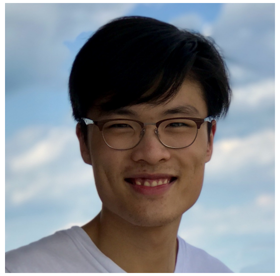
Ai2
Allen Institute for AI

Penn

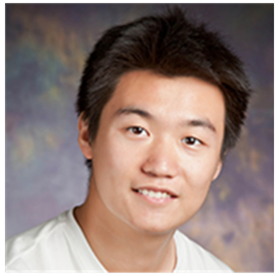
Cognitive Computation Group

Temporal Reasoning on Implicit Events from Distant Supervision



Ben Zhou

Kyle Richardson



Qiang Ning

Tushar Khot

Ashish Sabharwal

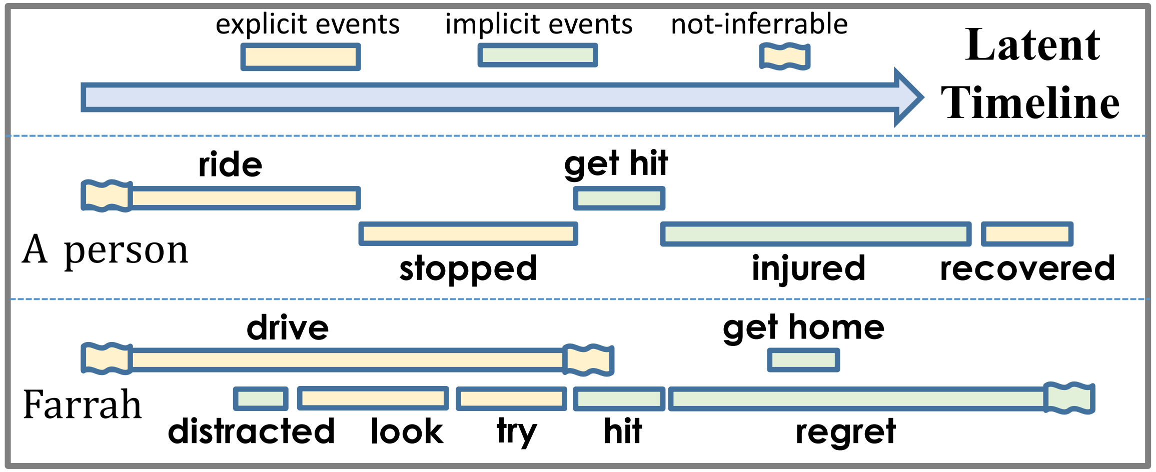
Dan Roth

Motivation

- Humans can construct latent timelines

Context Story

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

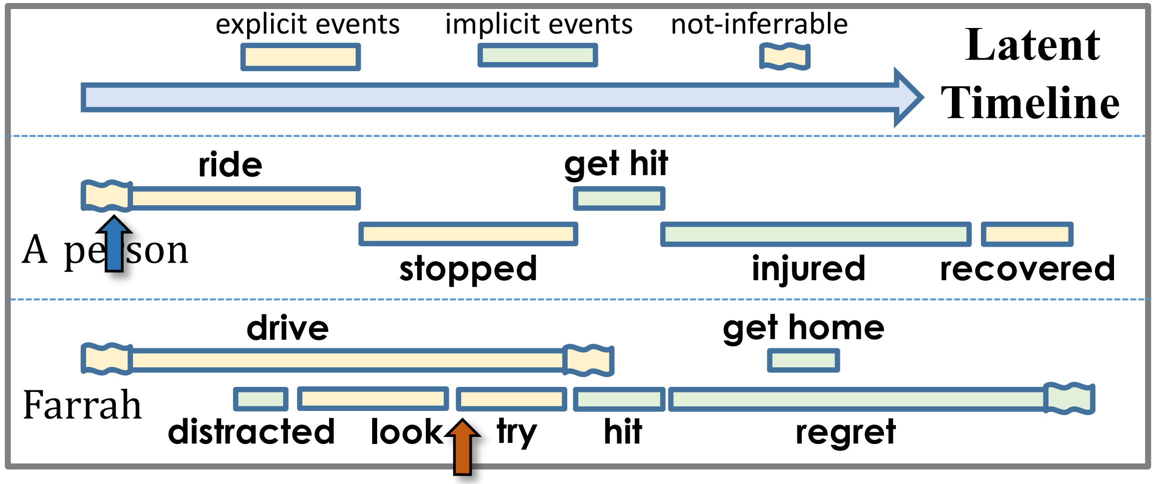


Motivation

- Humans can construct latent timelines
- On explicitly mentioned events
- Ride a bike <u>started before</u> Farrah brakes
- Ride a bike <u>ended before</u> Farrah brakes

Context Story

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

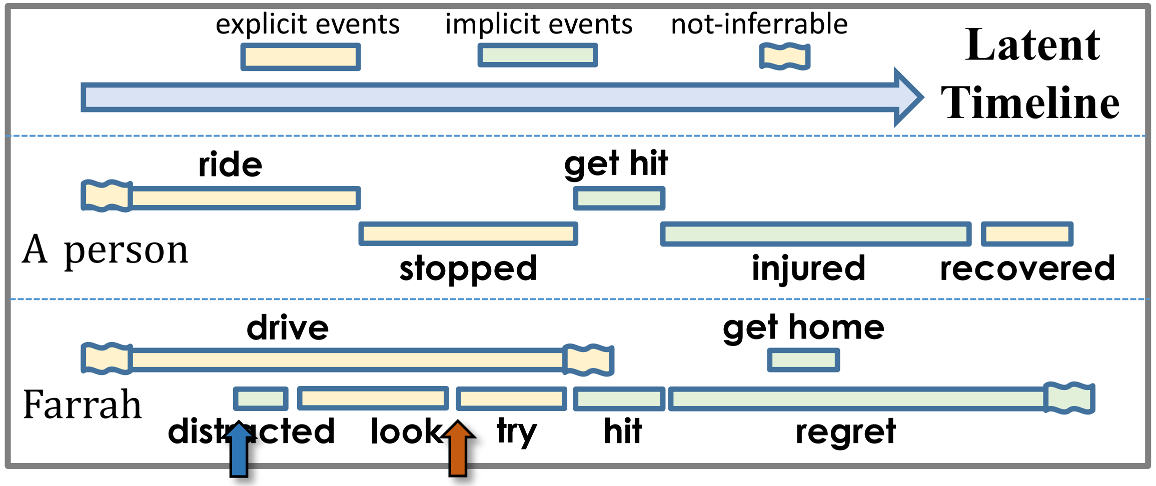


Motivation

- Humans can construct latent timelines
- Also on implicit events
- Farrah was distracted
  - □ Started before Farrah tries to brake

Context Story

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

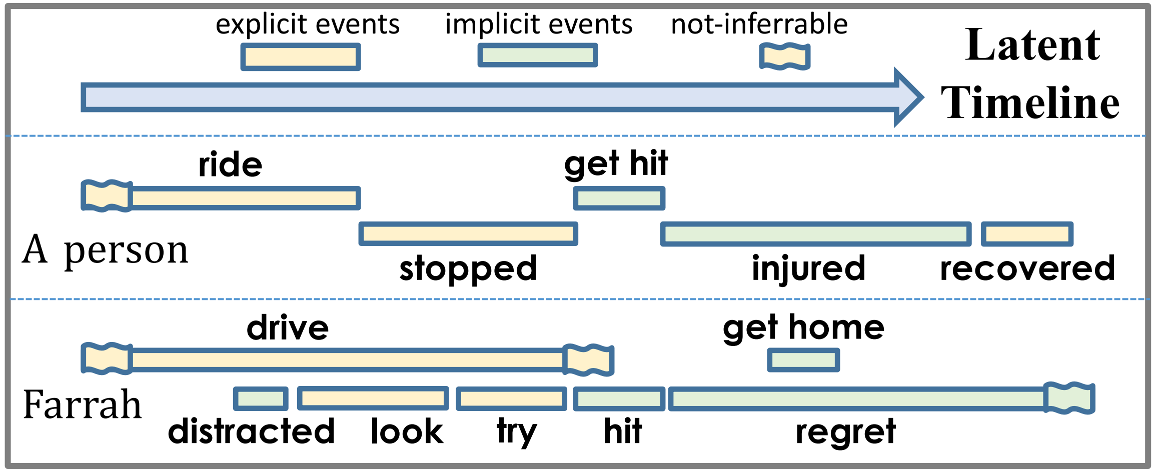


Motivation

- Humans can construct latent timelines
- On both explicit and implicit events
- Can fit any “unmentioned” events into the timeline
- *“Farrah’s phone rang while driving”*
- *“The person went to the hospital”*
- Such ability is not tested by existing temporal benchmarks

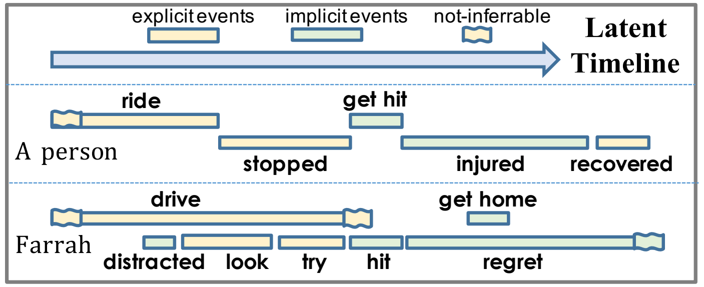
Context Story

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.



In this work…

- TRACIE (TempoRAl Closure InfErence)
  - □ A temporal relation benchmark with implicit events
  - □ Test both start time and end time
  - □ 5.5K entailment instances
  - □ RoBERTa-Large (cite): 71% binary accuracy
- Better models for implicit events and time
  - □ PatternTime
    - Trained on distant supervision collected automatically from textual patterns
  - □ SymTime
    - A neural-symbolic reasoning model on top of PatternTime
    - Symbolize interval-based algebraic operations
    - Decompose end time to start time and duration prediction



Tracie Instance

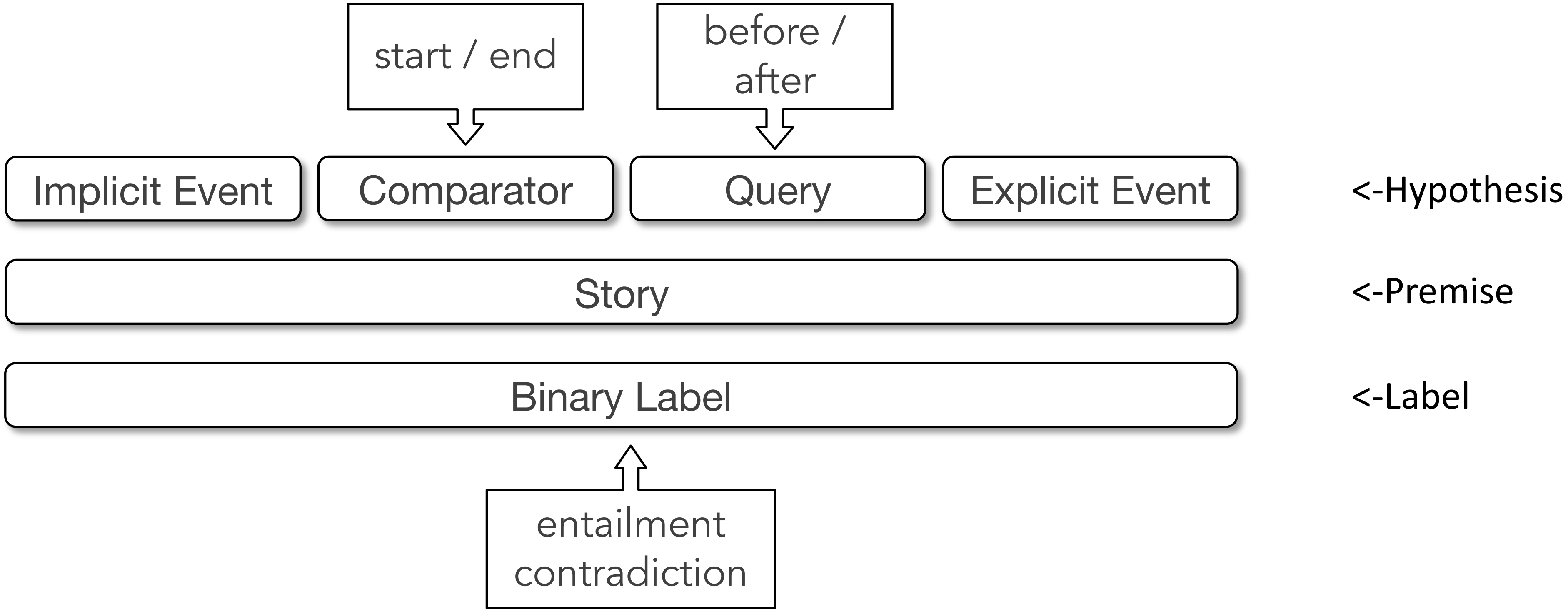
distracted starts *before* try starts ✅ entailment

distracted ends *after* try starts ❌ contradiction

…. many others

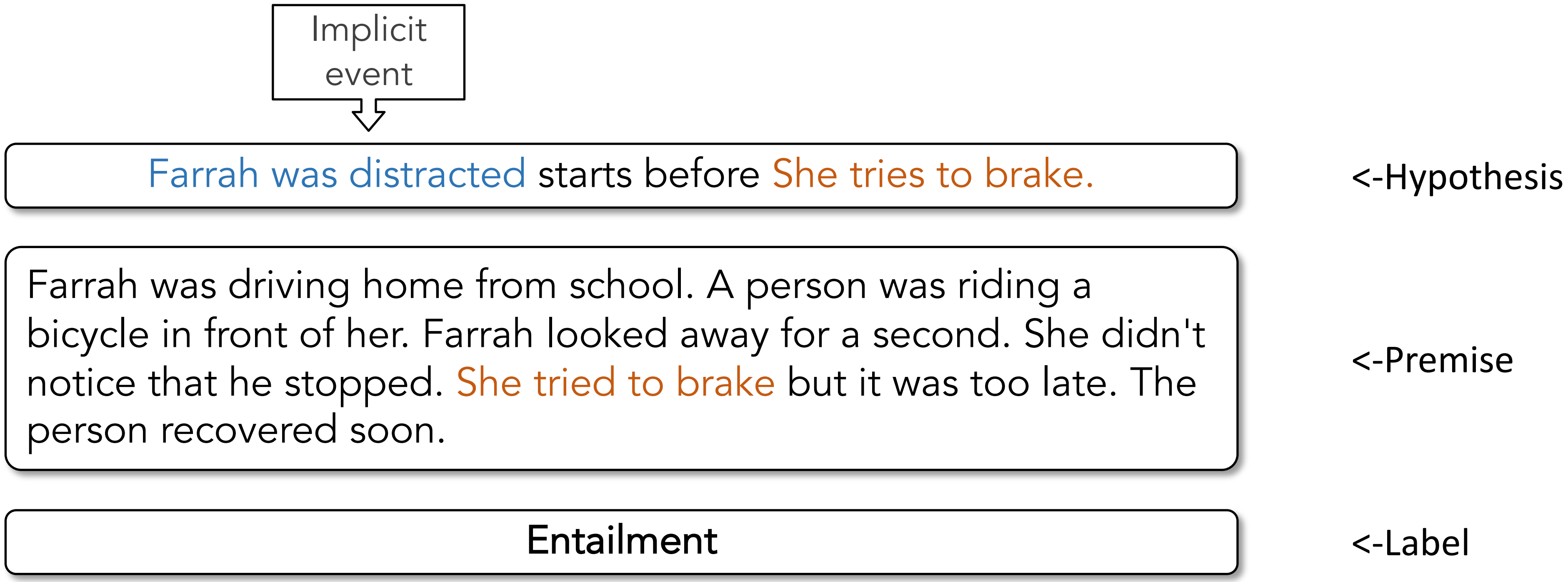
TRACIE: format

- A temporal benchmark on implicit events



TRACIE: format

- A temporal benchmark on implicit events



Implicit event

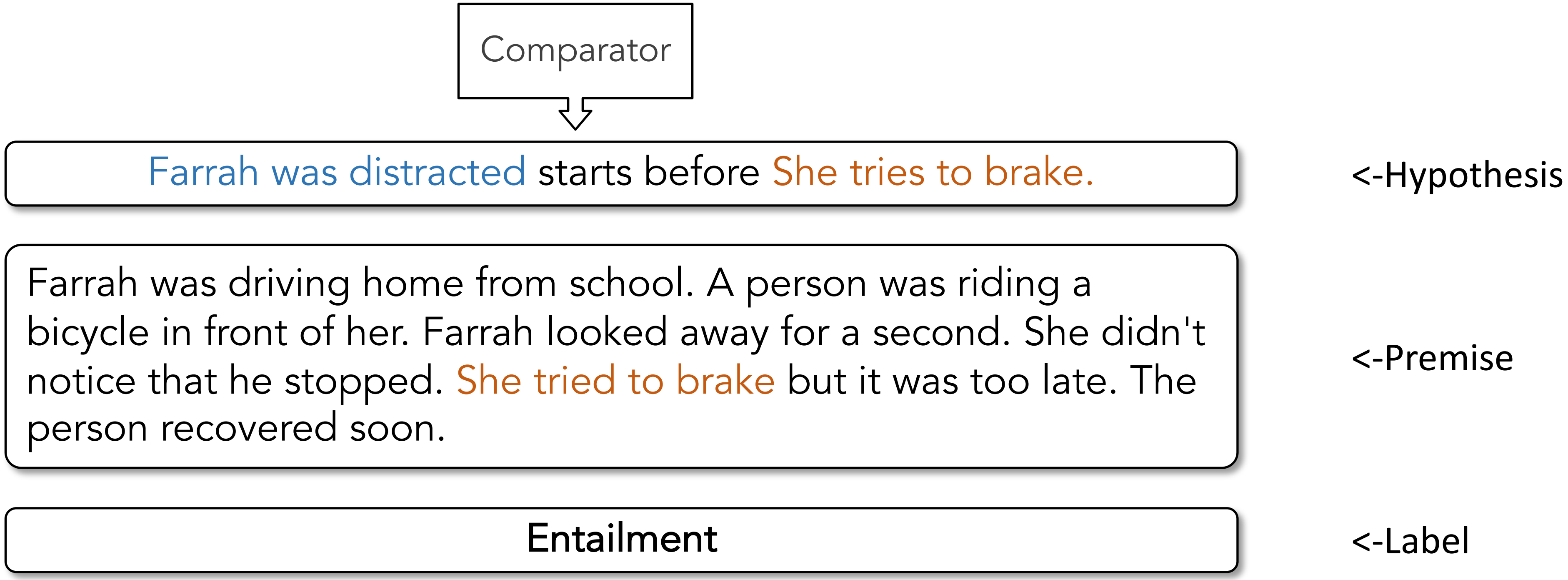
Farrah was distracted starts before She tries to brake. <-Hypothesis

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon. <-Premise

Entailment <-Label

TRACIE: format

- A temporal benchmark on implicit events



Comparator

Farrah was distracted starts before She tries to brake.

<-Hypothesis

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

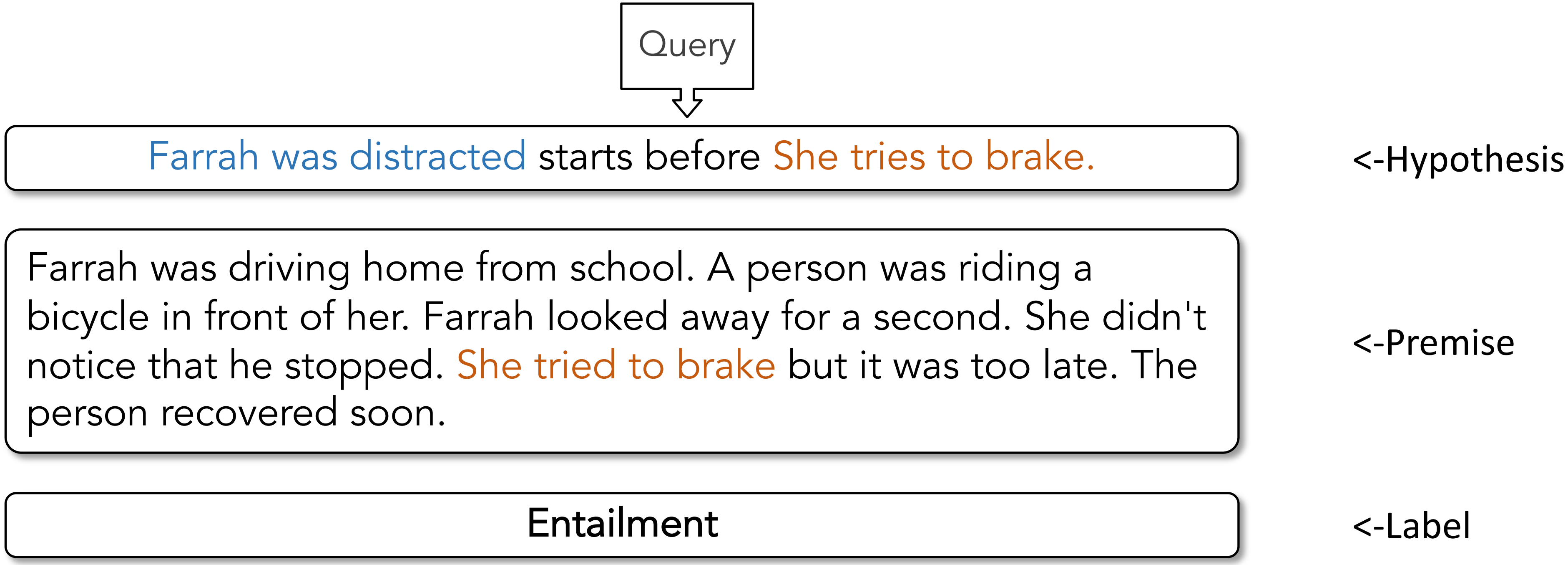
<-Premise

Entailment

<-Label

TRACIE: format

- A temporal benchmark on implicit events



Query

Farrah was distracted starts before She tries to brake.

<-Hypothesis

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

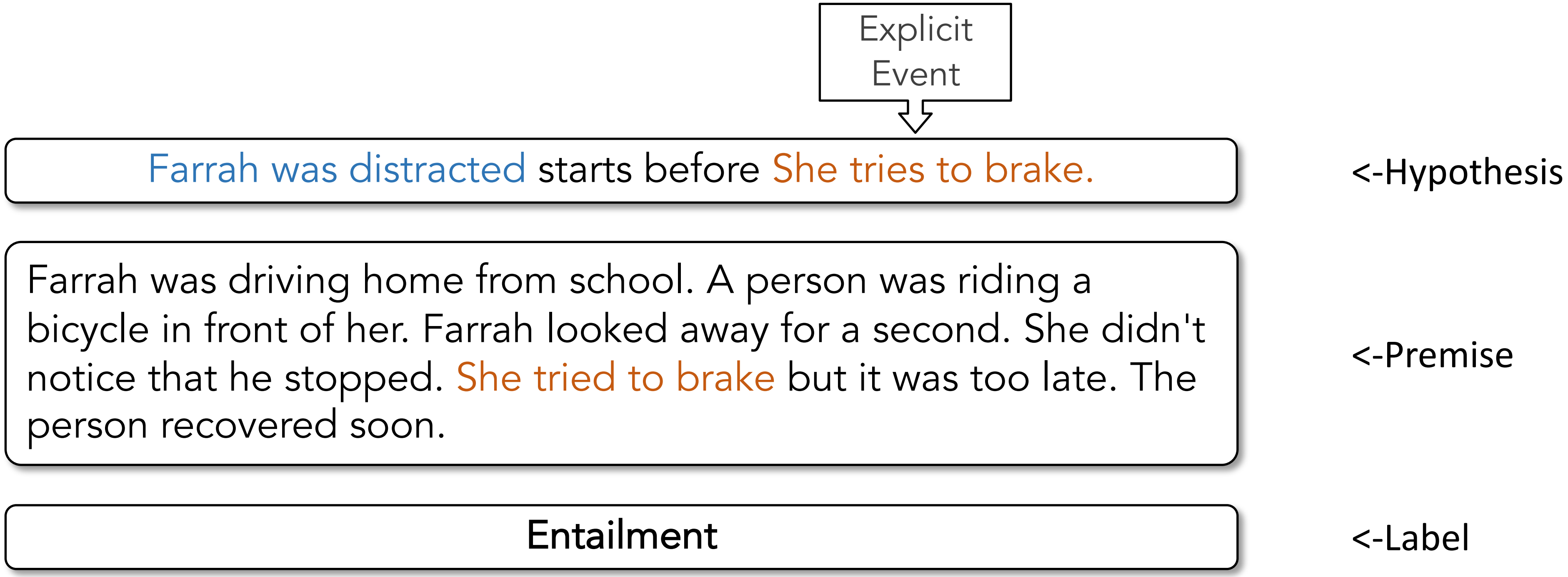
<-Premise

Entailment

<-Label

TRACIE: format

- A temporal benchmark on implicit events



Explicit Event

Farrah was distracted starts before She tries to brake. <-Hypothesis

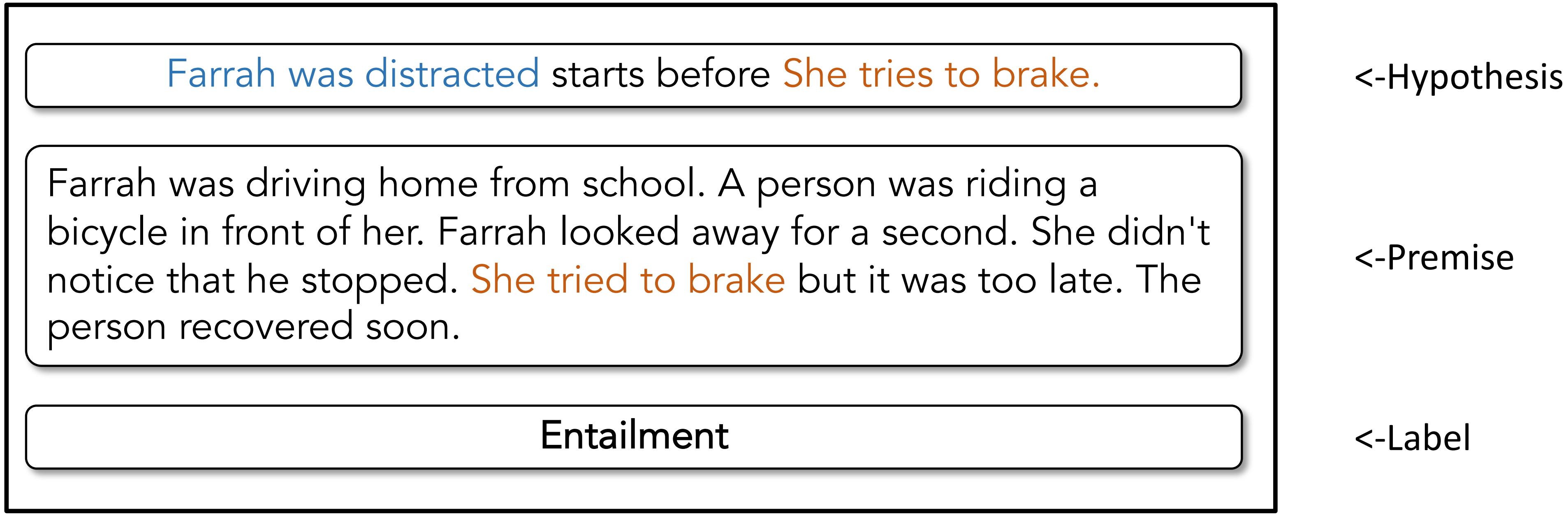
Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon. <-Premise

Entailment <-Label

TRACIE: format

- A temporal benchmark on implicit events

A TRACIE instance



Farrah was distracted starts before She tries to brake.

<-Hypothesis

Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

<-Premise

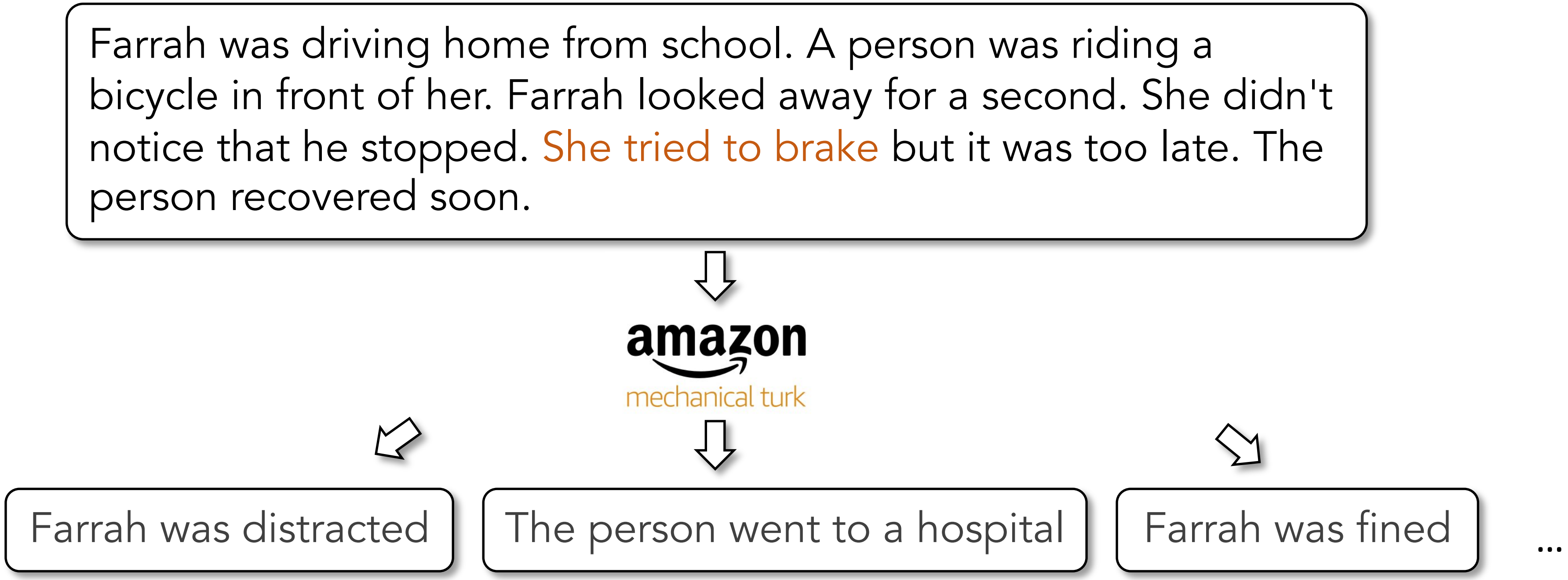
Entailment

<-Label

TRACIE: annotation

Stage 1: collect implicit events

- Sample context stories from ROCStories (cite)
- Annotators write implicit events in their own words



Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

Farrah was distracted

The person went to a hospital

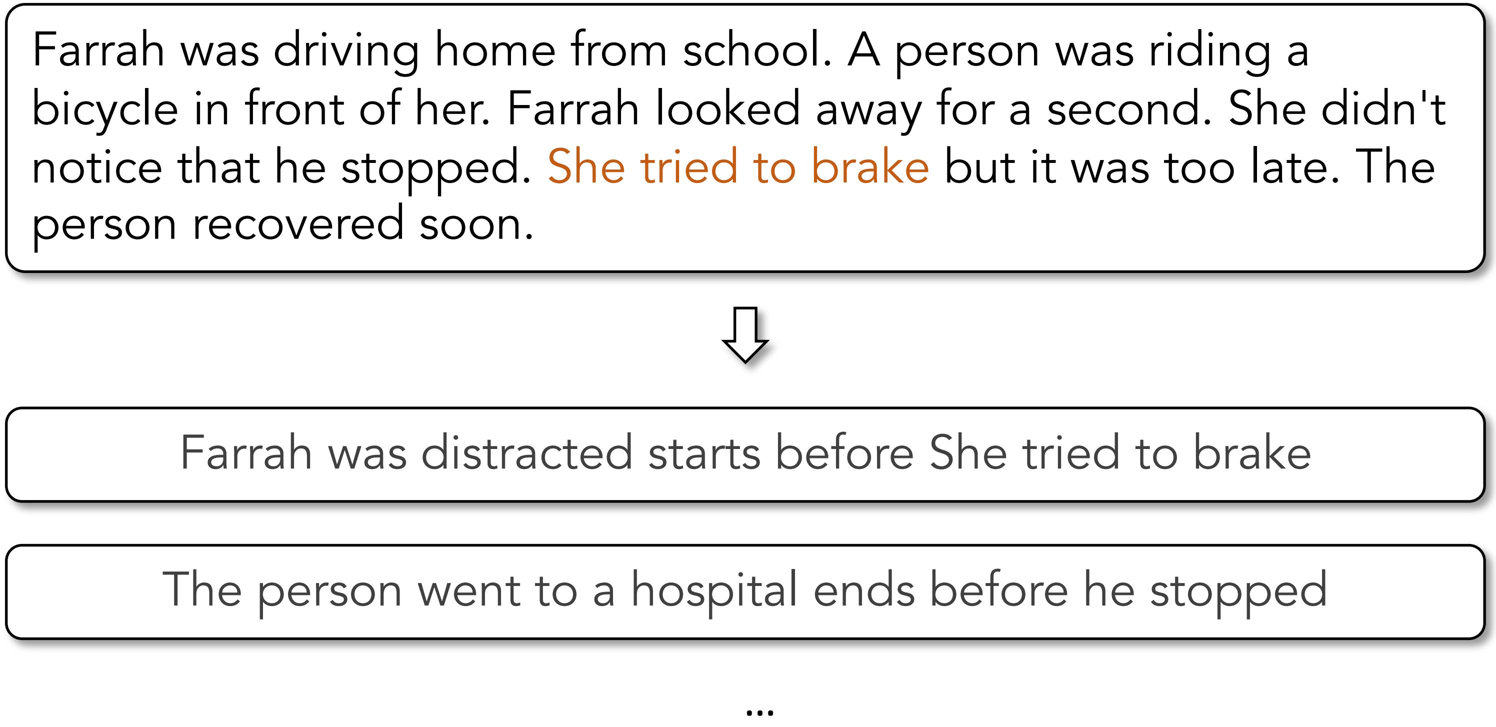
Farrah was fined

...

TRACIE: annotation

Stage 2: Generate (unlabaled) TRAICE instances

- Collect a pool of explicit events
  - □ Composed by both annotators’ rewriting and SRL extractions
- Randomly pair with explicit events and comparator/query



Farrah was driving home from school. A person was riding a bicycle in front of her. Farrah looked away for a second. She didn't notice that he stopped. She tried to brake but it was too late. The person recovered soon.

⇩

Farrah was distracted starts before She tried to brake

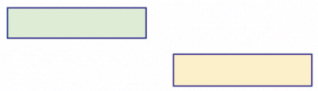
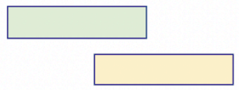
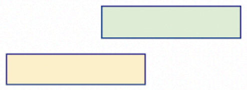
The person went to a hospital ends before he stopped

…

TRACIE: annotation

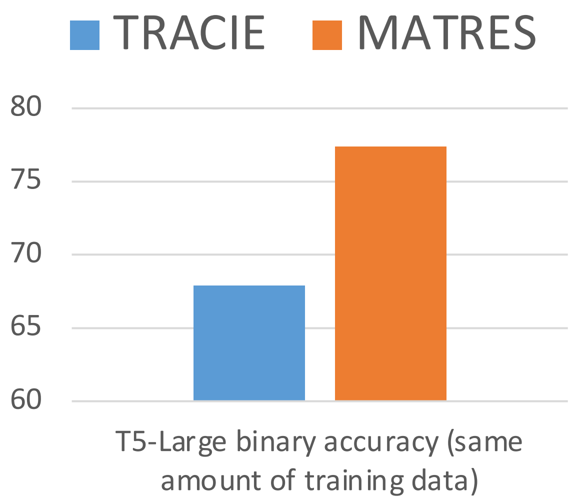
Stage 3: Annotate Binary Labels

- 4 annotators label each instance with a binary True/False label
- Label definition: compare an implicit event with an <u>explicit event's start time</u>
  - □ Improves annotator agreement
  - □ Makes the implicit event more groundable

Illustration	Allen's Relation	Tracie's Relation
	Precedes	Starts Before Ends Before
	Overlaps, Finished-by, Contains	Starts Before Ends After
	During, Finishes, Overlapped-by, Met-by, Preceded-by	Starts After Ends After

TRACIE: the dataset

- 5.5k instances
- 20%/80% train/test split
  - □ As a commonsense task, we should not ask a model to solely learn from in-domain supervision
- Uniform-prior split
  - □ Removes all prior knowledge regarding comparator-query-label distributions in training data
  - □ 51% binary accuracy for Bi-LSTM
  - □ ~70% binary accuracy for all existing pre-trained LMs
    - RoBERTa-large, T5-large, T5-3B



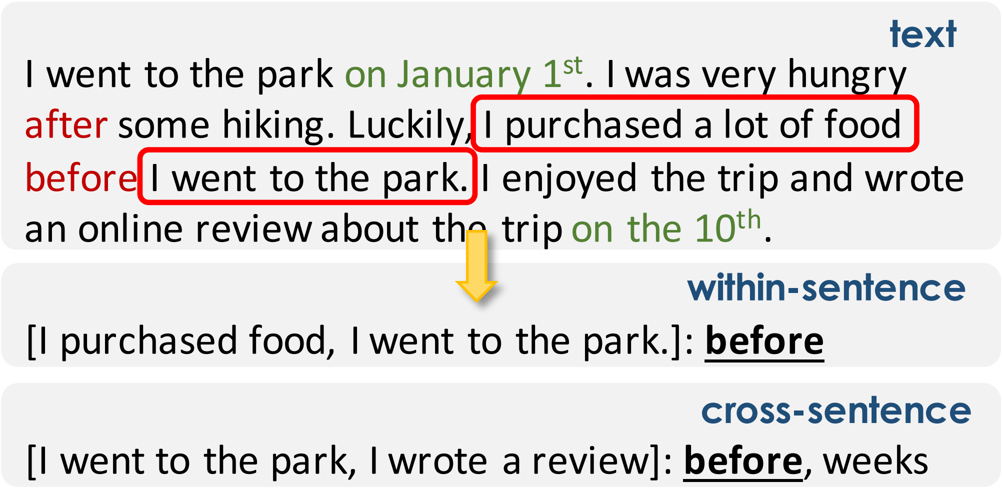
Our Models: overview

We propose two models:

- PatternTime ⇦
  - From distant supervision collected via textual patterns
- SymTime
  - Symbolic End-to-end Reasoning Model

PatternTime: Distant Supervision Collection

- We want to learn to compare start times
  - □ From unannotated free texts
- **Within-sentence extraction**
  - □ Not enough:
    - Does not address implicit events
    - Does not tell how far the two start times are



text

I went to the park on January 1st. I was very hungry after some hiking. Luckily, I purchased a lot of food before I went to the park. I enjoyed the trip and wrote an online review about the trip on the 10th.

within-sentence

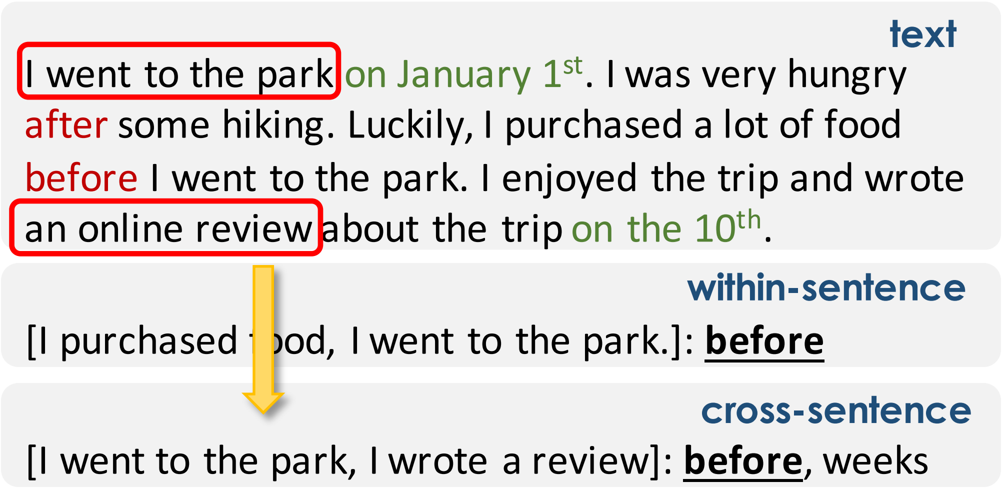
[I purchased food, I went to the park.]: **before**

cross-sentence

[I went to the park, I wrote a review]: **before**, weeks

PatternTime: Distant Supervision Collection

- We want to learn to compare start times
  - □ From unannotated free texts
- **Cross-sentence extraction**
  - □ Based on explicit temporal expressions
  - □ Independent of event locations
  - □ Produces relative distance between start times



text

I went to the park on January 1st. I was very hungry after some hiking. Luckily, I purchased a lot of food before I went to the park. I enjoyed the trip and wrote an online review about the trip on the 10th.

within-sentence

[I purchased food, I went to the park.]: **before**

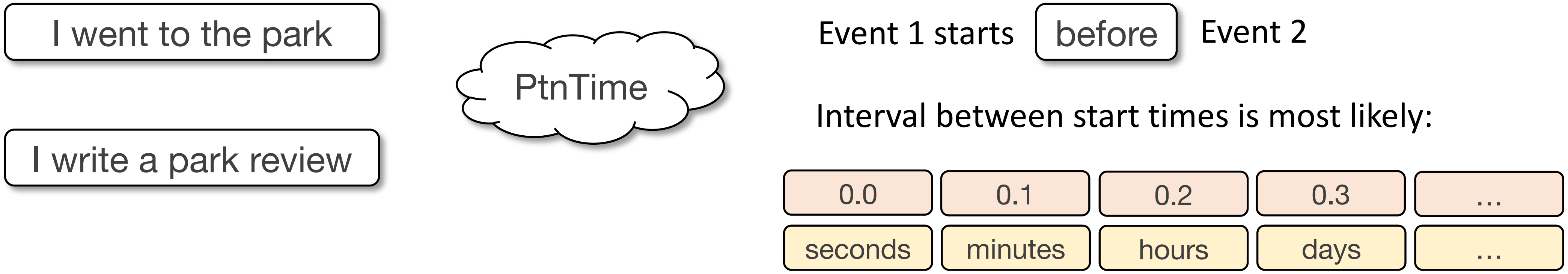
cross-sentence

[I went to the park, I wrote a review]: **before**, weeks

Learn with Distant Supervision

PatternTime

- A sequence-to-sequence model
  - □ Train on 1.5M distant supervision instances
- Input: two event phrases
- Output:
  - □ A binary label indicating which event starts earlier
  - □ Probabilities over duration units indicating the interval between two start times



Event 1 starts before Event 2

Interval between start times is most likely:

0.0	0.1	0.2	0.3	…
seconds	minutes	hours	days	…

Our Models: overview

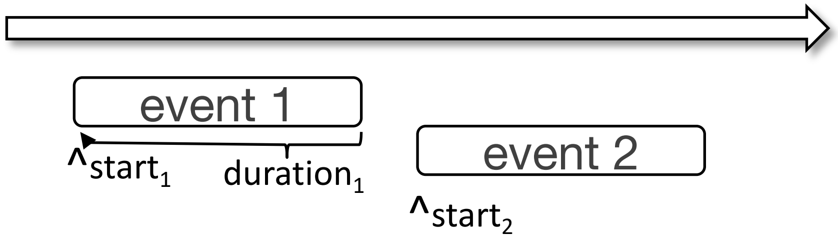
We propose two models:

- PatternTime
  - □ From distant supervision collected via textual patterns
- SymTime ⇦
  - □ Symbolic End-to-end Reasoning Model

Symbolic Reasoning Model

SymTime

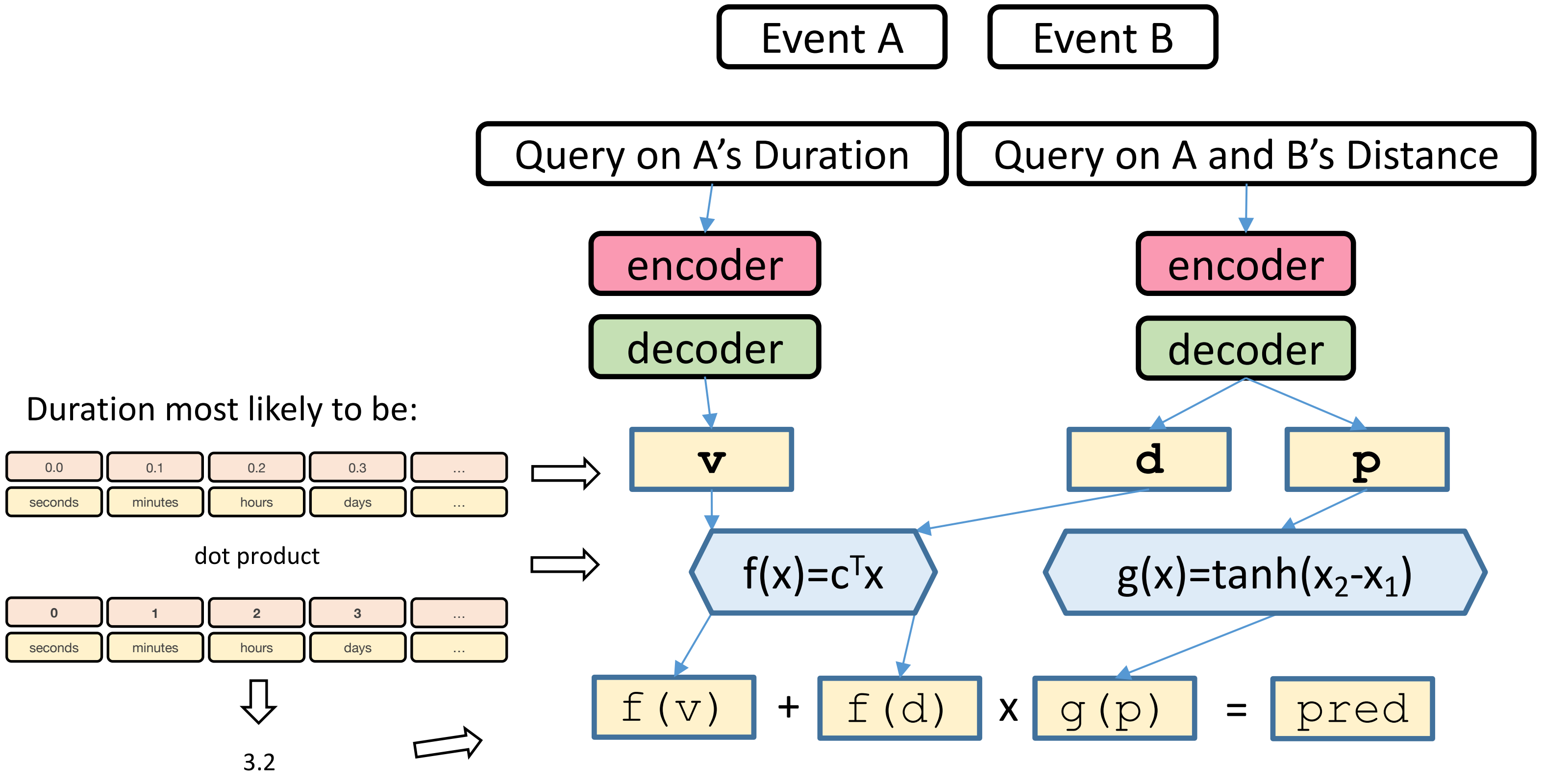
comparator l	**relation** $r_l(\mathbf{e}_1, \mathbf{e}_2)=$	
ends	**before**	if $\mathbf{end}_1 < \mathbf{start}_2$
	after	*otherwise*
starts	**before**	if $\mathbf{start}_1 < \mathbf{start}_2$
	after	*otherwise*



- Comparator=start: solvable with PatternTime
- Comparator=end:
  - □ start$_1$ + duration$_1$? start$_2$
  - □ $\underbrace{\text{duration}_1}$? $\underbrace{\text{start}_2 - \text{start}_1}_{\text{PatternTime}}$

Another model trained with distant supervision from a previous work (Zhou et al. 2020)

SymTime



Event A

Event B

Query on A’s Duration

Query on A and B’s Distance

encoder

decoder

encoder

decoder

v

d

p

$f(x)=c^Tx$

$g(x)=\tanh(x_2-x_1)$

f(v) + f(d) x g(p) = pred

Duration most likely to be:

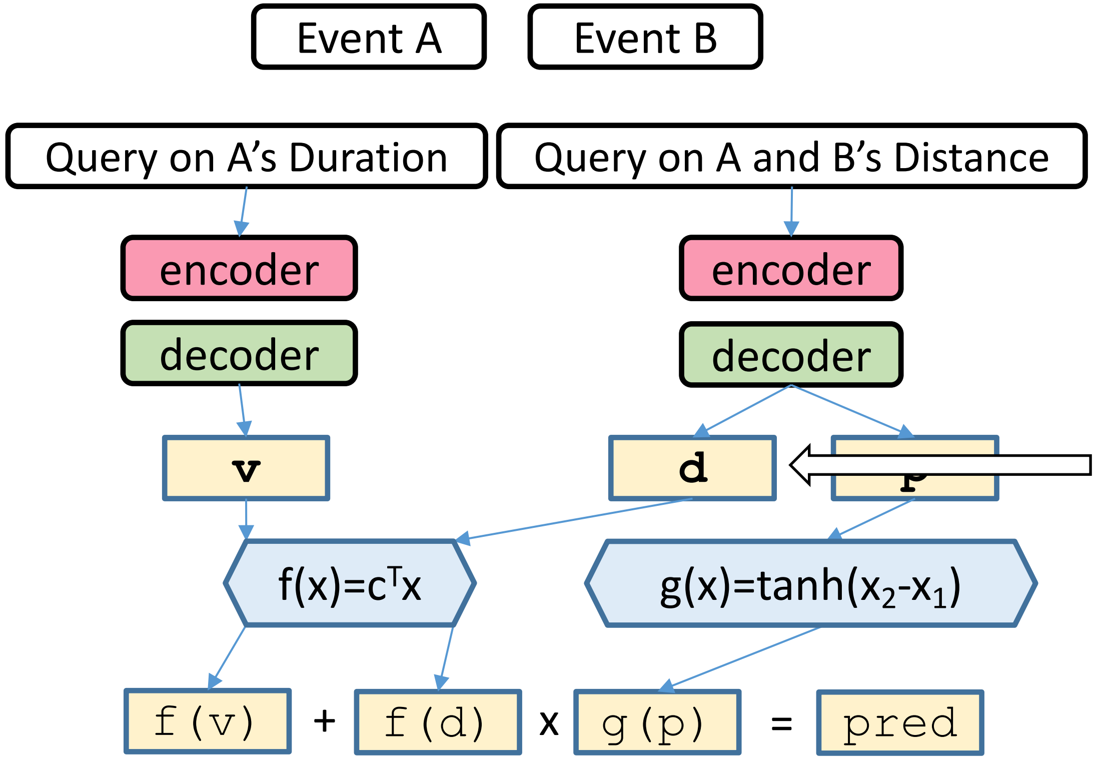
0.0	0.1	0.2	0.3	...
seconds	minutes	hours	days	...

dot product

0	1	2	3	...
seconds	minutes	hours	days	...

3.2

SymTime



Start time interval most likely to be:

0.0	0.1	0.2	0.3	...
seconds	minutes	hours	days	...

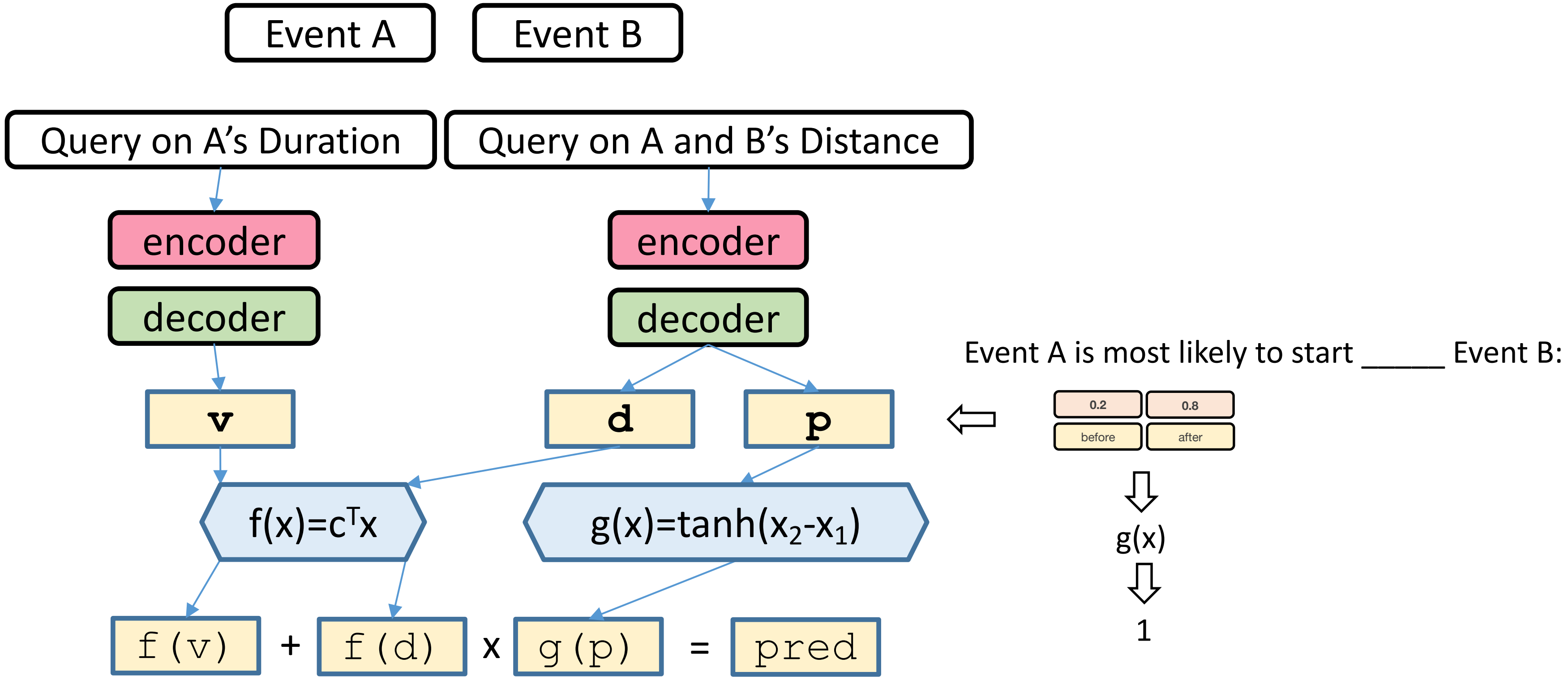
dot product

0	1	2	3	...
seconds	minutes	hours	days	...

⇩

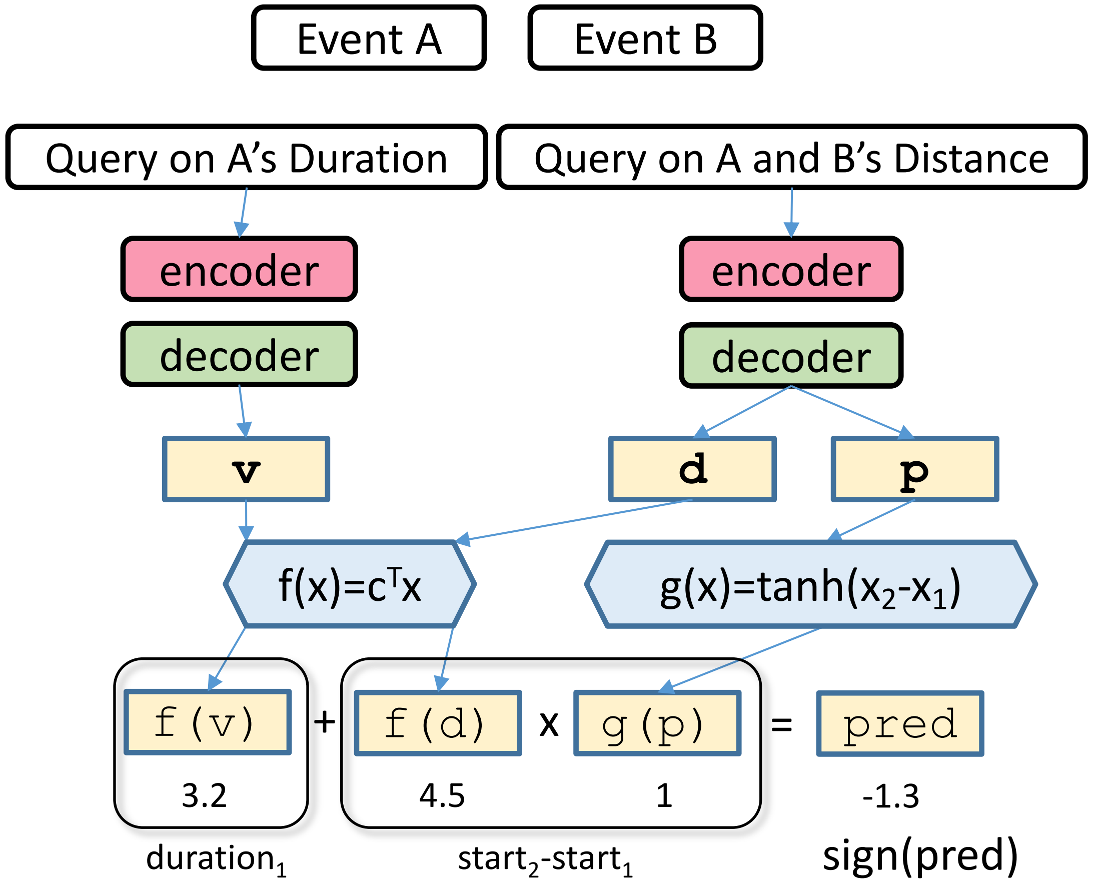
4.5

SymTime



Event A
Event B
Query on A's Duration
Query on A and B's Distance
encoder
decoder
encoder
decoder
v
d
p
f(x)=c^{T}x
g(x)=tanh(x$_2$-x$_1$)
f(v)
+
f(d)
x
g(p)
=
pred
Event A is most likely to start _____ Event B:
0.2
0.8
before
after
g(x)
1

SymTime


Event A
Event B
Query on A's Duration
Query on A and B's Distance
encoder
encoder
decoder
decoder
v
d
p
f(x)=c^{T}x
g(x)=tanh(x$_2$-x$_1$)
f(v)
+
f(d)
x
g(p)
=
pred
3.2
4.5
1
-1.3
duration$_1$
start$_2$-start$_1$
sign(pred)

Experiments

- On uniform-prior training data

🟦 T5-Large 🟧 T5-Matres ⬜ PatternTime 🟨 SymTime 🟦 T5-3B

⇧

Our baseline LM
Main Comparison

Experiments: TRACIE

- On uniform-prior training data

🟦 T5-Large 🟧 T5-Matres ⬜ PatternTime 🟨 SymTime 🟦 T5-3B

⇧

finetuned on MATRES

Experiments: TRACIE

- On uniform-prior training data

🟦 T5-Large 🟧 T5-Matres ◼ PatternTime 🟨 SymTime 🟦 T5-3B

Our proposed models

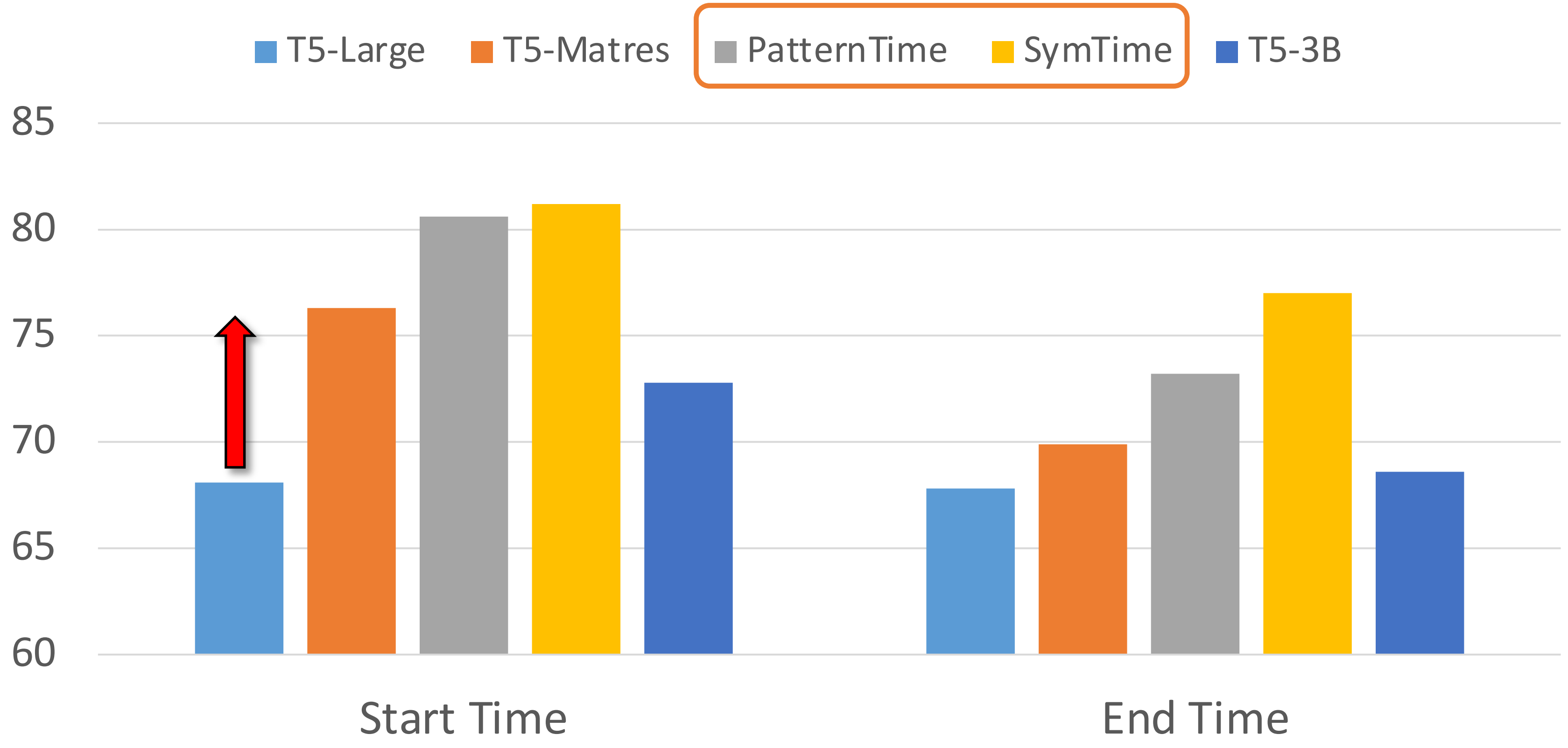
Experiments: TRACIE

- On uniform-prior training data

A Larger T5

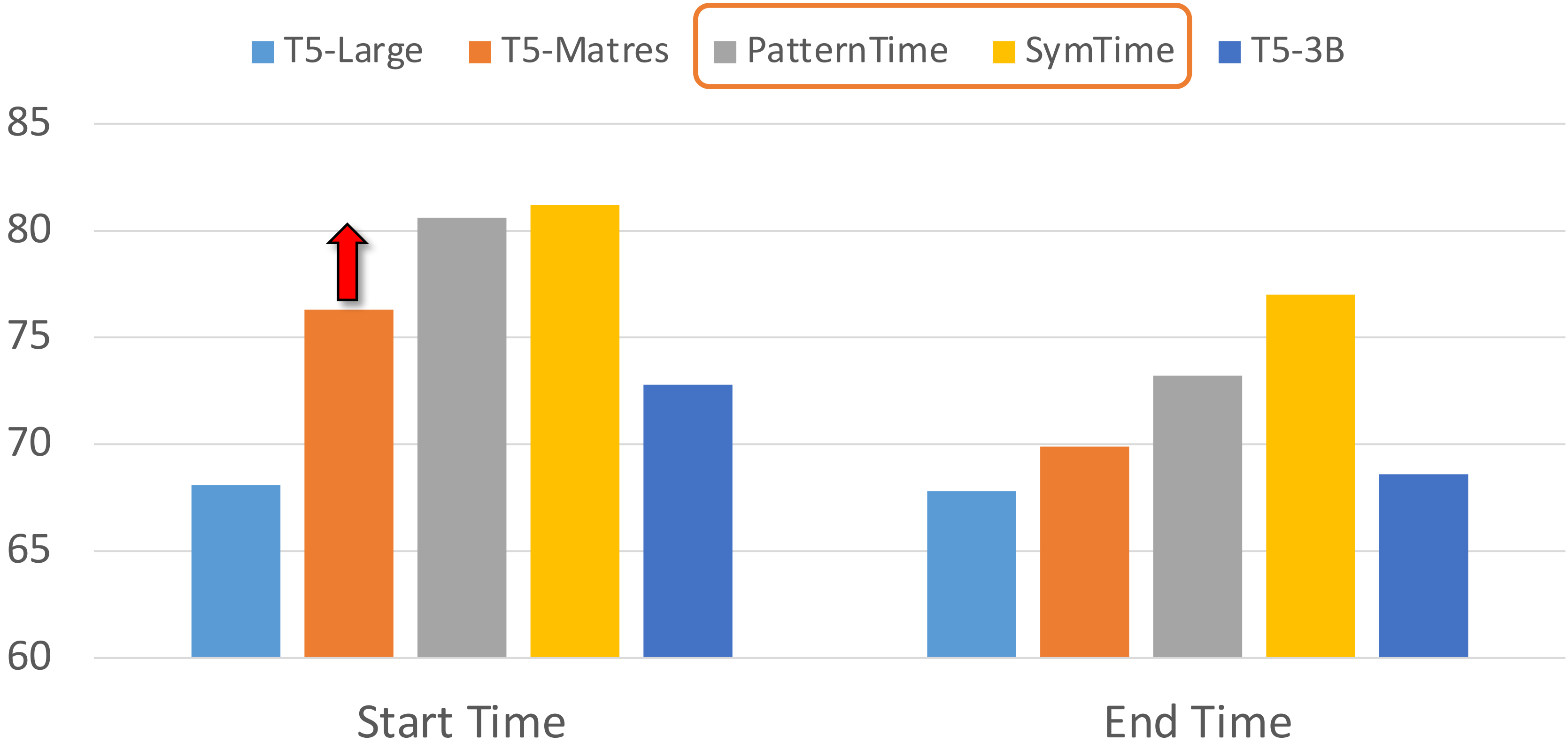
Experiments: TRACIE

- On uniform-prior training data



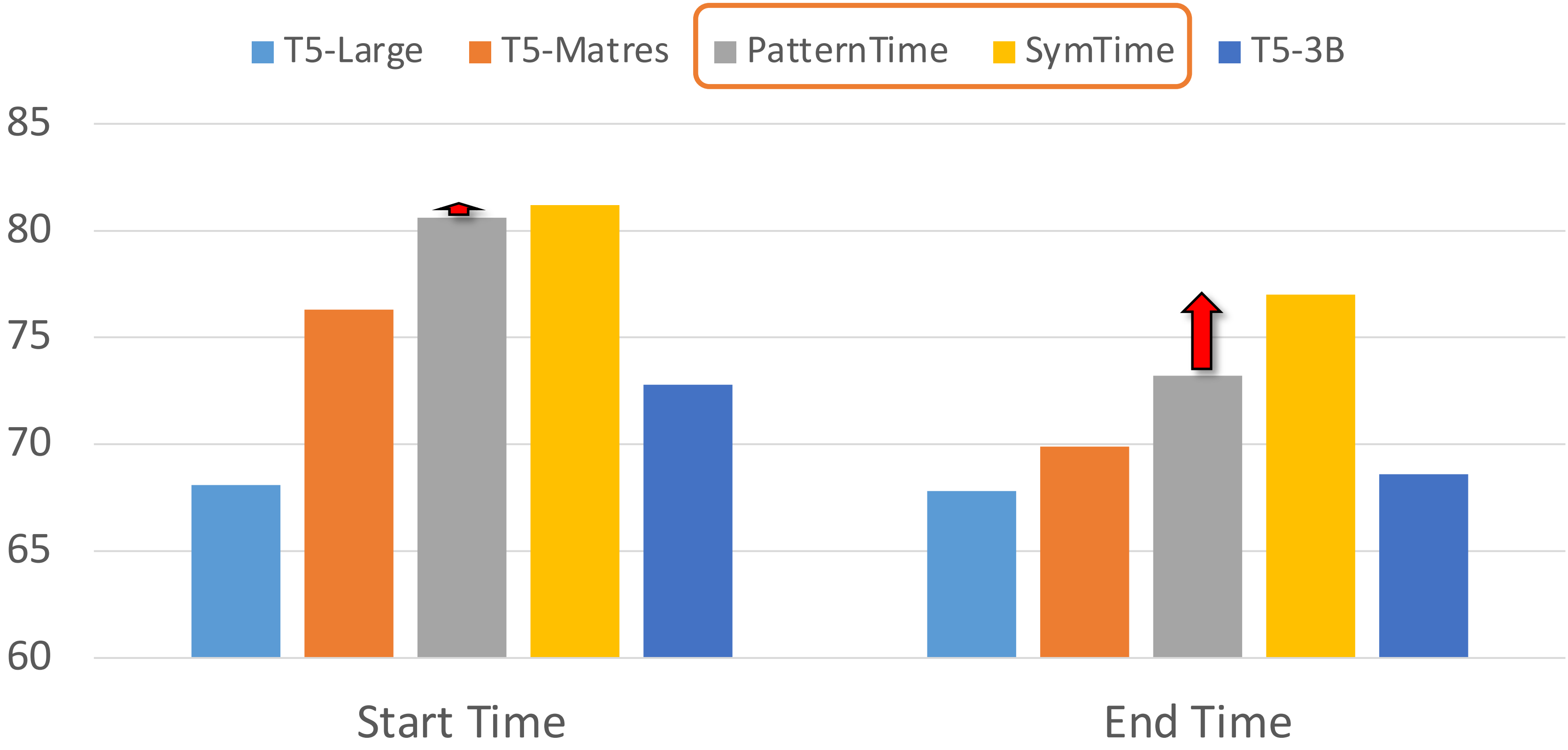
Experiments: TRACIE

- On uniform-prior training data



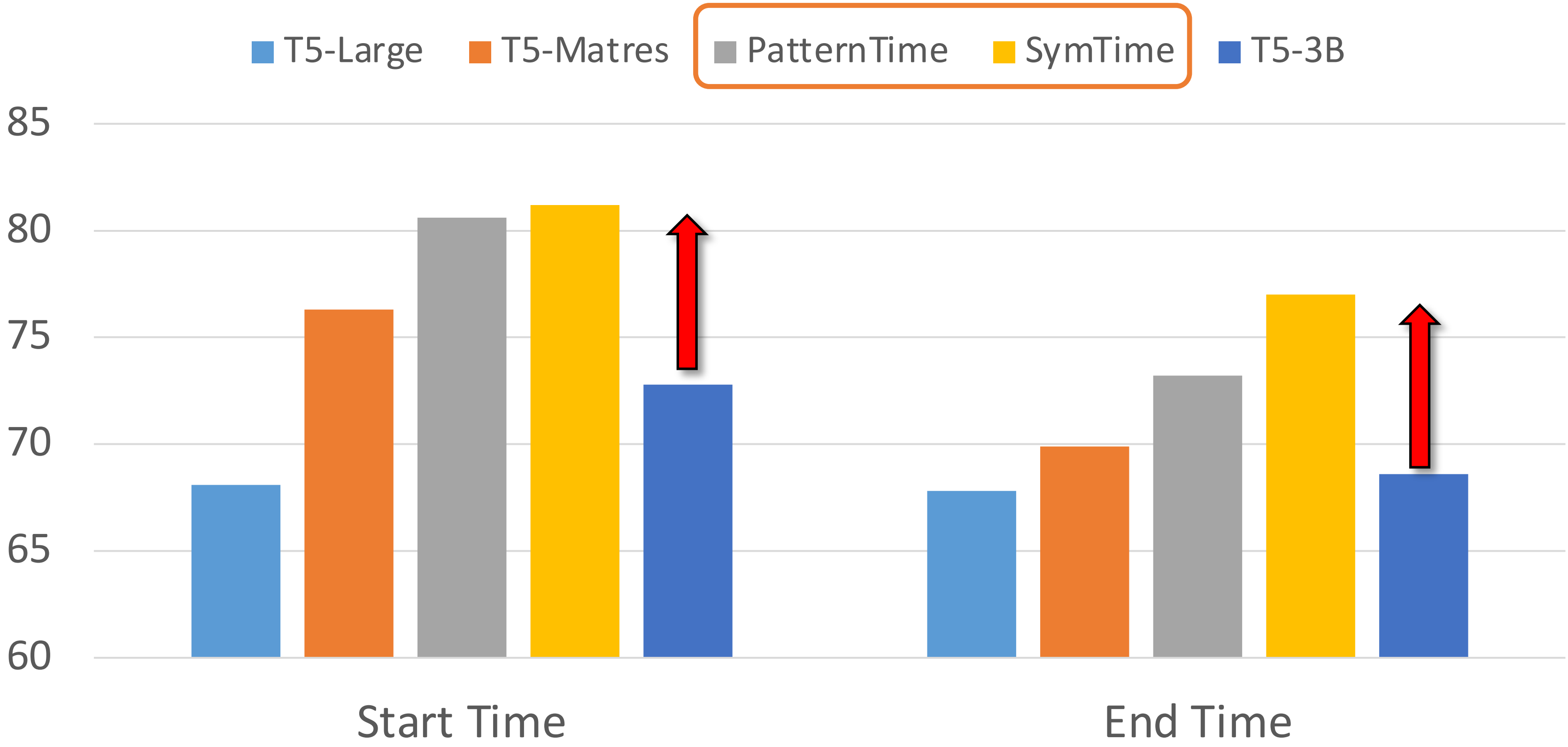
Experiments: TRACIE

- On uniform-prior training data



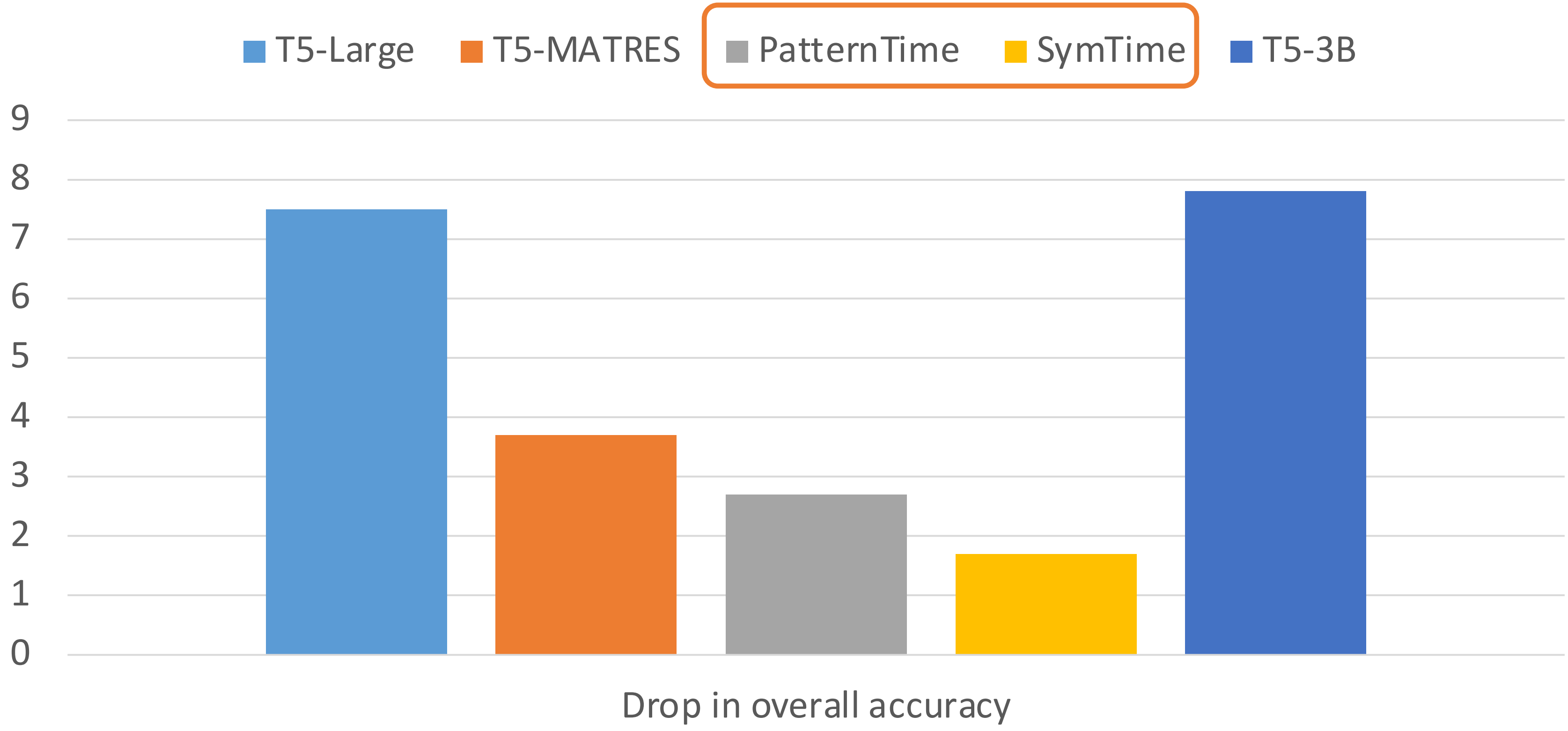
Experiments: TRACIE

- On uniform-prior training data



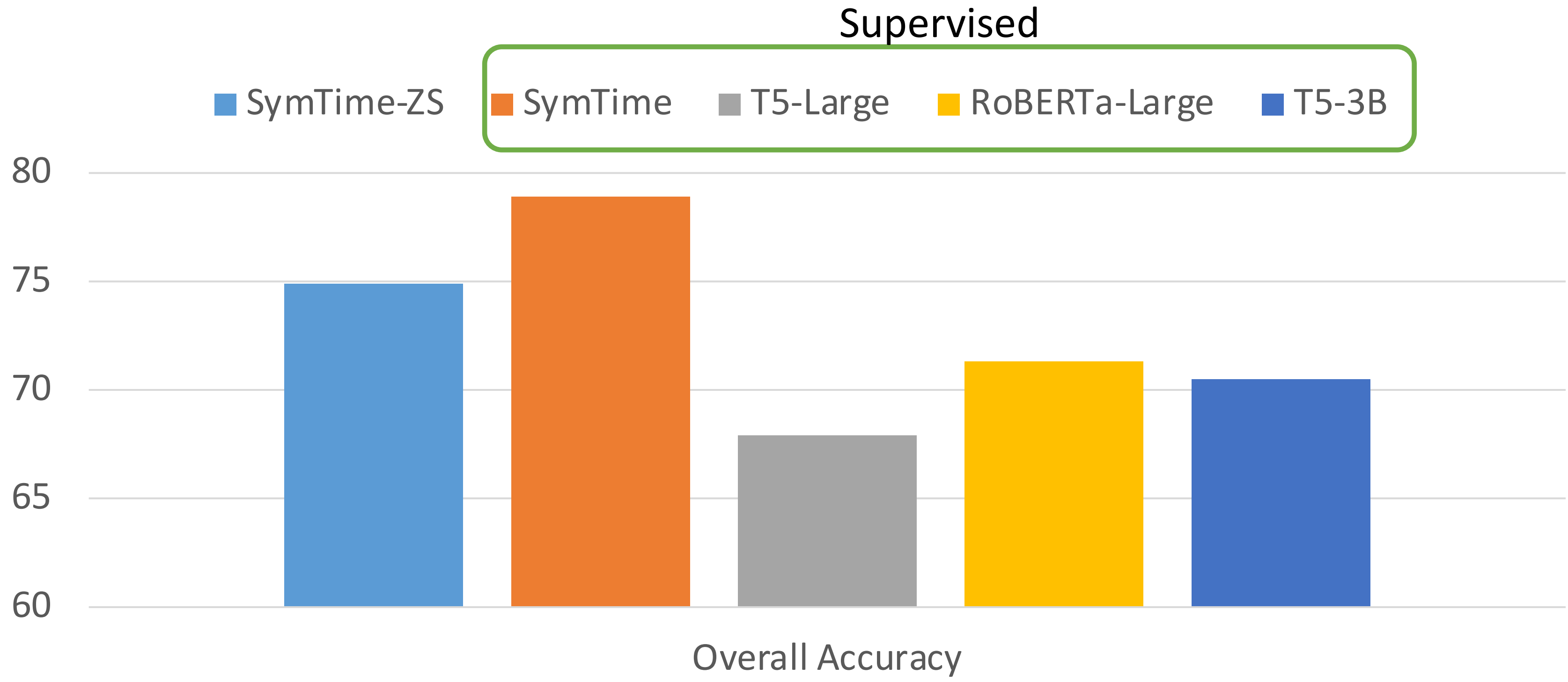
Experiments: TRACIE

- Uniform-prior v. IID training data
- Same test set



Experiments: TRACIE

- SymTime as a zero-shot model (Symtime-ZS)
  - □ Because models are initialized by distant supervision
  - □ Uses no TRAICE supervision
- On uniform-prior training data



Conclusion

- We present TRACIE
  - A temporal benchmark on implicit events
  - 5.5k NLI queries about start and end time
- We present PatternTime
  - Trained from automatically extracted distant supervision
  - Within/cross-sentence extraction for implicit event understanding
- We present SymTime
  - Symbolically combine start time and duration
  - Improves over all baselines
  - Does well even without task-specific supervision

- More experiments and discussions in the paper!
- Thank you!

code, data and paper