two Events

Police used tear gas

People were angry

Time

Police **used tear gas**…People **were angry** at the police.

Two Events

Police used tear gas

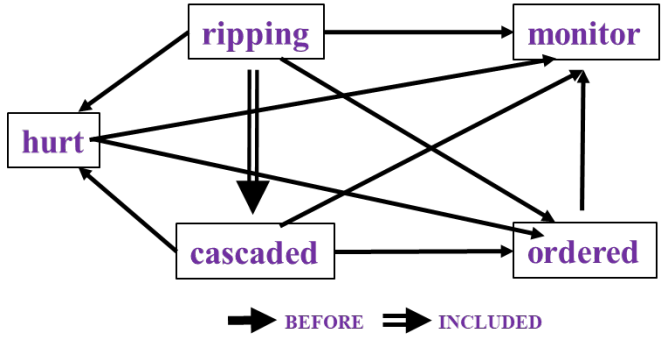
People were angry

Time

In natural language, we rarely see explicit **timestamps**, so we have to figure out the temporal order **from cues in the text**.

Time and Events [Ning et al. *SEM'2018; ACL'18, EMNLP'18, EMNLP'19]

- In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.

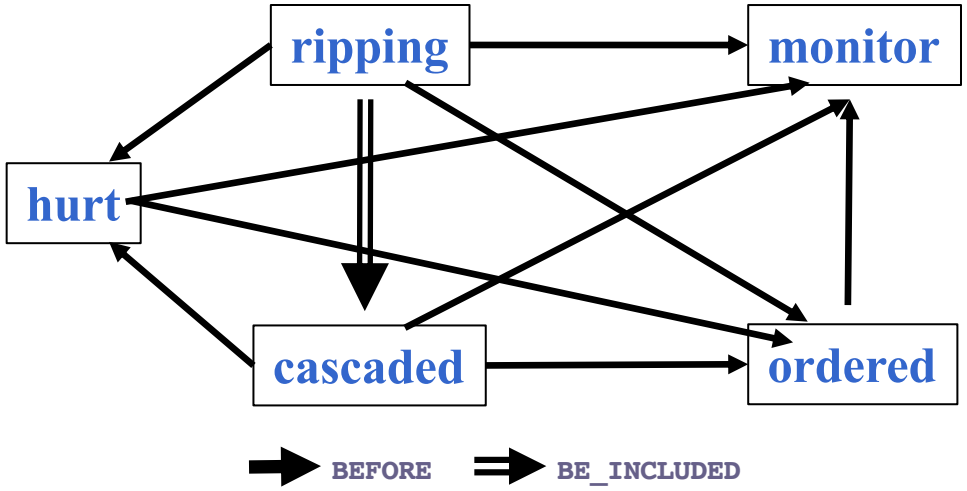


- Very difficult task— hinders exhaustive annotation ($O(N^2)$ edges)
- But, it's rather easy to get partial annotation – some relations.
- And, we have **strong expectations** from the output
  - Transitivity
  - Some events tend to precede others, or follow others

Multiple Events Form a Situation

The task: label the edges of the temporal graphs.

- In Los Angeles that lesson was brought home Friday when tons of earth cascaded down a hillside, ripping two houses from their foundations. No one was hurt, but firefighters ordered the evacuation of nearby homes and said they'll monitor the shifting ground until March 23rd.



This task is difficult:

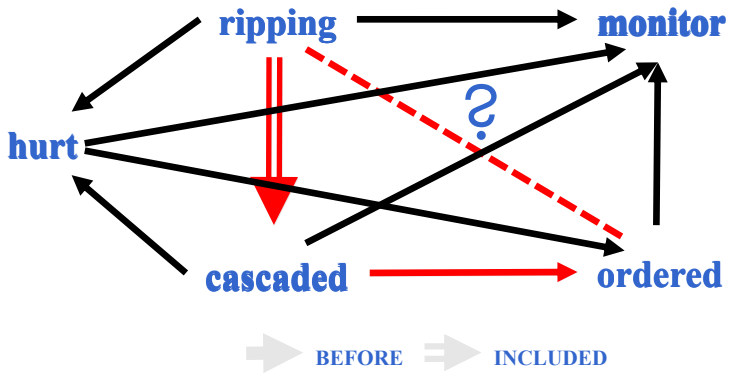
Temporal relation extraction	Literature F1 (%)
w/ gold events	low 50's
w/o gold events	low 30's

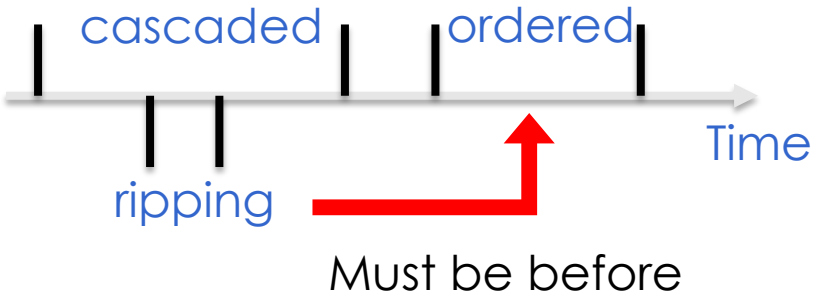
2015 results

Temporal graphs are structured

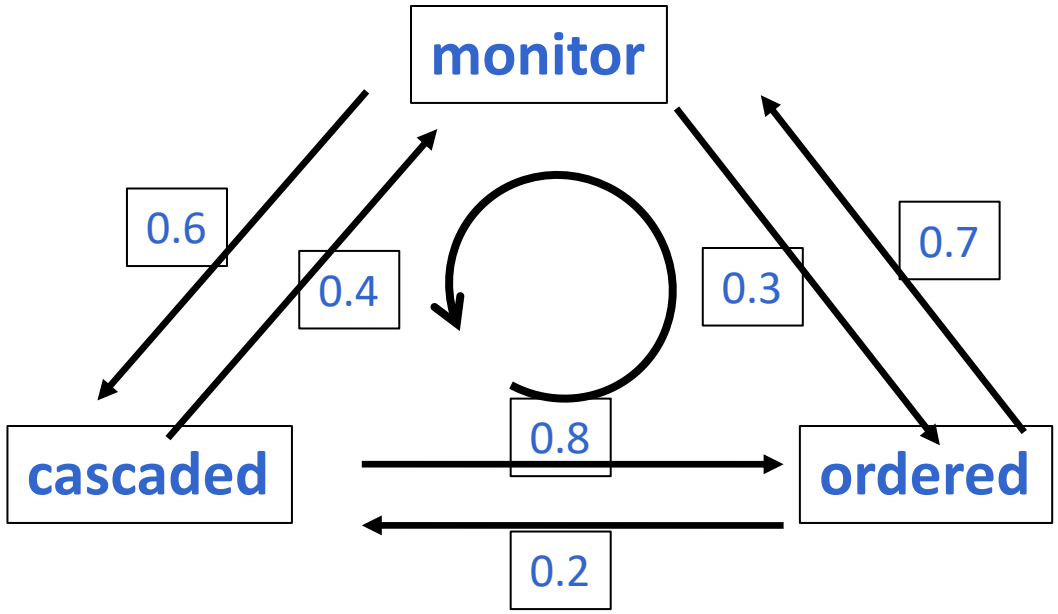
Due to transitivity, TempRels are not independent

First step: global inference



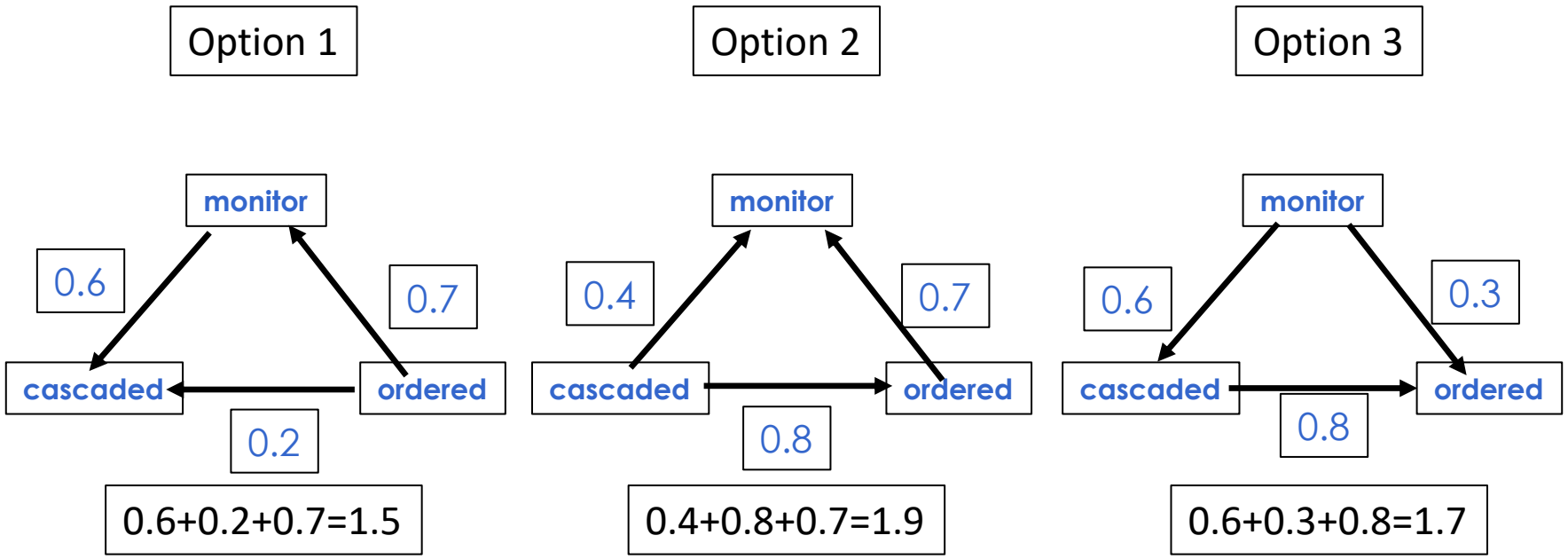


I. Global inference (a toy example)



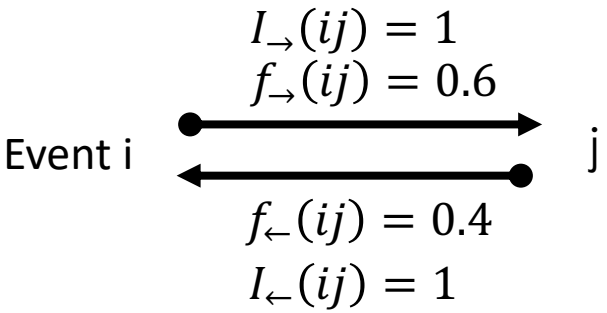
Resulting temporal graph is not feasible

Global inference (a toy example)



We should not only select the assignment with the **best score**, but also one that **does not violate our constraints** (here: transitivity). Formulated as an ILP (Roth & Yih 2004)

Global inference via ILP



Integer Linear Programming (ILP)

real variable

$$\hat{I} = arg\max_{I} \sum_{i<j} \sum_{r} f_r(ij)\ I_r(ij)$$

Boolean variable

s.t. $\forall i, j, k$

$$\sum_{r} I_r(ij) = 1, \qquad I_{r1}(ij) + I_{r2}(jk) - I_{r3}(ik) \leq 1$$

Uniqueness

Transitivity (no loops)

We're maximizing the score of an entire graph **while enforcing transitivity constraints**.

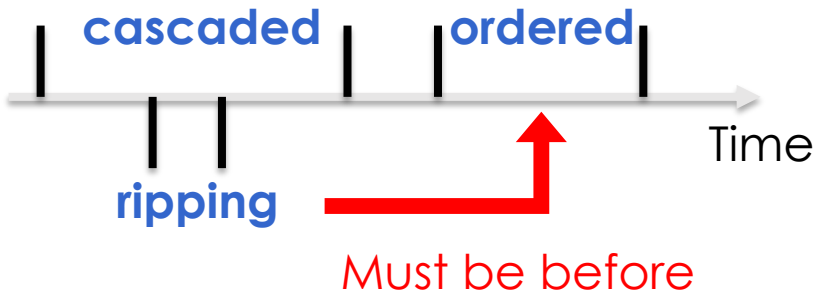
Global Inference is essential. But, **how should we train the models** $I_r(ij)$?

II. Local learning is not sufficient

tons of earth **cascaded** down a hillside,

…**ripping** two houses…firefighters **ordered** the evacuation of nearby homes…

- Q: (**ripping**, **ordered**)=? (difficult even for humans)
- Annotation says "before", but training this way without accounting for the other labels – eg., the **relation** to "cascaded" misleads the model and leads to <u>overfitting</u>



- **Jointly training** the relations is essential
  - ☐ Structured Learning

III. Temporal Common Sense

- More than 10 people have (**event1:**), police said. A car (**event2**:) on Friday in a group of men.

- Which event came first, **event1** or **event2**?
- Hard to tell.

- What about now?
  - Context is important, but not sufficient
  - Humans have good priors about which event “usually” happens before another.
- We would like to acquire it, and use it to enhance our temporal relation identification

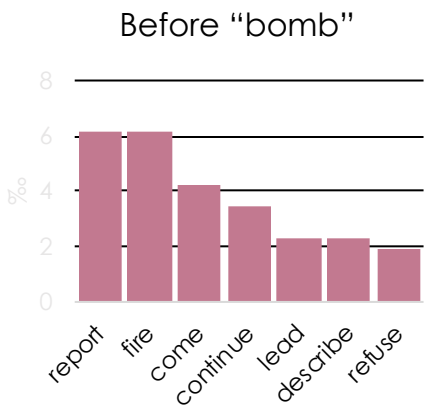
TemProb: Temporal Relation Probabilistic Knowledge Base

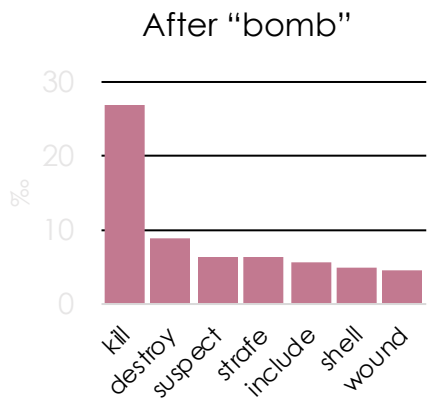
- Run initial system on New York Times 1987-2007, #Articles~1M
- Identify events; identify temporal order
- 80M temporal relations
- Noisy statistics is sufficient to give good priors.

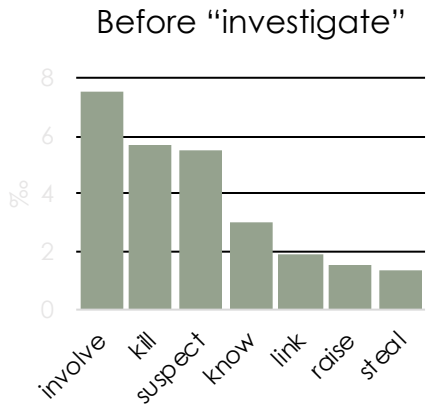
Example pairs		Temporal Before (%)	Temporal After (%)
Text Before	**Text After**		
Ask	Help	86	9
Attend	Schedule	1	82
Accept	Propose	10	77
Die	Explode	14	83

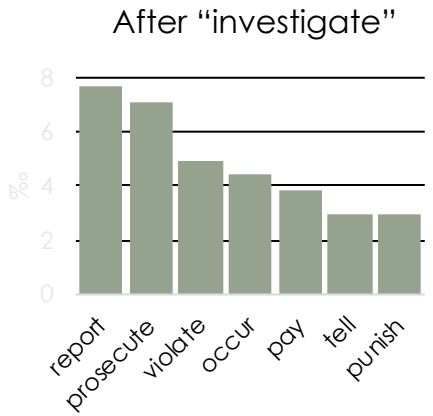
Priors on order are often different than order of occurrence in text

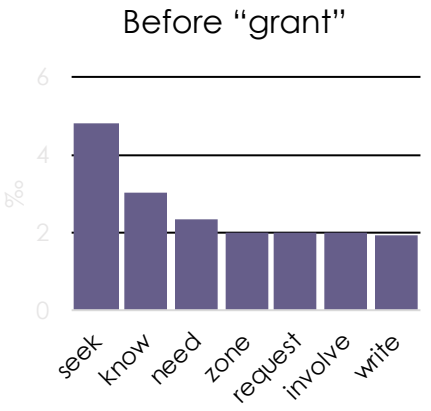
Event distributions from TemProb

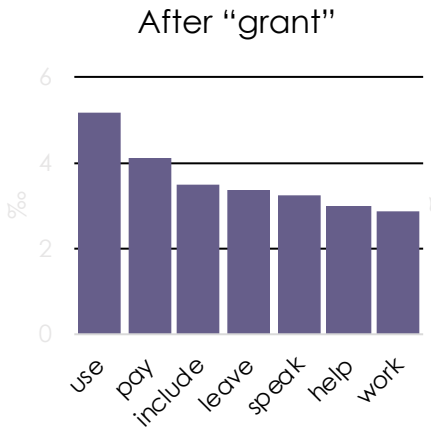


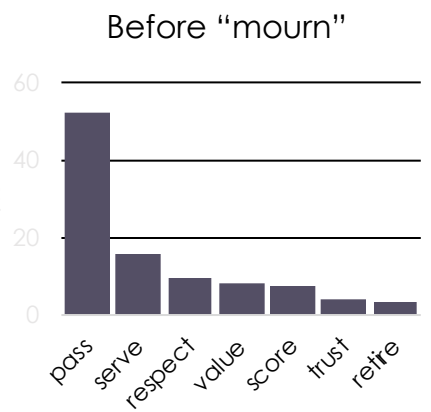


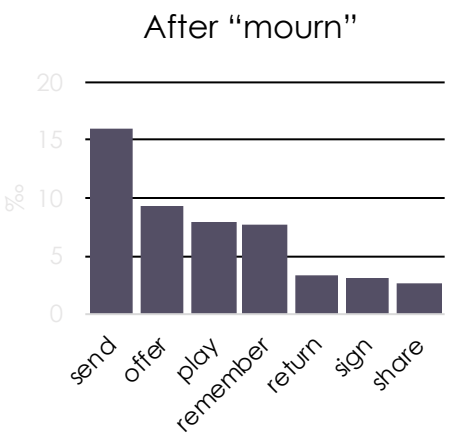




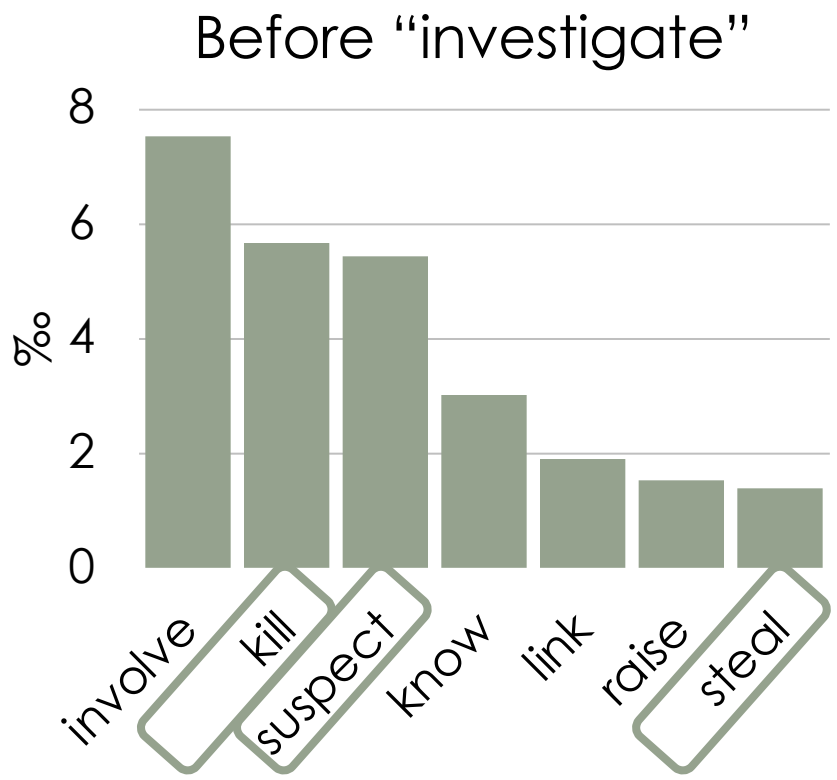


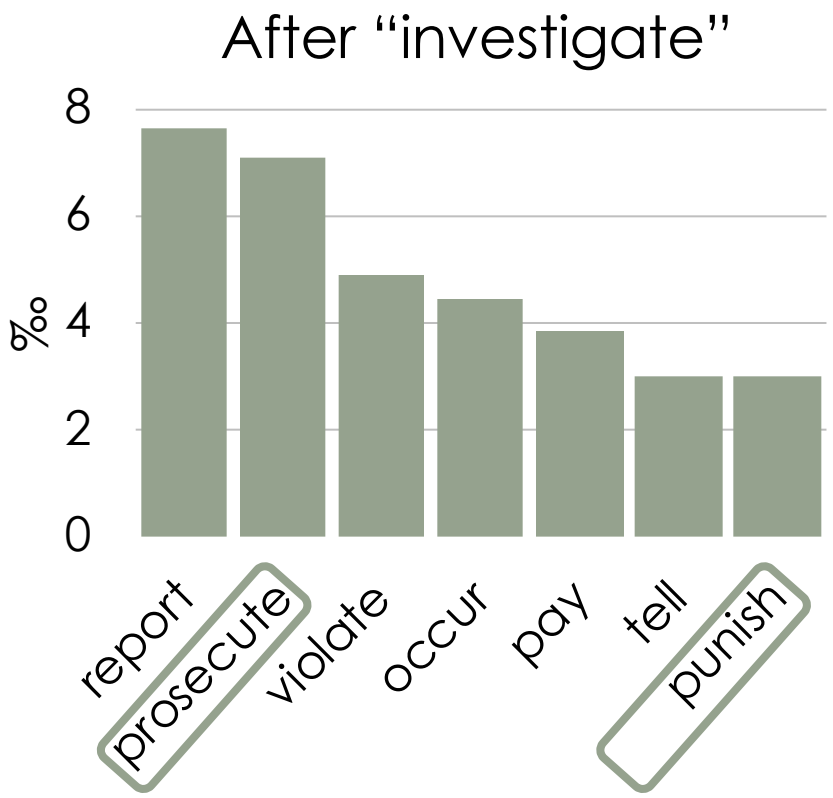




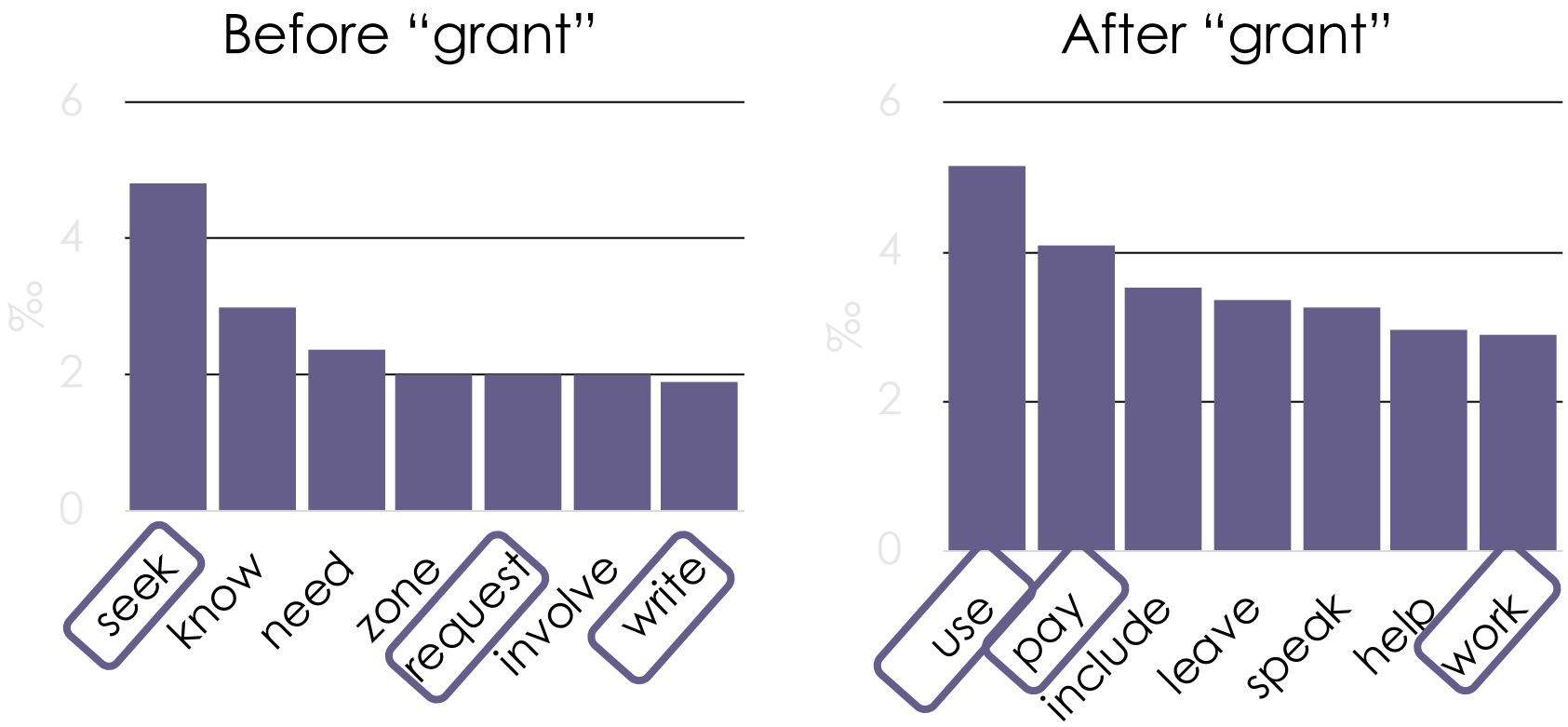


Event distributions from TemProb

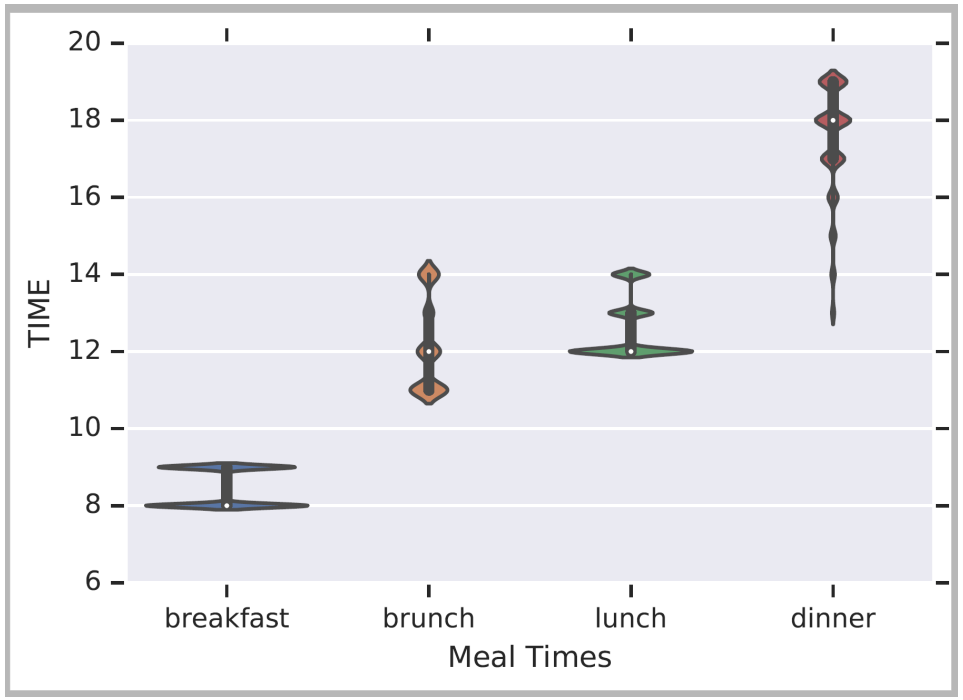




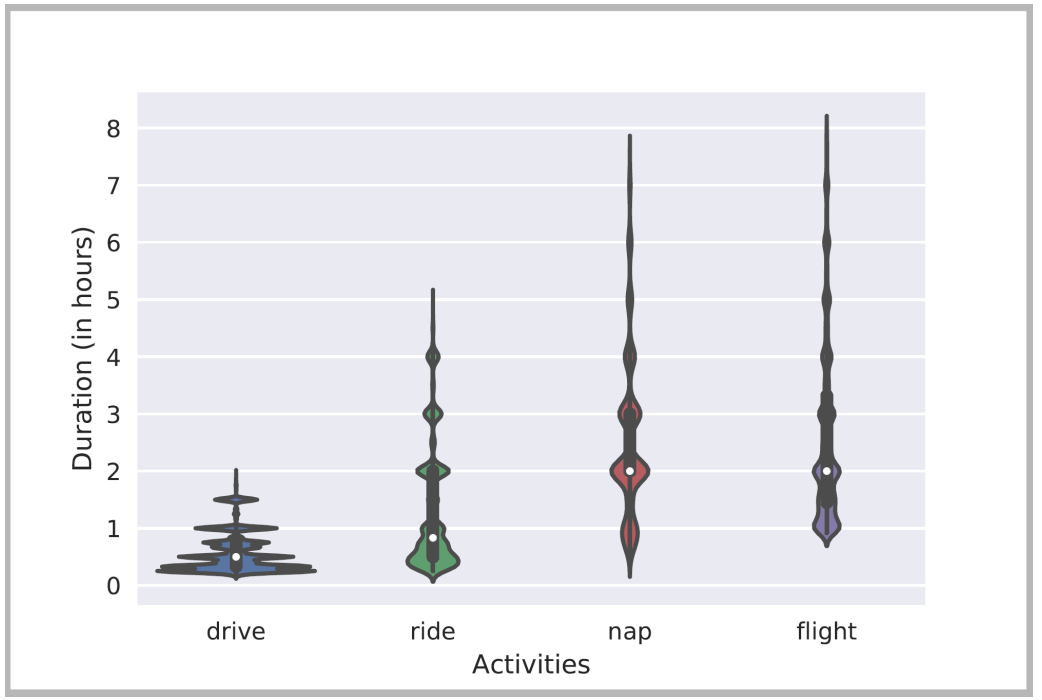
Event distributions from TemProb



Many Forms of Temporal Common Sense



Typical Time



Duration

- Frequency, Periodicity [ACL'19]
- Not being used in this talk

III. Making use of TemProb

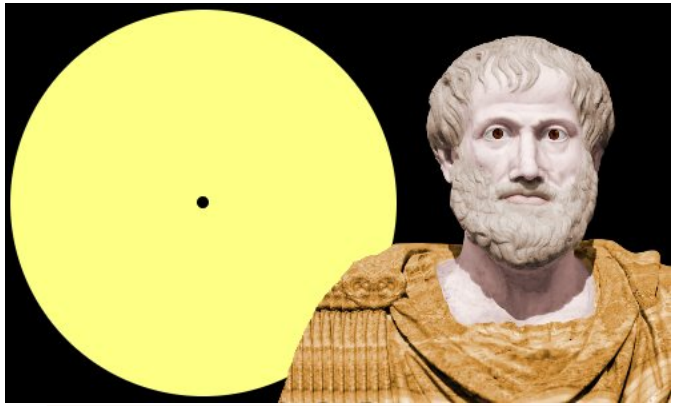
- Let $C(v_i, v_j, r)$ be the number of appearances of (v_i, v_j) classified to be relation r.
- For each pair of verb events,
- $h_r(v_i, v_j) = \frac{C(v_i, v_j, r)}{\sum_{r'} C(v_i, v_j, r')}$ is the **prior probability of the pair having relation r**.

- **Learning:** Use the prior probability as an additional feature and retrain our system.
- **Inference:** Use as a regularization term in the objective.

$$\hat{I} = arg\max_{I} \sum_{i<j} \sum_{r} (f_r(ij) + \boldsymbol{h_r(ij)}) I_r(ij)$$

It's Time for Reasoning: **Outline**

- Reasoning about **events** and **time** in natural language
  - Temporal ordering of events
    - Learning & Inference paradigms to support reasoning
  - Temporal common sense
  - Reasoning & Supervision paradigms



Did Aristotle have a laptop?

- Reflects an important move in NLU from **sentence level** to **situation level**
- Addresses issues in combining learning and reasoning, and **supervision.**

- More about Temporal Common Sense

- Initial thoughts on additional Reasoning paradigms
  - Decomposition, and computing functions over sets of variables

Putting it all together: A Neural Approach

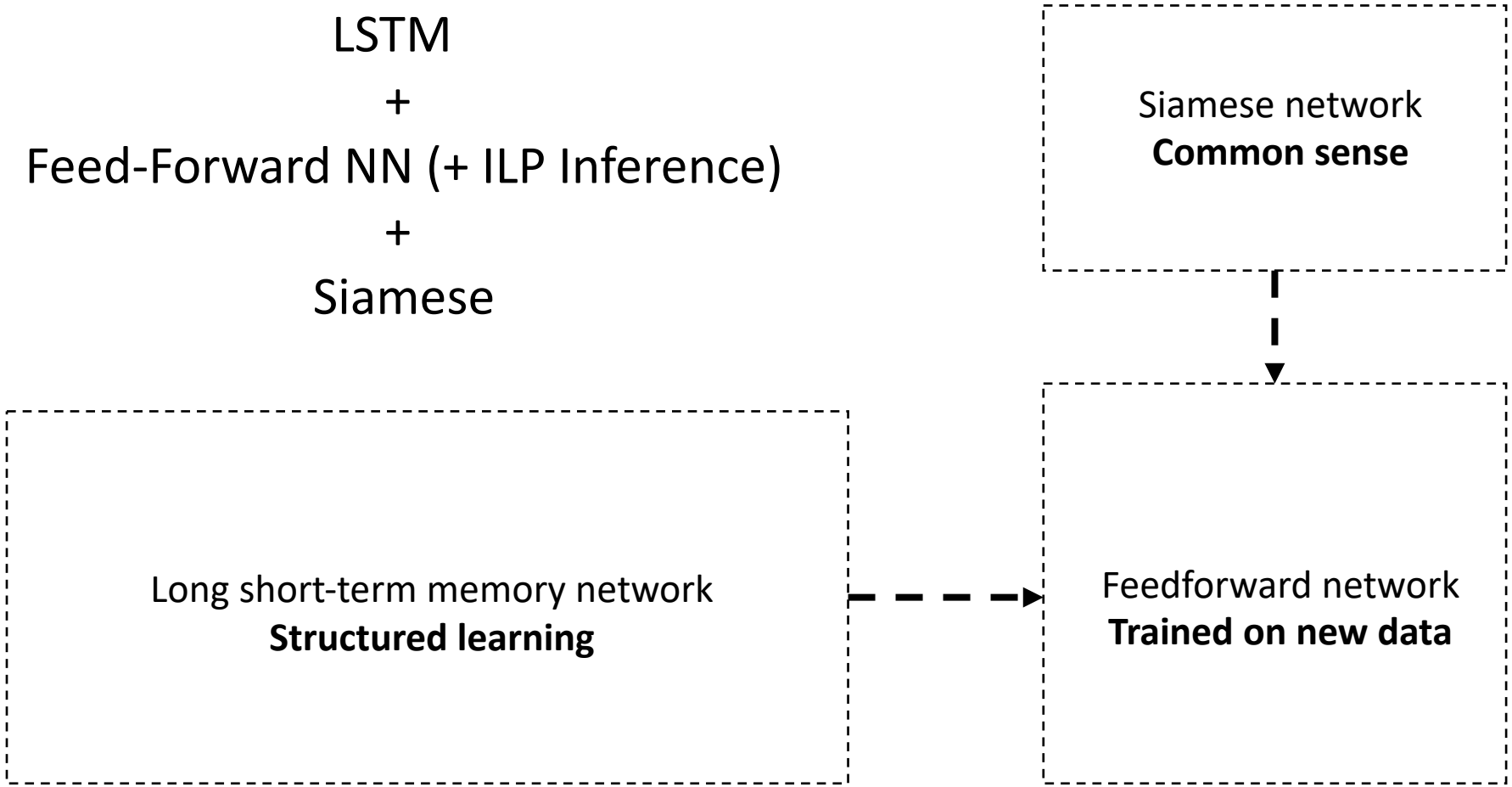
LSTM

+

Feed-Forward NN (+ ILP Inference)

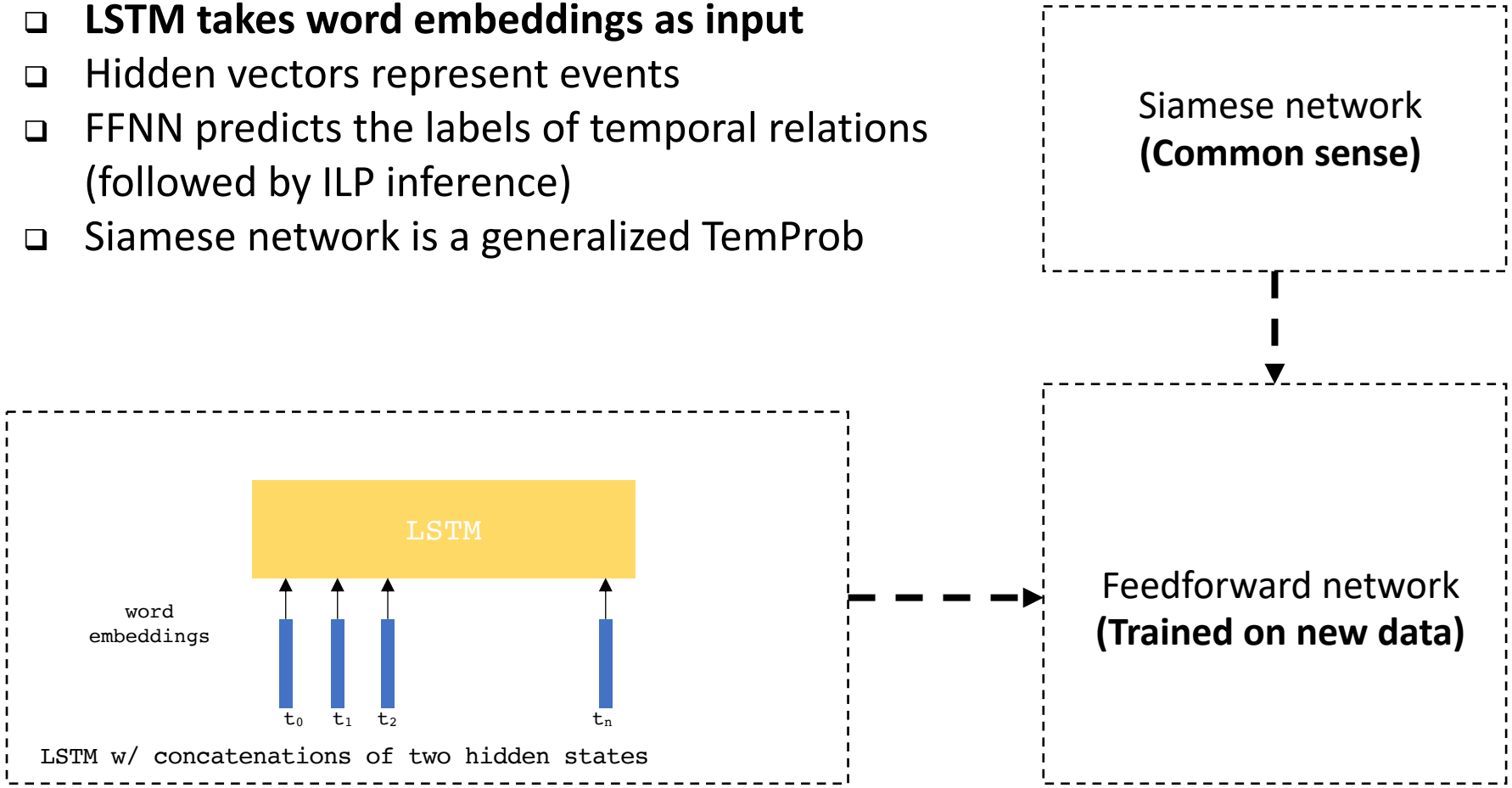
+

Siamese



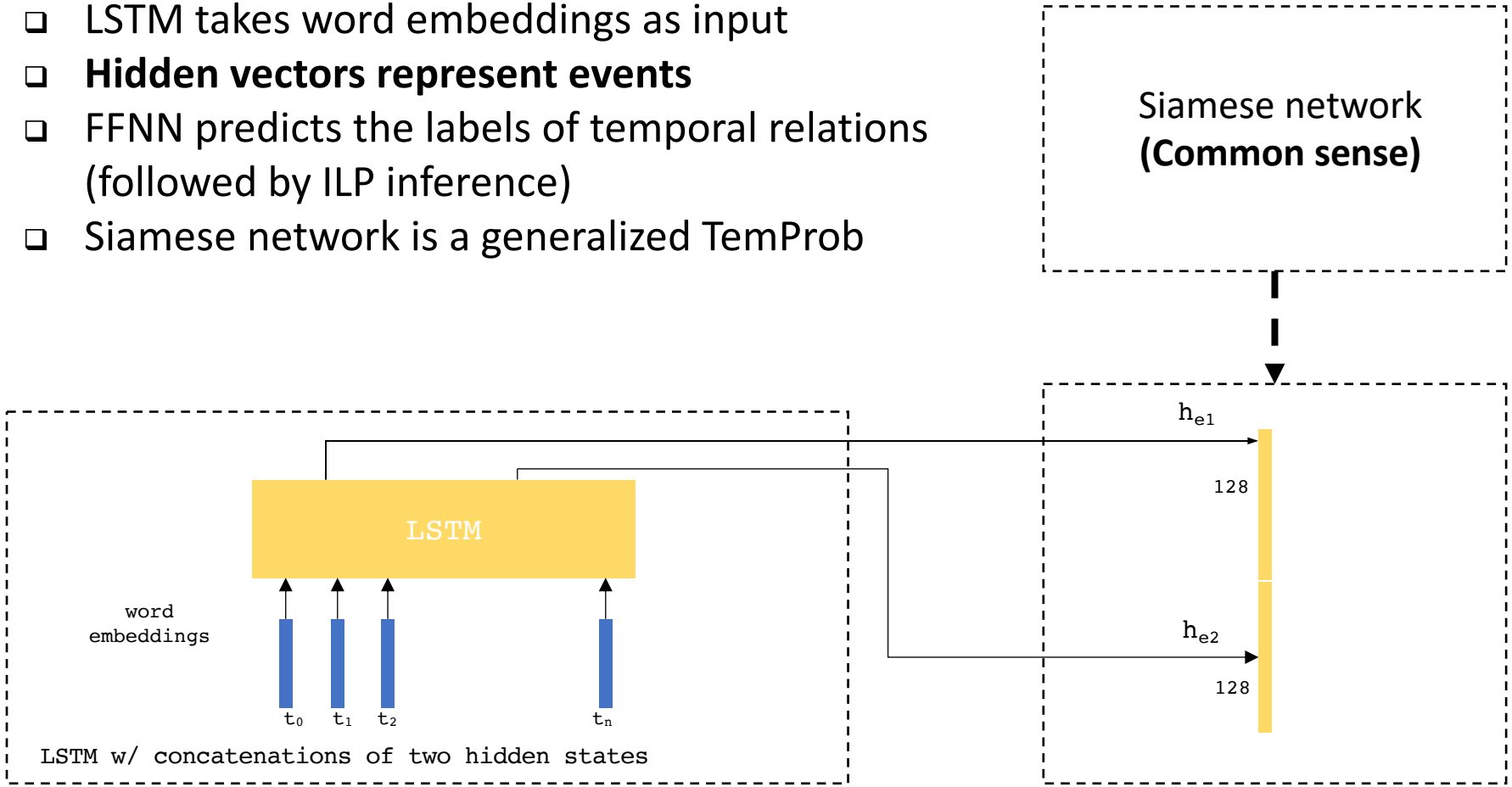
Putting it all together: A Neural Approach

- **LSTM takes word embeddings as input**
- Hidden vectors represent events
- FFNN predicts the labels of temporal relations (followed by ILP inference)
- Siamese network is a generalized TemProb



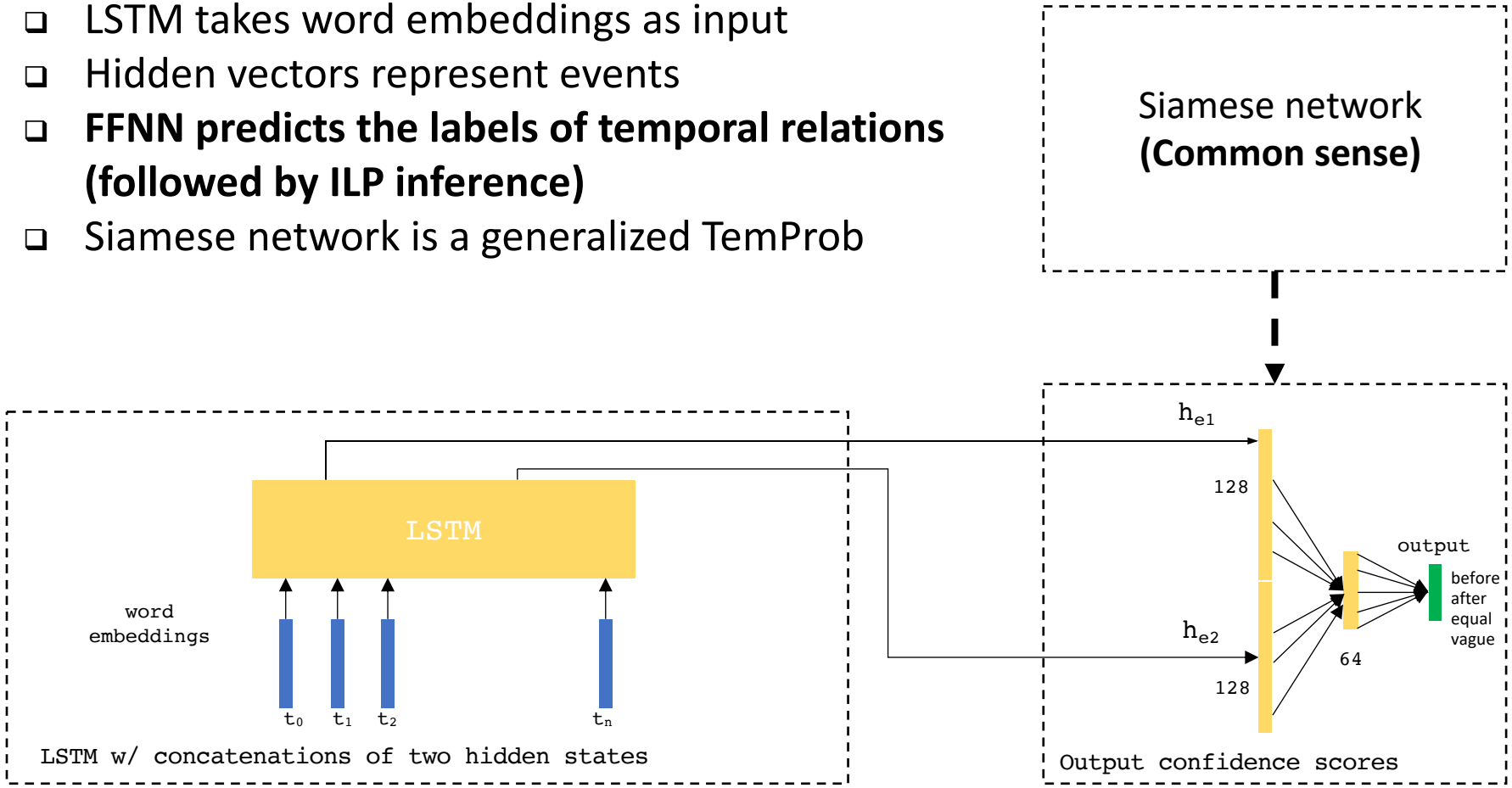
Putting it all together: A Neural Approach

- LSTM takes word embeddings as input
- **Hidden vectors represent events**
- FFNN predicts the labels of temporal relations (followed by ILP inference)
- Siamese network is a generalized TemProb



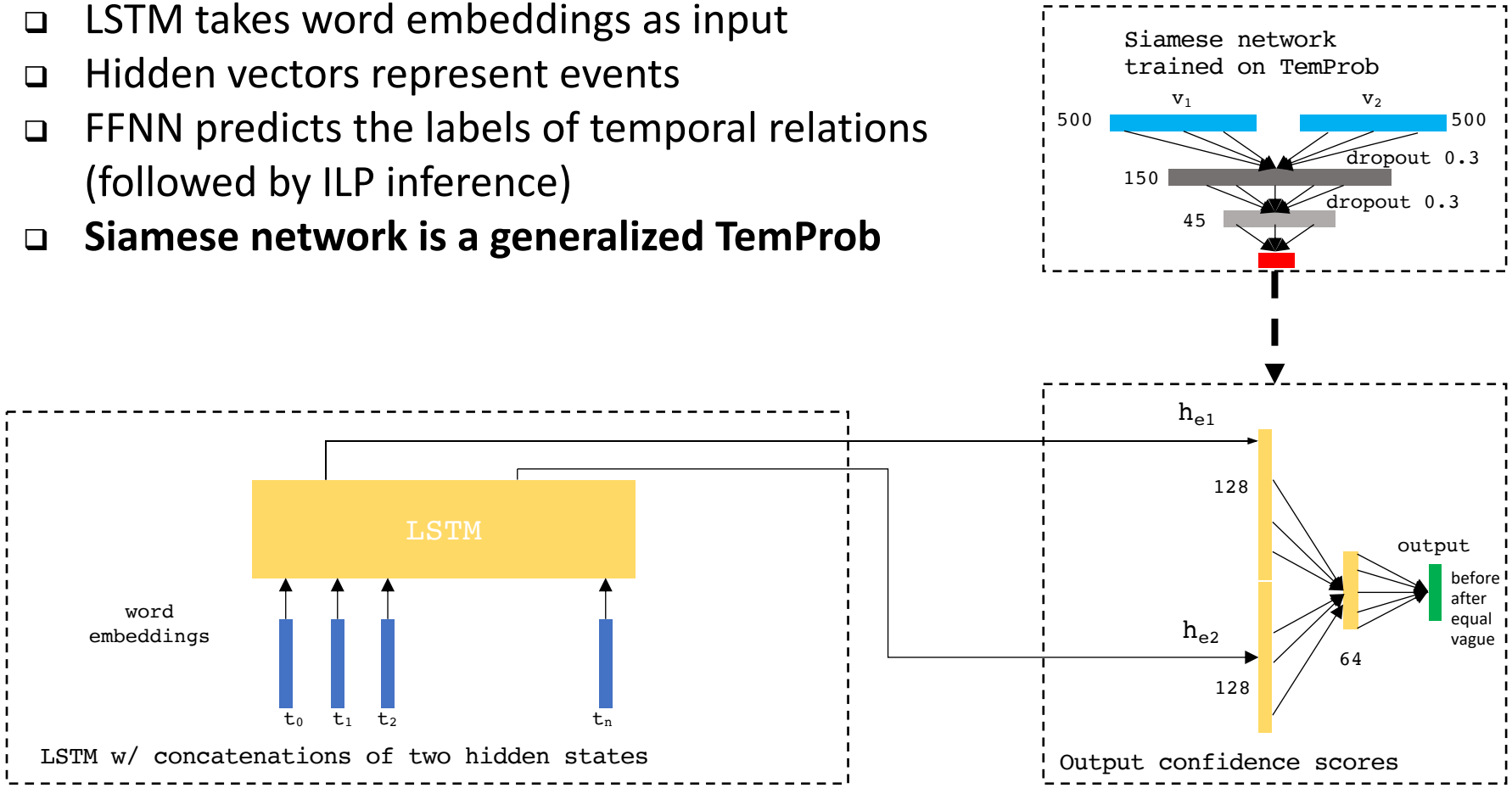
Putting it all together: A Neural Approach

- LSTM takes word embeddings as input
- Hidden vectors represent events
- **FFNN predicts the labels of temporal relations (followed by ILP inference)**
- Siamese network is a generalized TemProb



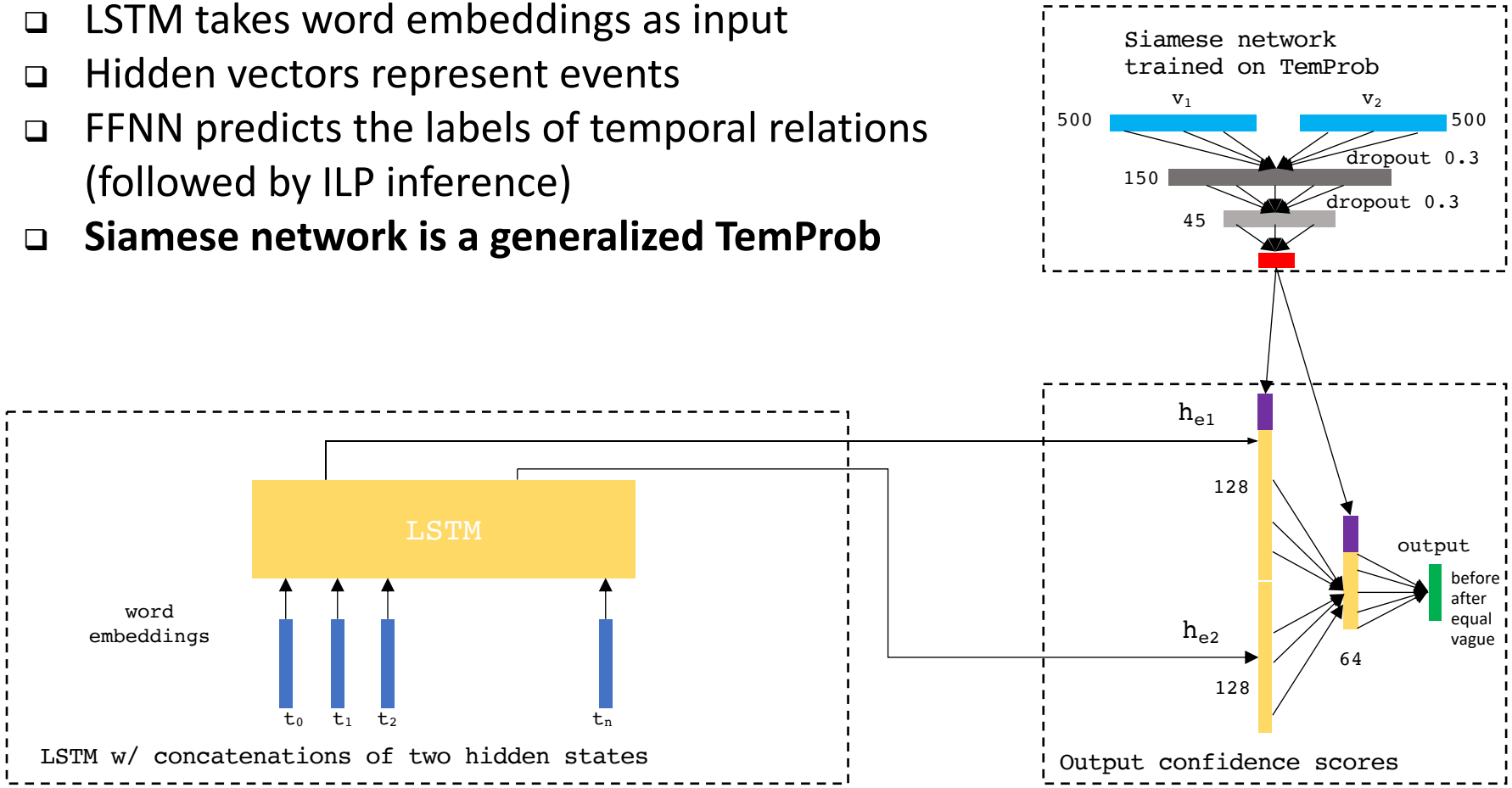
Putting it all together: A Neural Approach

- LSTM takes word embeddings as input
- Hidden vectors represent events
- FFNN predicts the labels of temporal relations (followed by ILP inference)
- **Siamese network is a generalized TemProb**

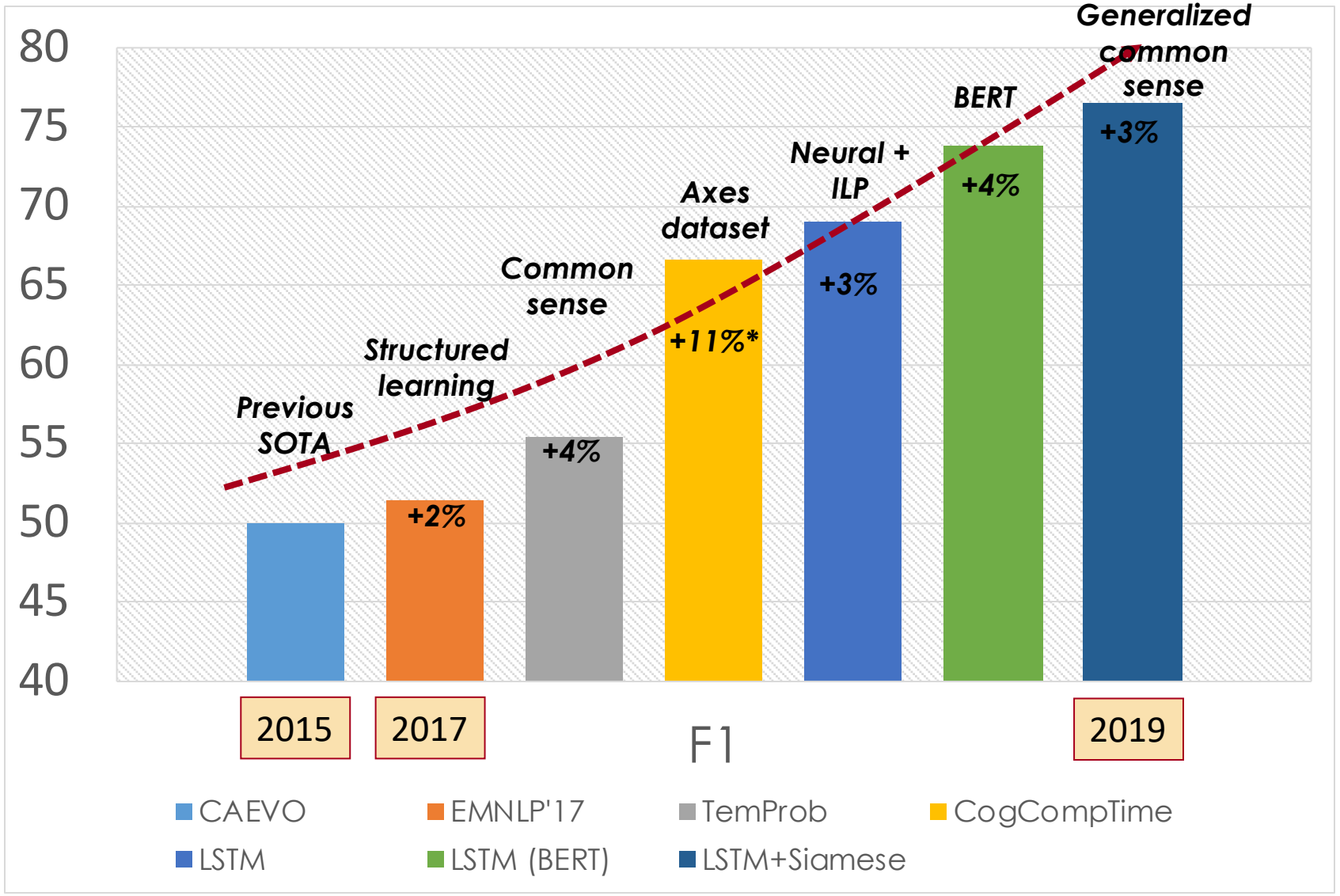


Putting it all together: A Neural Approach

- LSTM takes word embeddings as input
- Hidden vectors represent events
- FFNN predicts the labels of temporal relations (followed by ILP inference)
- **Siamese network is a generalized TemProb**



A Revolution in Temporal Reasoning



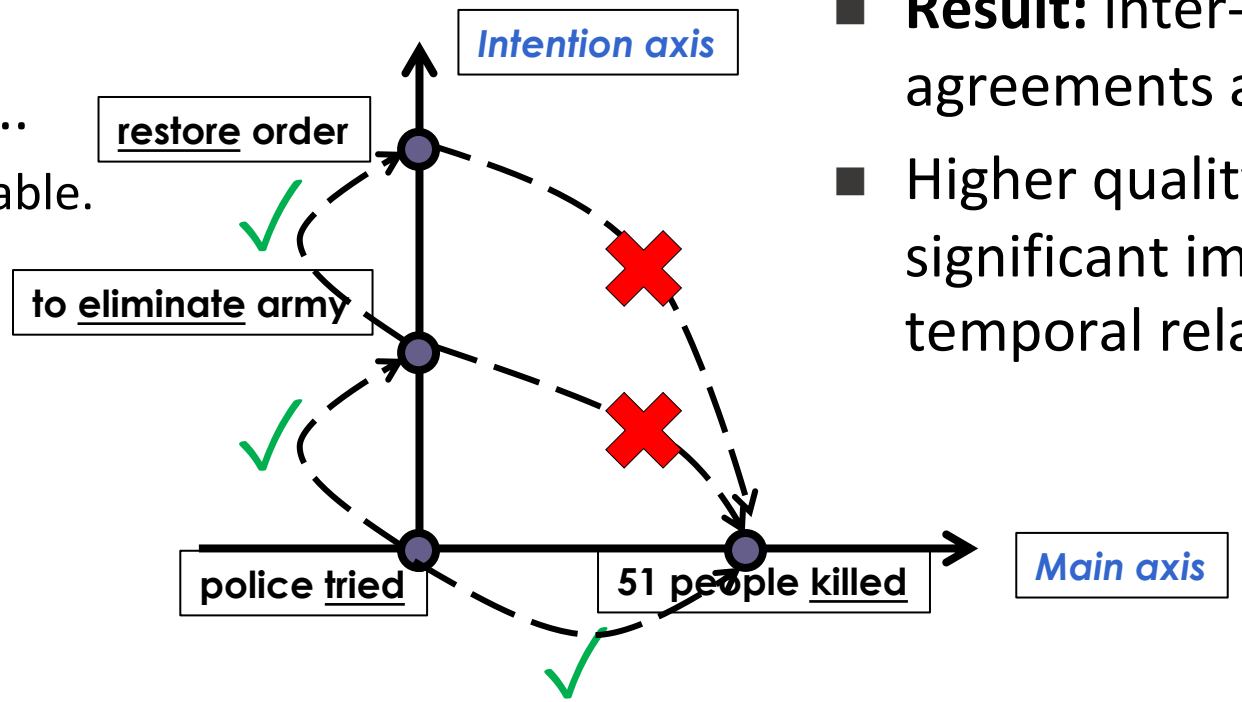
Thesis work of Qiang Ning (2019)

Multi-Axis View of Events [ACL'18]

- The improvements involves another conceptually important step:
  - We suggest that not all events are comparable
  - **Events reside on multiple axes**.

*Police **tried** to **eliminate** the pro-independence army and **restore** order. At least 51 people were **killed** in clashes between police and citizens in the troubled region.*

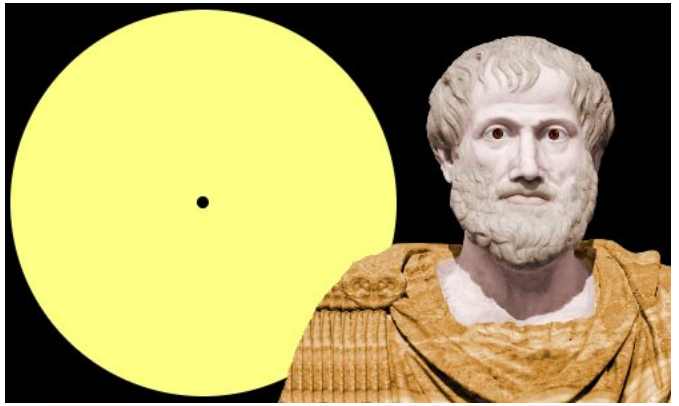
- **Events could be actual, hypothetical, intentions**,..
  - Not all events are comparable.



- **Result:** inter-annotator agreements are much higher.
- Higher quality training data, with significant improvement in temporal relation identification

It's Time for Reasoning: **Outline**

- Reasoning about **events** and **time** in natural language
  - Temporal ordering of events
    - Learning & Inference paradigms to support reasoning
  - Temporal common sense
  - Reasoning & Supervision paradigms



Did Aristotle have a laptop?

- Reflects an important move in NLU from **sentence level** to **situation level**
- Addresses issues in combining learning and reasoning, and **supervision.**

- More about Temporal Common Sense

- Initial thoughts on additional Reasoning paradigms
  - Decomposition, and computing functions over sets of variables

What kinds of supervision signals do can we get/need?

Inducing Semantics

- Inducing semantic representations and making decisions that depend on it require learning and, in turn, supervision.
- Standard machine learning methodology:
  - Given a task
  - Collect data for the task and annotate it
  - Learn a model **[doesn't matter how]**
- We will never have enough annotated data to train all the models, for all the tasks we need, this way.
  - We don't even know what are "all the tasks"
  - Most of what **we** learn we don't learn by "training" on many examples
- Current methodology is not scalable and, often, makes no sense
  - Annotating for complex tasks is difficult, costly, and sometimes impossible.
    - Most decisions we care about are too sparse to be trained for directly

It's ok to do supervised learning.
But what about tasks that cannot be supervised directly?

In most interesting cases, learning should be (and is) driven by **incidental supervision** signals [Roth AAAI'17]

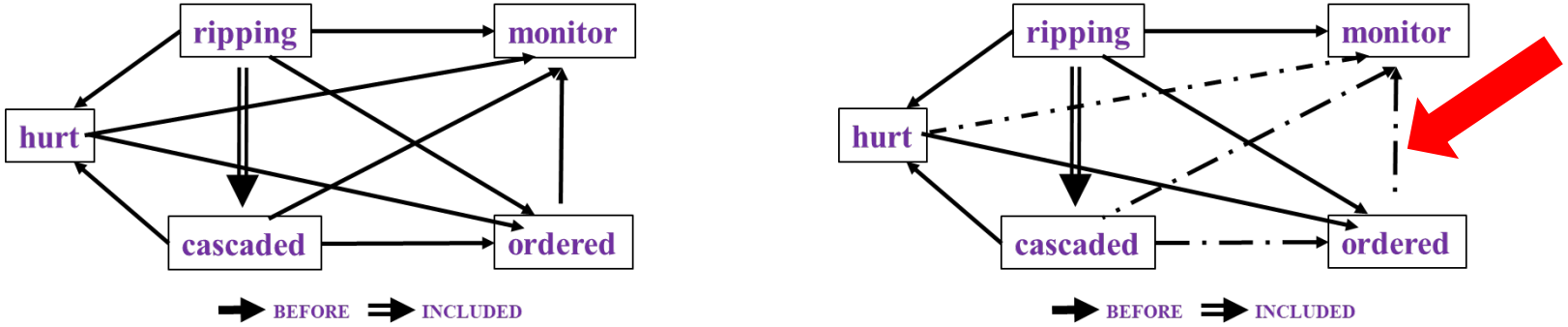
Incidental supervision: How to understand, acquire and use signals that were not put there to help a specific target task.

Today: how does the ability to reason (a little bit) helps supervision?

Temporal Relations [Ning et al. *SEM'2018; ACL'18, EMNLP'18]

Two key questions: **(i)** What level of supervision is **really needed**?
(ii) Algorithmic Approach: how to gain from it?

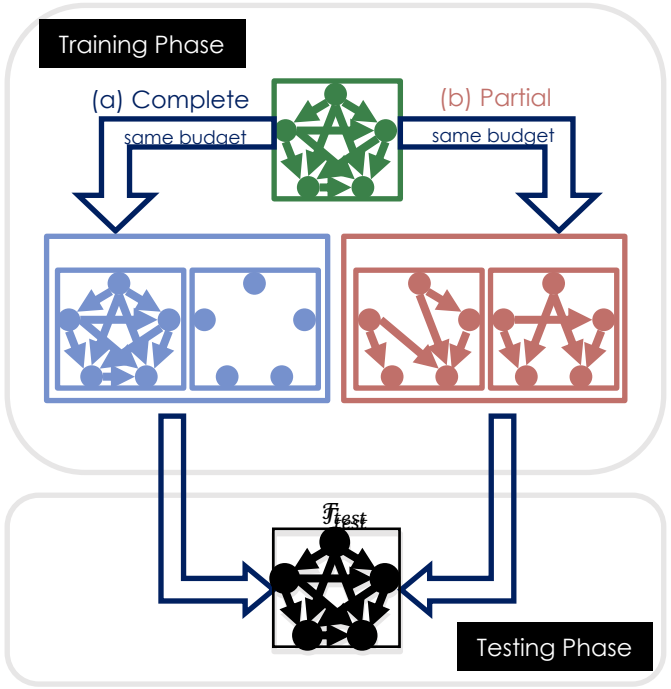
- In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.

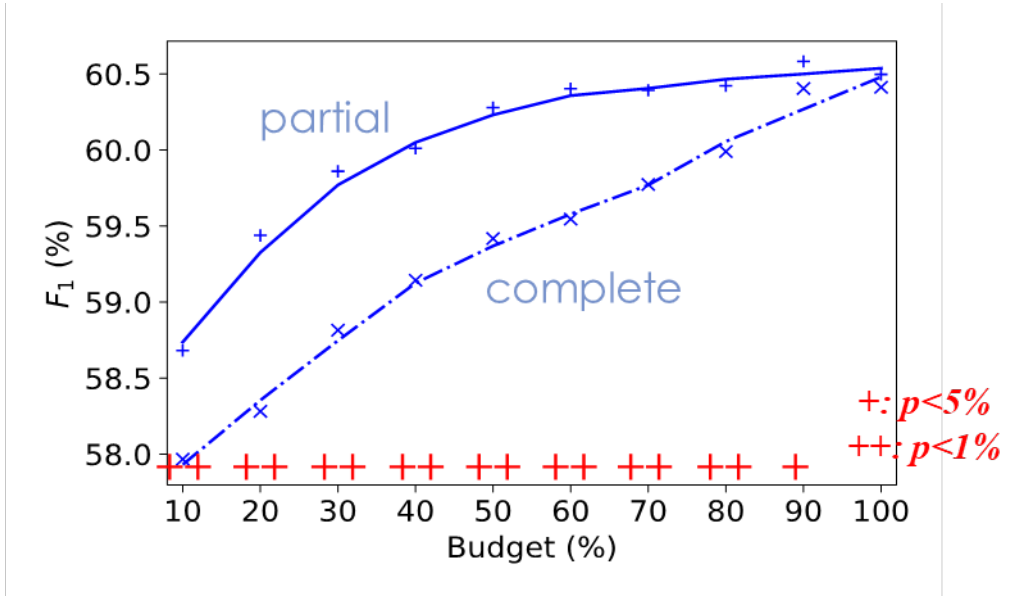


- Very difficult task— hinders exhaustive annotation ($O(N^2)$ edges)
- But, it's rather easy to get partial annotation – some relations.
- And, we have **strong expectations** from the output
  - Transitivity
  - Some events tend to precede others, or follow others [Ning et. al., NAACL'18]

Intuition:
- If the **structure** is tight, there is no need to annotate all the variables.
- A partial set of signals can supervise all.

Partial or complete, that's the question [NAACL'19]





- When we have a fixed budget, partial structures indeed lead to better performance.
- How should we quantify the quality of "partial"?

Structure

- **Structure:** a vector of random variables: $Y = [Y_1, Y_2, \dots, Y_d] \in C(\mathcal{L}^d)$
  - $\mathcal{L}$ is the label set
  - $Y \in C(\mathcal{L}^d) \subseteq \mathcal{L}^d$ represents the constraints imposed by this type of structure.
- **(Generalized) Annotation**:
  - $\boldsymbol{k}$ out of d variables are labeled → a subset of $C(\mathcal{L}^d)$
  - Let $\boldsymbol{f_k}$ **be the size of the feasible subset**
  - $f_0 = |C(\mathcal{L}^d)| \geq f_1 \geq f_2 \geq \cdots \geq f_d = 1$

Constrained output; not all assignments are possible.

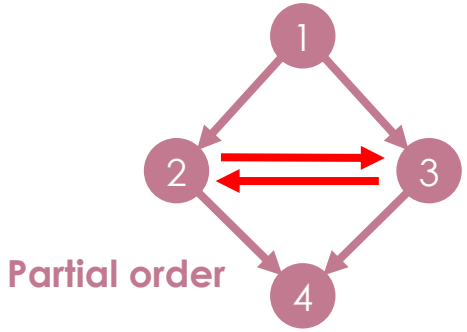
Intuition:
- If the structure is tight, there is no need to annotate all the variables.
- A partial set of signals can supervise all.

No annotation

Complete annotation

- **Define the benefit of k labels:** $\boldsymbol{I_k \triangleq \log |C(\mathcal{L}^d)| - E[\log f_k]}$

How much of $C(\mathcal{L}^d)$ has been **disqualified** by k labels



Partial order

1→2→3→4

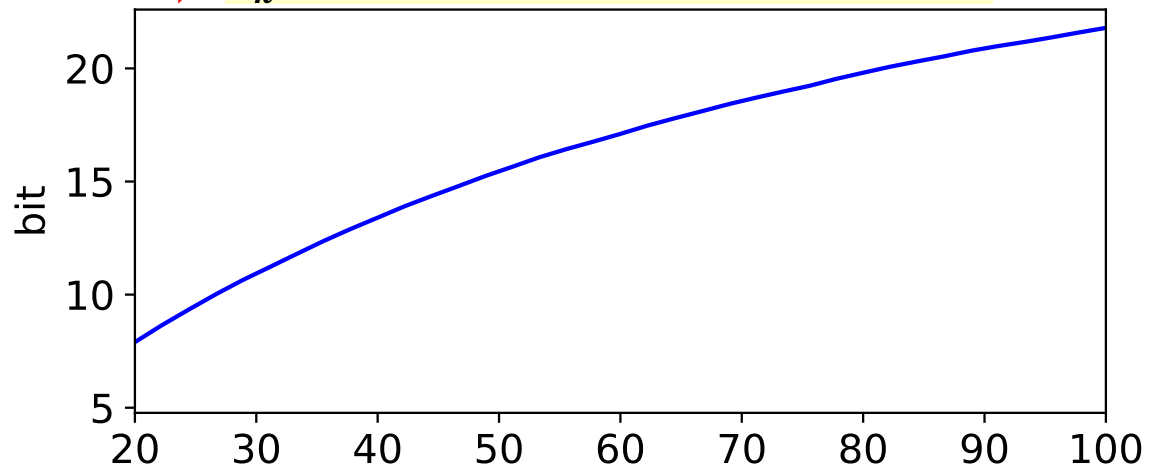
Or

1→3→2→4

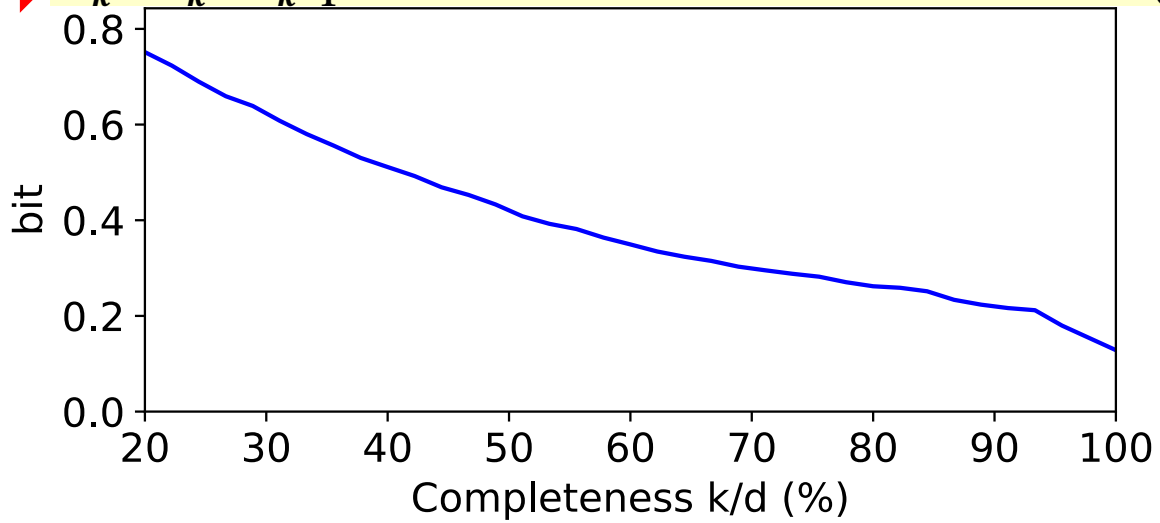
$f_k = 2$

Diminishing Return of New Labels

I_k: The benefit of k labels is concave



$\Delta_k = I_k - I_{k-1}$: The benefit of a new label is diminishing



Improvement Consistent With Tightness Analysis By I_k

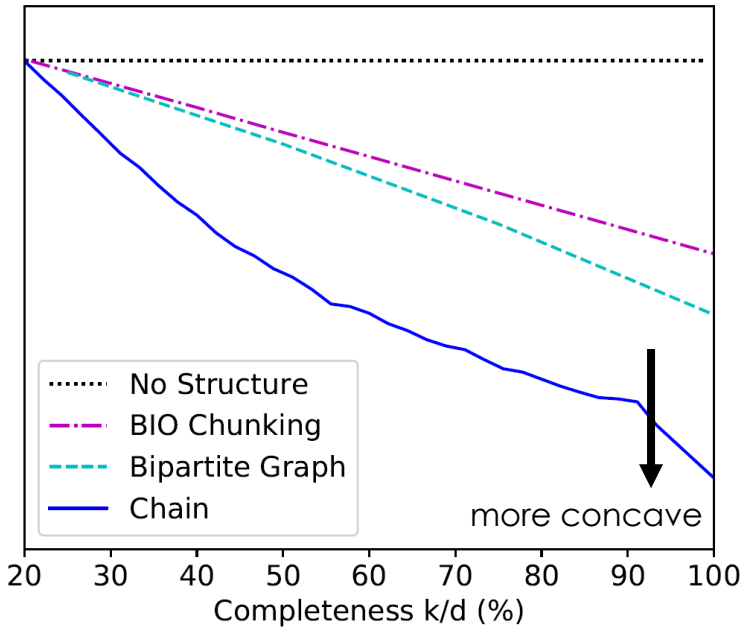
Algorithmically:

- Results are consistent with tightness of the structure
- A version of the Constraint-Driven Learning (CoDL) Algorithm [Chang et al. 2012]

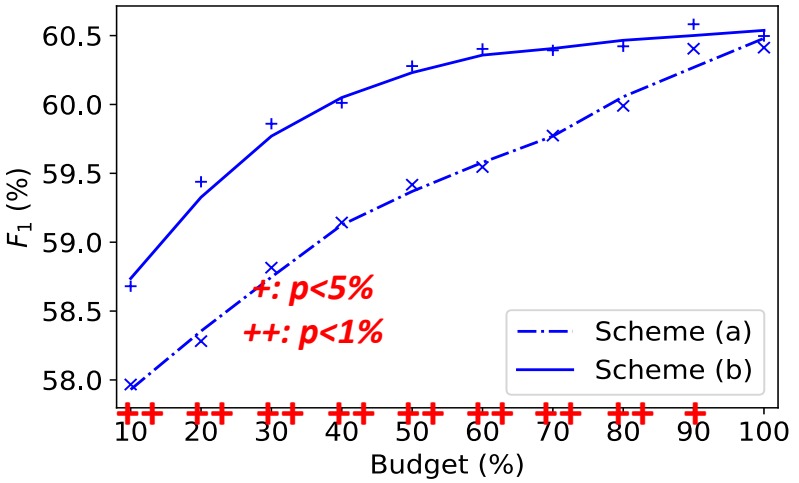
Indeed:

- If the structure is tight, there is no need to annotate all the variables.
- A partial set of signals can supervise all.

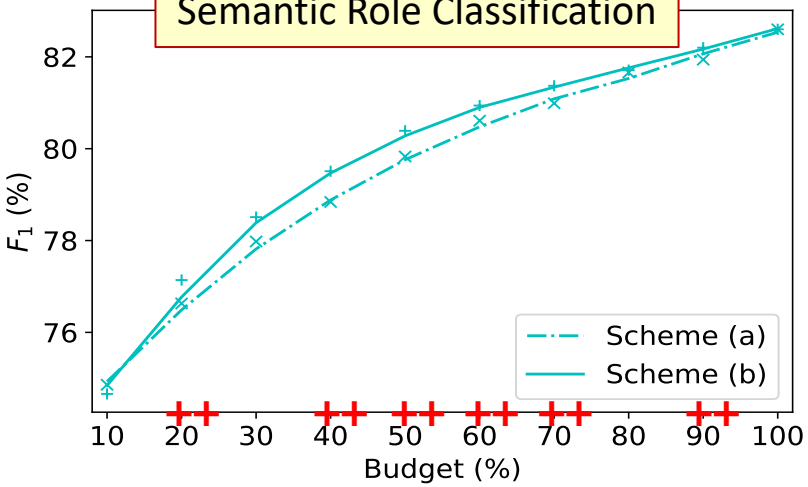
$\Delta_k = I_k - I_{k-1}$: The benefit of a new label



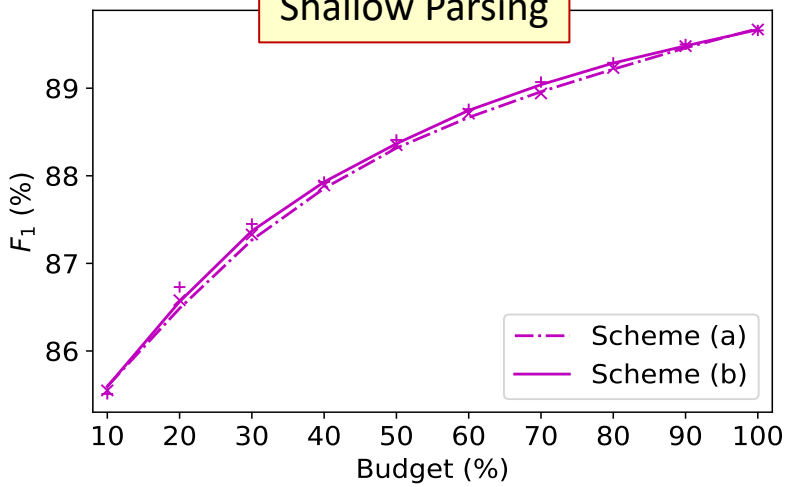
Chain:
Temporal Relation Extraction



Bipartite Graph:
Semantic Role Classification



Seq Tagging
Shallow Parsing



What is I_k Actually?

- Definition: A $\boldsymbol{k}$-partial annotation A_k is a vector of random variables $\boldsymbol{A_k} = [A_{k,1}, A_{k,2}, \ldots, A_{k,d}] \in (\mathcal{L} \cup \sqcap)^d$, where $\sqcap$ is a special character for no label yet, such that
  - $\sum_{i=1}^{d} \mathbb{I}(A_{k,i} \neq \sqcap) = k$
  - $P(Y|A_k = a_k) = P(Y|Y_j = a_{k,j}, j \in \mathcal{J})$, where $\mathcal{J} = \{j : a_{k,j} \neq \sqcap\}$
  - A_k means k variables in Y are labeled, and they are correct

- **Theorem:** I_k is the mutual information between Y and A_k when both Y and the k variables labeled in A_k follow uniform distributions.

How Reasoning Helps Learning

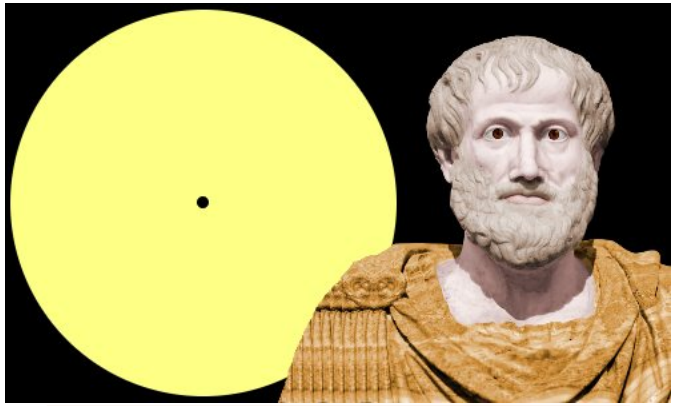
- It provides a reduction in the uncertainty of a target structure Y, by the annotation random process A
  - Here we reasoned from the "partial" annotation to the complete one.

- More generally, we argue:
  - Any signal that has non-zero mutual information with Y can be viewed as "annotation"
  - Since it allows us to "reason" from it to the complete annotation needed.
- Points out a way to understand and quantify the value of indirect supervision signals.
  - Refine the theory
  - Rather than annotate a data set at the events and relations level, answer some questions relative to it.
  - Rather than annotating topics of documents, use Wikipedia to "understand" the topic means; then classify
  - …….

Work in Progress

It's Time for Reasoning: **Outline**

- Reasoning about **events** and **time** in natural language
  - Temporal ordering of events
    - Learning & Inference paradigms to support reasoning
  - Temporal common sense
  - Reasoning & Supervision paradigms



Did Aristotle have a laptop?

- Reflects an important move in NLU from **sentence level** to **situation level**
- Addresses issues in combining learning and reasoning, and **supervision.**

- More about Temporal Common Sense

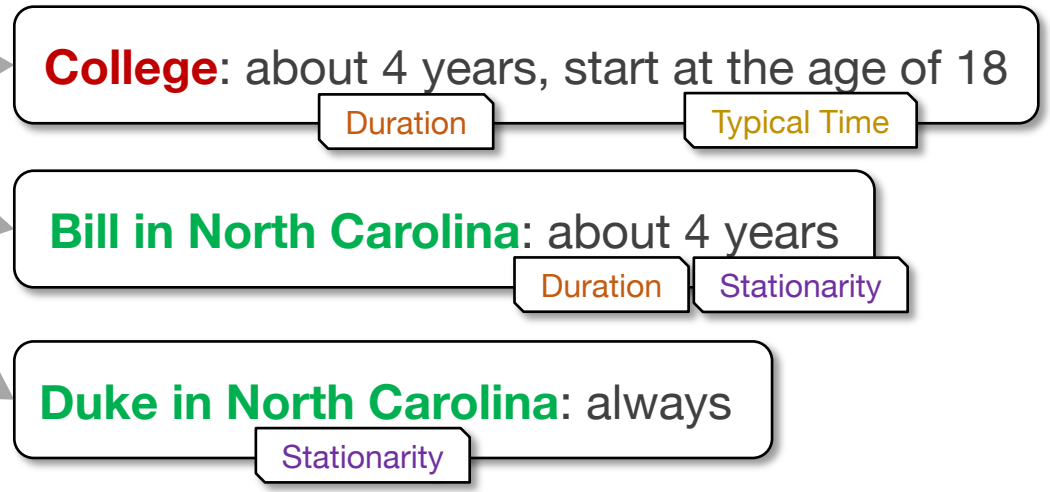
- Initial thoughts on additional Reasoning paradigms
  - Decomposition, and computing functions over sets of variables

Temporal Common Sense [EMNLP'19 + In Progress]

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

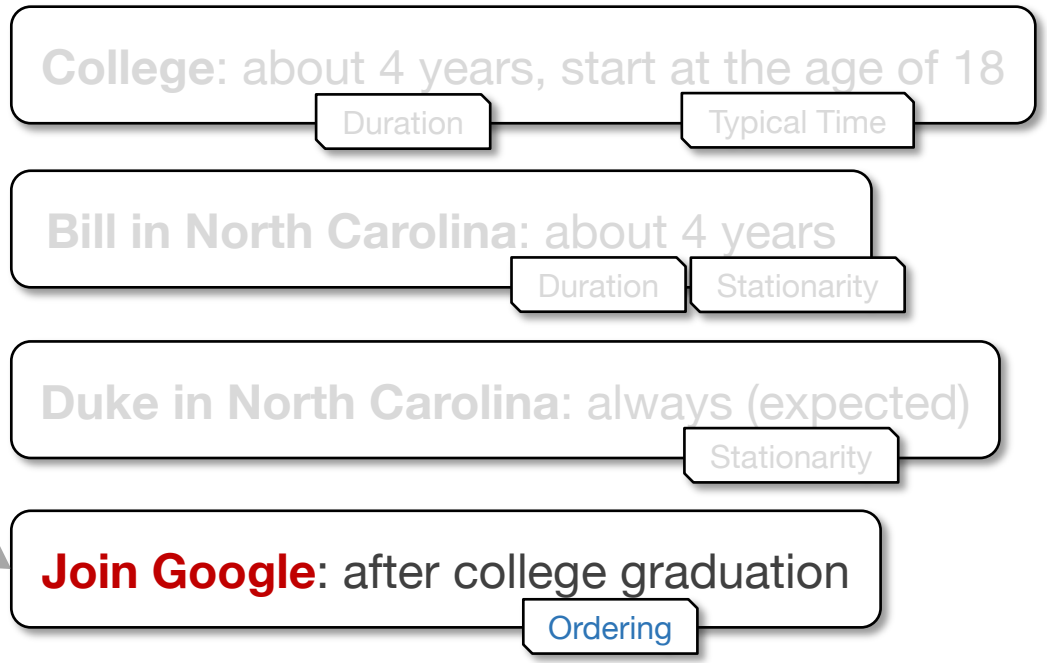
Temporal Common Sense

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.



Temporal Common Sense

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

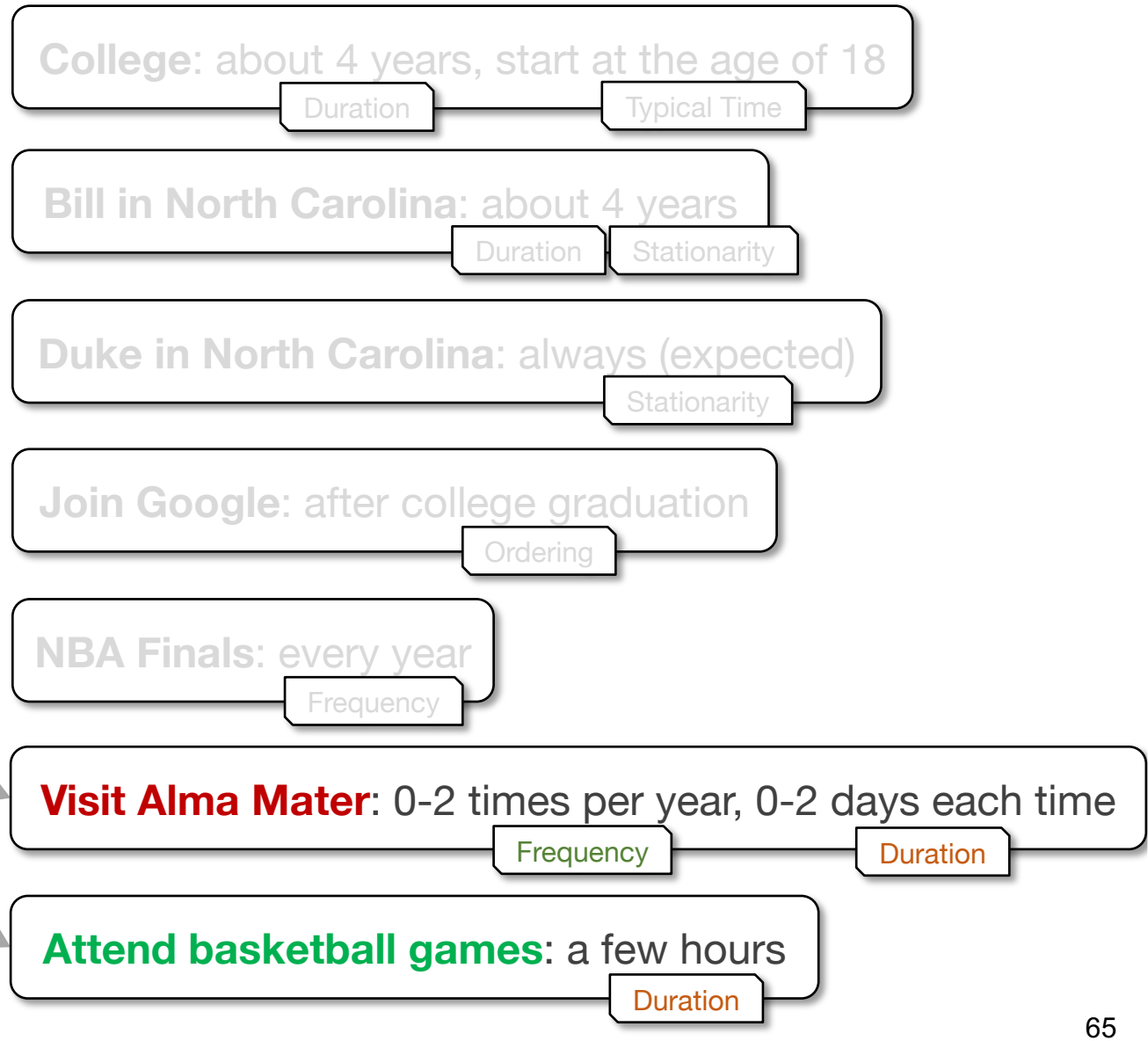


Temporal Common Sense

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

Temporal Common Sense

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.



Temporal Commonsense

My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.

- Human infer temporal common sense that helps them to better understand the story.
- This is reflected in the ability to answer questions about temporal aspects.

- **Q: How old is Bill?**
- A: Around 25.
- R: 3 + 4 + 18

- **Q: How long will take Bill to fly to Duke?**
- A: A few (1-5) hours.
- R: Duke is always in NC, Bill is now in CA

- **Q: How often would he visit Duke in the future?**
- A: A few (<5) times a year.

- Q: **Which one happened first, went or joined?**
- A: Went.

MC-TACO🌮: A Temporal Common Sense Dataset

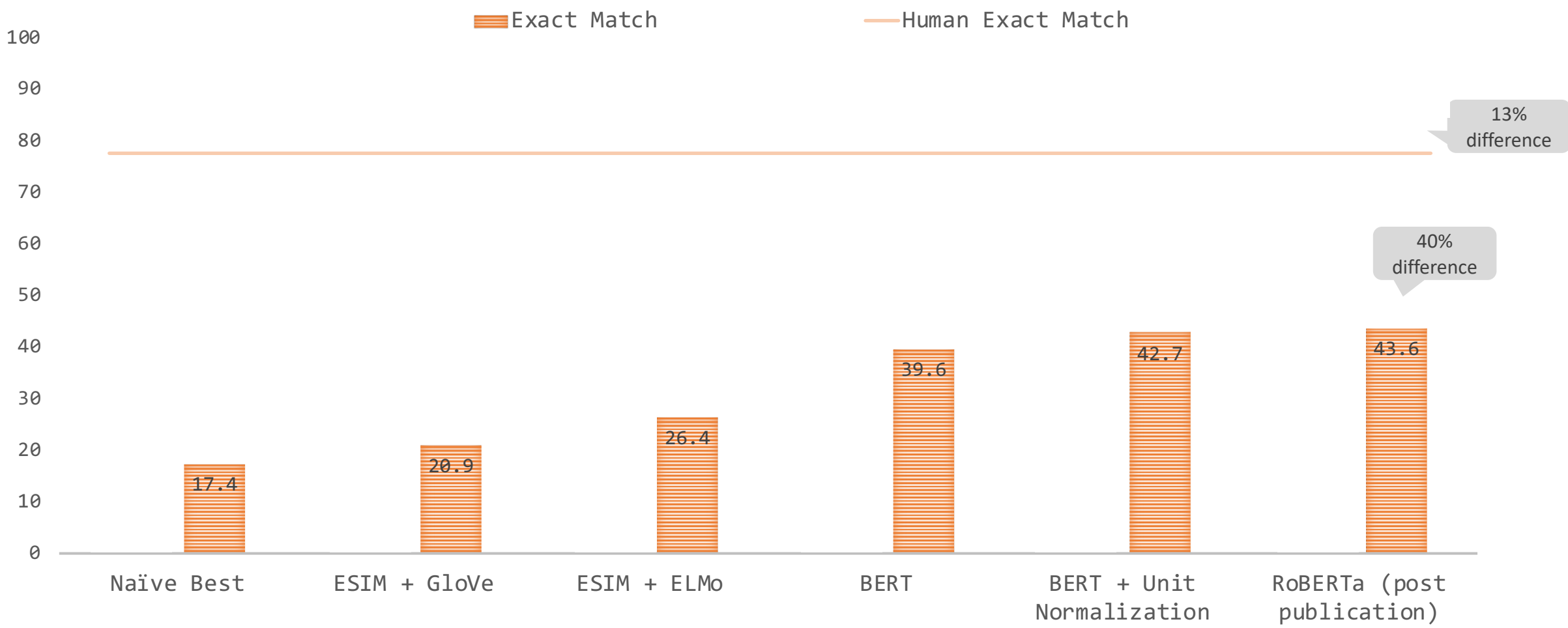
- MC-TACO 🌮 (**m**ultiple **c**hoice **t**empor**a**l **co**mmon-sense) **:**
  - In a given scenario – addressing multiple aspects/options of temporal commons sense
  - Input:

			Gold	Prediction	
He went to Duke University.	How long did it take him to graduate?	4 years			✓
He went to Duke University.	How long did it take him to graduate?	10 days			✓
		3.5 years			X
		16 hours			✓
		1 century			✓

  - Task: Decide whether each answer is plausible.
  - Metrics:
    - Exact Match: the percentage of question of which **all** candidates
    - F1: The F1 score of "plausible"
  - Statistics:
    - 1,893 questions
    - 13,225 question-answer pairs

F1: 66.7
Exact Match: 0.0

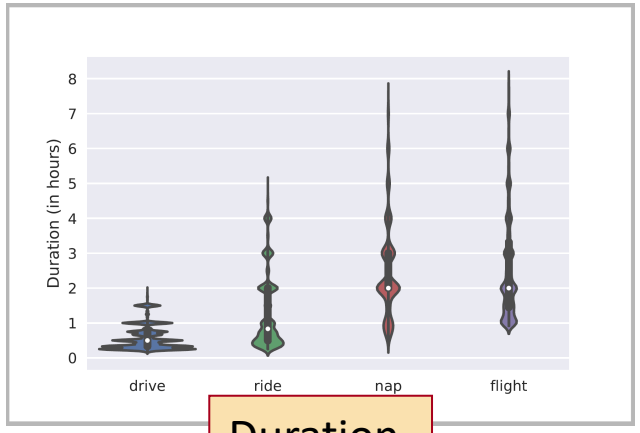
Results: We are Far



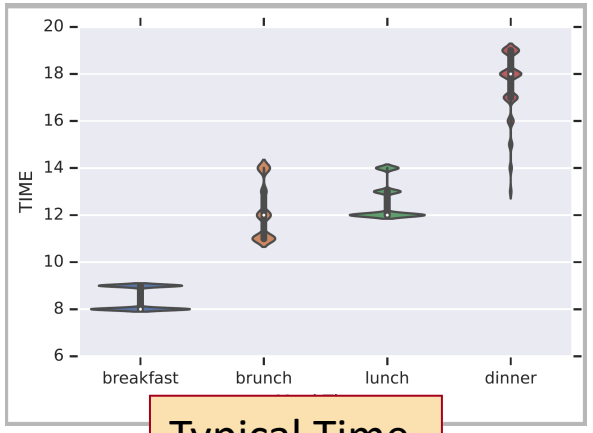
ESIM: Enhanced LSTM for Natural Language Inference (Chen et al., 2016)
GloVe: Global Vectors for Word Representation (Pennington et al., 2014)
ELMo: Deep contextualized word representations (Peters et al., 2018)
BERT: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2019)
RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)

What Do We Know?

- We can estimate some temporal aspects well



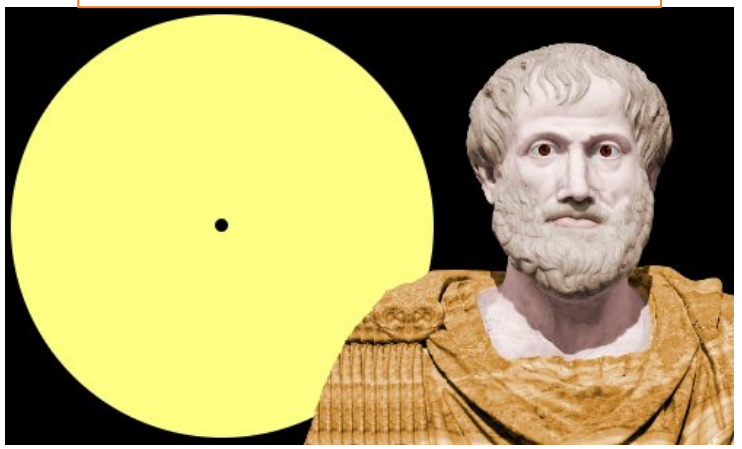
Duration



Typical Time

[ACL'19 + In Progress]

Did Aristotle have a laptop?



- But we don't know if Aristotle had a laptop
  - Not even if "we can make it to dinner before the movie".
- **Decompose or not decompose?**
  - **The strategy:** What do we need to know in order to answer
  - **What functions** are to be computed over components?
- How/What to **learn**?

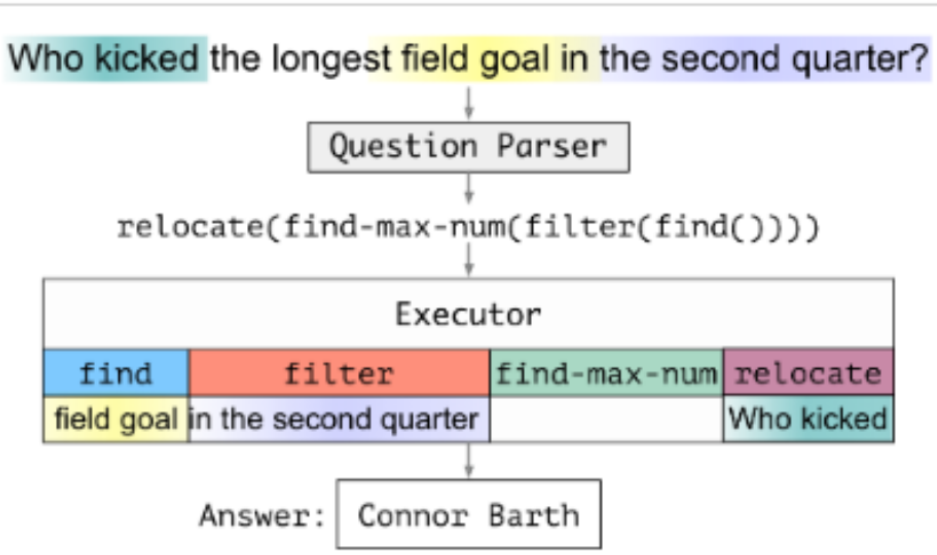
Reasoning over paragraphs

In the first quarter, Buffalo trailed as Chiefs QB Tyler Thigpen completed a 36-yard TD pass to RB Jamaal Charles. The Bills responded with RB Marshawn Lynch getting a 1-yard touchdown run. In the second quarter, Buffalo took the lead as kicker Rian Lindell made a 21-yard and a 40-yard field goal. Kansas City answered with Thigpen completing a 2-yard TD pass. Buffalo regained the lead as Lindell got a 39-yard field goal. The Chiefs struck with kicker Connor Barth getting a 45-yard field goal, yet the Bills continued their offensive explosion as Lindell got a 34-yard field goal, along with QB Edwards getting a 15-yard TD run. In the third quarter, Buffalo continued its poundings with Edwards getting a 5-yard TD run, while Lindell got himself a 48-yard field goal. Kansas City tried to rally as Thigpen completed a 45-yard TD pass to WR Mark Bradley, yet the Bills replied with Edwards completing an 8-yard TD pass to WR Josh Reed. In the fourth quarter, Edwards completed a 17-yard TD pass to TE Derek Schouman.

Who kicked the longest field goal in the second quarter?

Decomposition for Reasoning about Text[In Submission]

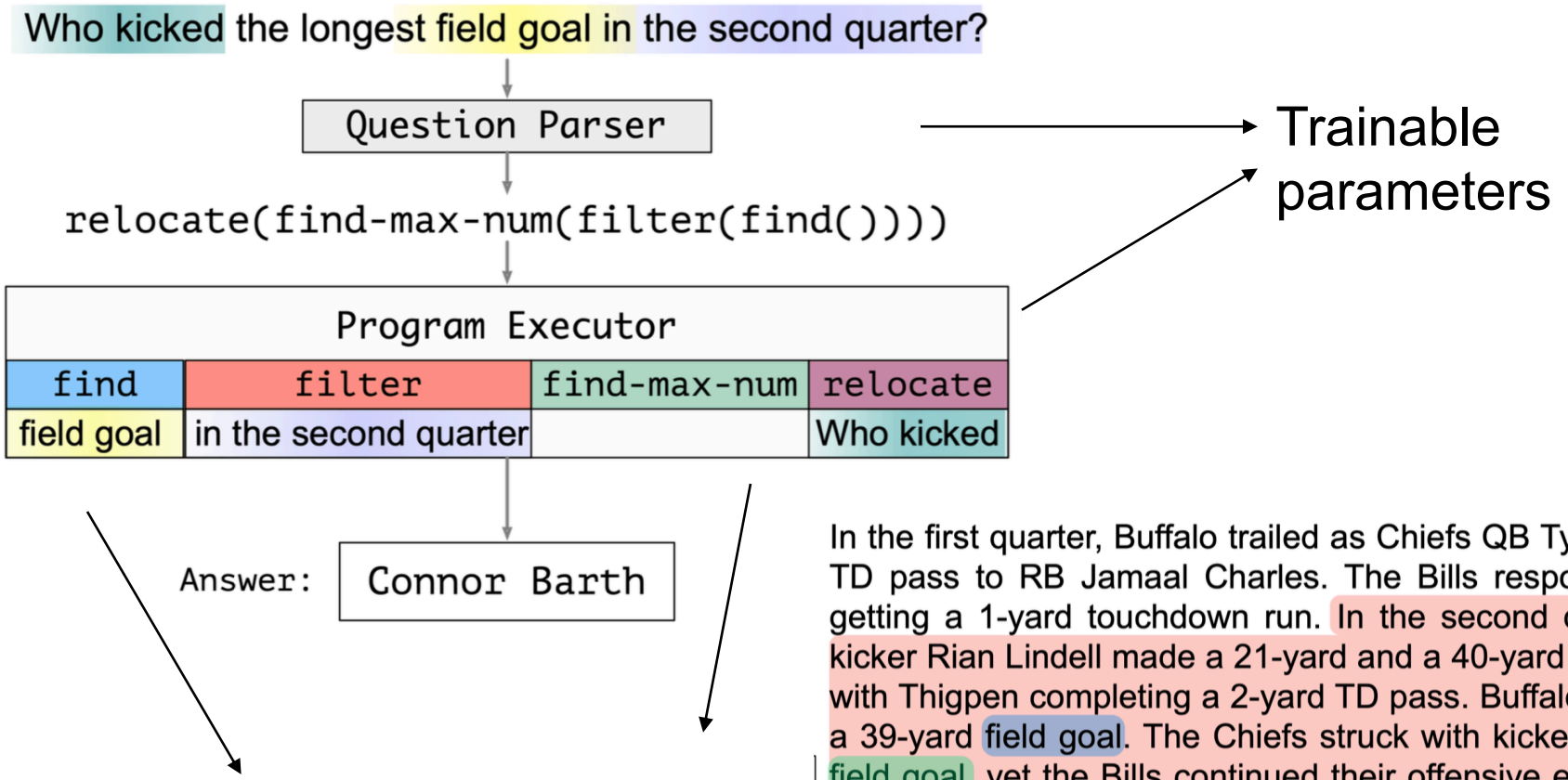
- We need to induce a program, with some executable modules at the leaves
  - "This phrase" indicates a field goal
  - "This player" scored it; "this is the team" that scored it
  - "This is the length" of the field goal
- Reason symbolically over it.
  - Imagine more expressive functions



Colors represent levels of attention over phrases

In the first quarter, Buffalo trailed as Chiefs QB Tyler Thigpen completed a 36-yard TD pass to RB Jamaal Charles. The Bills responded with RB Marshawn Lynch getting a 1-yard touchdown run. In the second quarter, Buffalo took the lead as kicker Rian Lindell made a 21-yard and a 40-yard field goal. Kansas City answered with Thigpen completing a 2-yard TD pass. Buffalo regained the lead as Lindell got a 39-yard field goal. The Chiefs struck with kicker Connor Barth getting a 45-yard field goal, yet the Bills continued their offensive explosion as Lindell got a 34-yard field goal, along with QB Edwards getting a 15-yard TD run. In the third quarter, Buffalo continued its poundings with Edwards getting a 5-yard TD run, while Lindell got himself a 48-yard field goal. Kansas City tried to rally as Thigpen completed a 45-yard TD pass to WR Mark Bradley, yet the Bills replied with Edwards completing an 8-yard TD pass to WR Josh Reed. In the fourth quarter, Edwards completed a 17-yard TD pass to TE Derek Schouman.

Neural Module Networks for Text



In the first quarter, Buffalo trailed as Chiefs QB Tyler Thigpen completed a 36-yard TD pass to RB Jamaal Charles. The Bills responded with RB Marshawn Lynch getting a 1-yard touchdown run. In the second quarter, Buffalo took the lead as kicker Rian Lindell made a 21-yard and a 40-yard field goal. Kansas City answered with Thigpen completing a 2-yard TD pass. Buffalo regained the lead as Lindell got a 39-yard field goal. The Chiefs struck with kicker Connor Barth getting a 45-yard field goal, yet the Bills continued their offensive explosion as Lindell got a 34-yard field goal, along with QB Edwards getting a 15-yard TD run. In the third quarter, Buffalo continued its poundings with Edwards getting a 5-yard TD run, while Lindell got himself a 48-yard field goal. Kansas City tried to rally as Thigpen completed a 45-yard TD pass to WR Mark Bradley, yet the Bills replied with Edwards completing an 8-yard TD pass to WR Josh Reed. In the fourth quarter, Edwards completed a 17-yard TD pass to TE Derek Schouman.

Challenges

Training:

- End-to-End Learning in this context is challenging
- You have to know something
  - A limited amount of heuristically-obtained modules provides sufficient inductive bias for accurate learning

Reasoning

- We introduce modules that perform symbolic reasoning (such as arithmetic, sorting, counting) over numbers in a probabilistic and differentiable manner.
- But how to extend it to more functions?

At this point, we can get state-of-the art and interpretable results on a relevant fraction of the DROP dataset.

Recap

Modified version of a question for AI2's DROP dataset

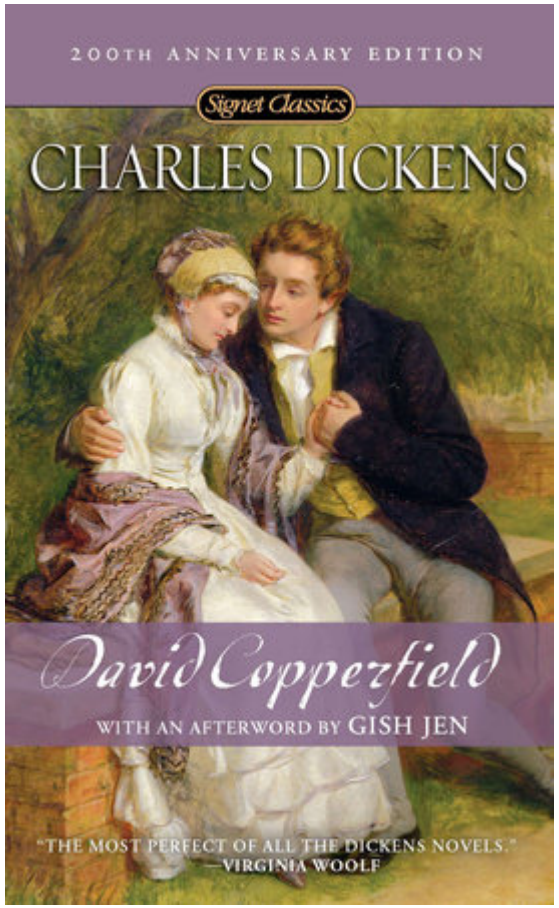
Coming off a road win over the Cowboys, the Redskins traveled to Lincoln Financial Field for a Week 5 NFC East duel with the Philadelphia Eagles. In the first quarter, the Redskins trailed early **as RB Brian Westbrook scored on a 9-yard TD run** and **the Eagles DeSean Jackson returned a punt 68 yards for a touchdown.** Washington still trailed at half time 14:9, **with field goals from Shaun Suisham**. In the third quarter, the Redskins took the lead on a trick play as WR Antwaan Randle El threw an 18-yard **TD pass to TE Chris Cooley**. In the fourth quarter, the Redskins increased their lead when **Clinton Portis scored on a 4-yard TD run**. The Eagles managed one more score in the final quarter for a **final score of 17:23**.

- What are the computational tasks that we should think about?

CliffsNotes

- Multiple natural language documents
  - Small units of text or large units of texts
  - Reading news about an event/situation **over time** and/or from **multiple sources**
  - **Reading a book**

  - The novel features the character David Copperfield, his journey of change and growth from infancy to maturity, as many people enter and leave his life and he passes through the stages of his development. (**Fiction, and you know it**)
  - London and England in the 19-th century; socio-economic state, child exploitation; schools, prisons, emigration to Australia (**true historical facts**)
- What are the computational tasks that we should think about?



Conclusion

- What is Reasoning?
- Who is doing the Reasoning?