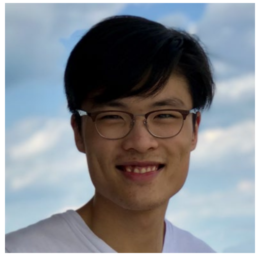
Temporal Commonsense

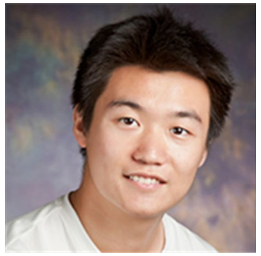
Dan Roth

Department of Computer & Information Science

University of Pennsylvania

With Ben Zhou, Qiang Ning, Daniel Khashabi





ACL'20

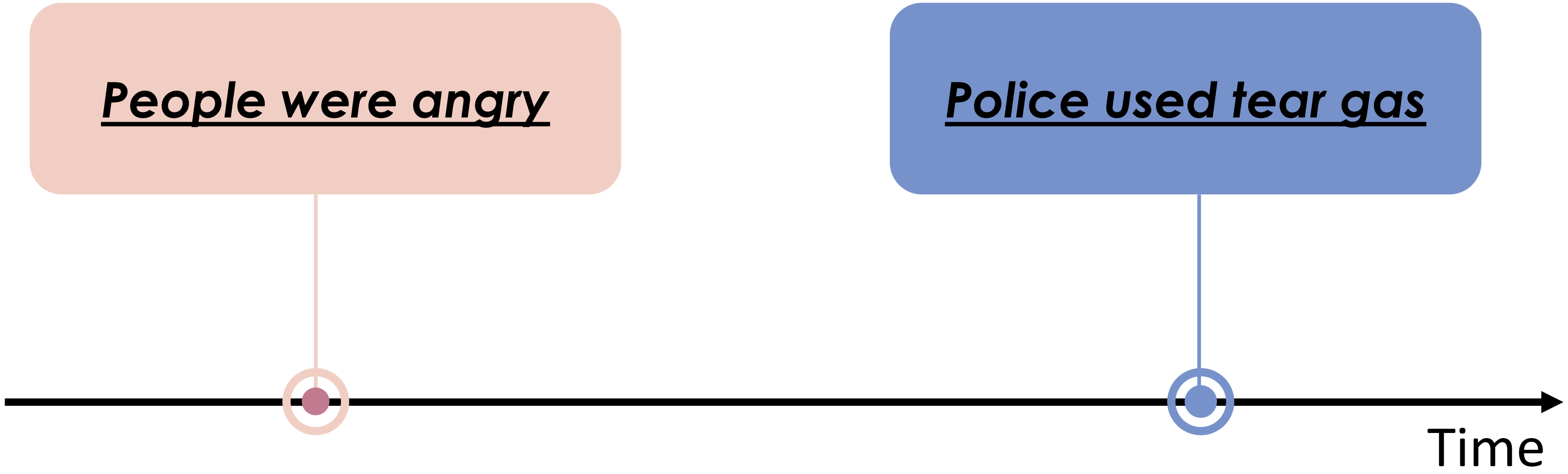
July 2020

Understanding Time is Important

People were angry

Police used tear gas

Understanding Time is Important



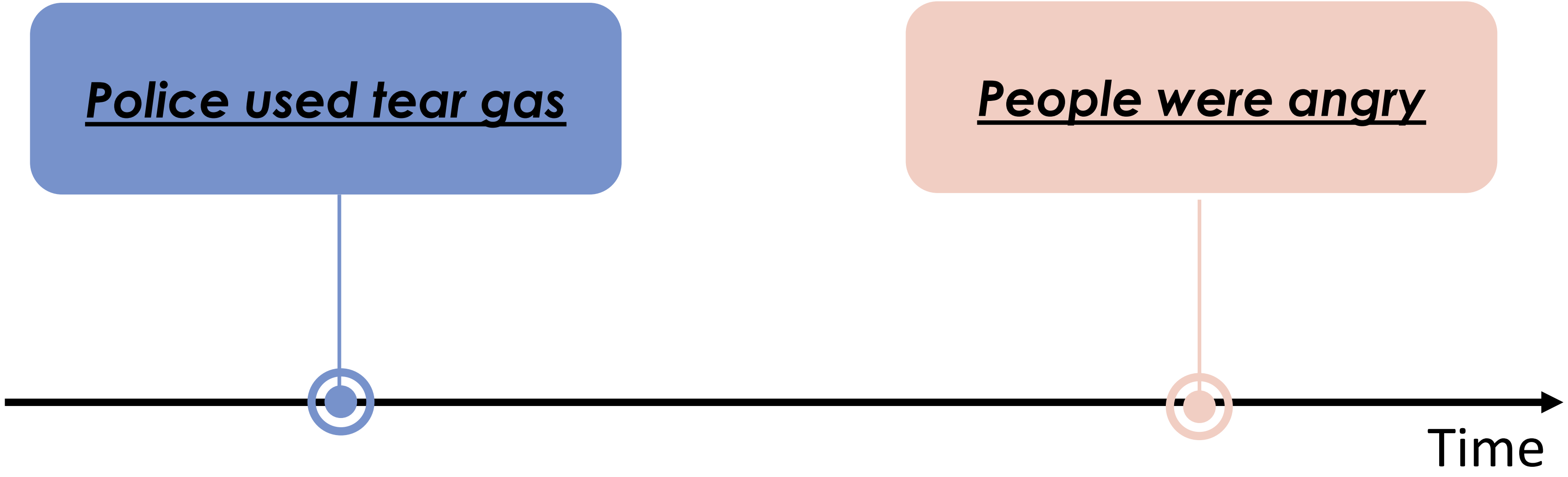
People were angry

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).

Understanding Time is Important



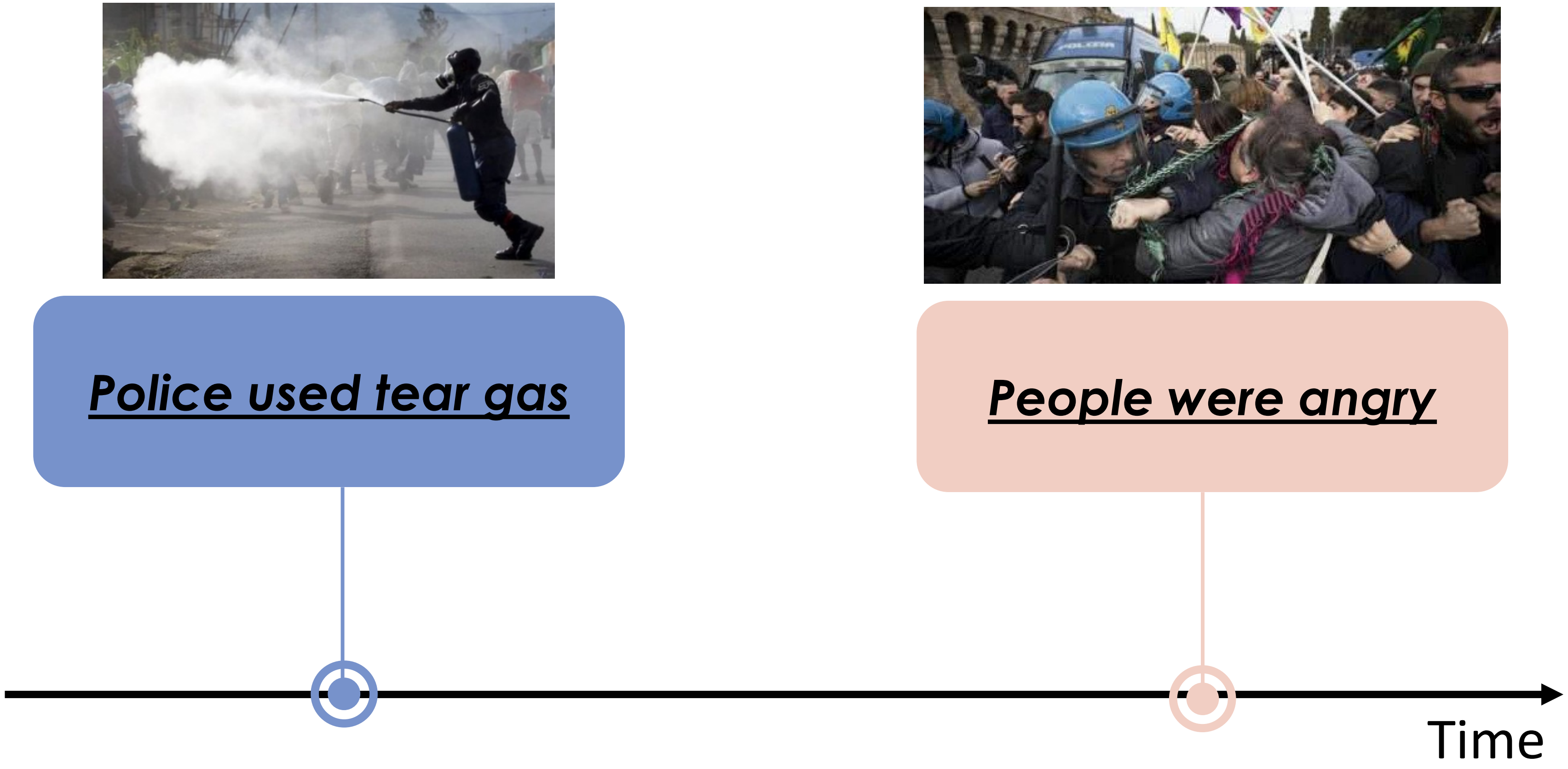
Police used tear gas

People were angry

Time

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

Understanding Time is Important



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**.

Understanding Time

- Natural Language rarely communicates explicit temporal information

Police used tear gas

People were angry

- Vagueness with respect to time is inherent in natural language
  - □ But some of it can be handled using inference and (commonsense) knowledge

Understanding Time

- Natural Language rarely communicates explicit temporal information

> Police used tear gas starting **at 7pm on Saturday and stopped at 7:30;**….
> People were angry at the police **between 7:01 and 9pm**.

Police used tear gas

People were angry

- Vagueness with respect to time is inherent in natural language
  - □ But some of it can be handled using inference and (commonsense) knowledge

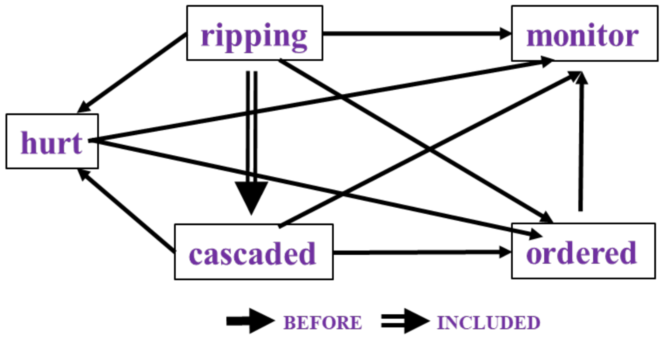
Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

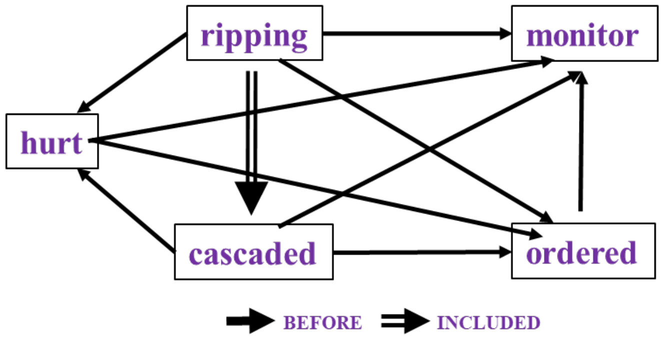
- In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.



Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

- 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.

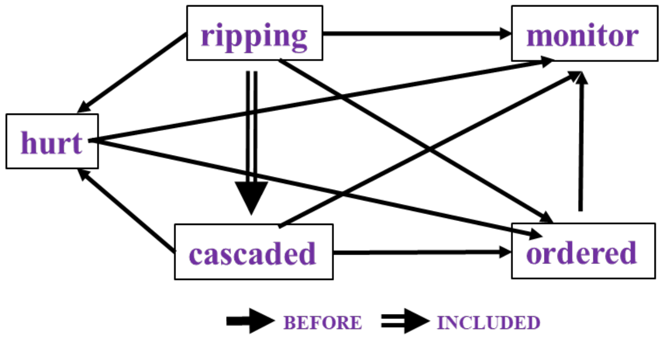


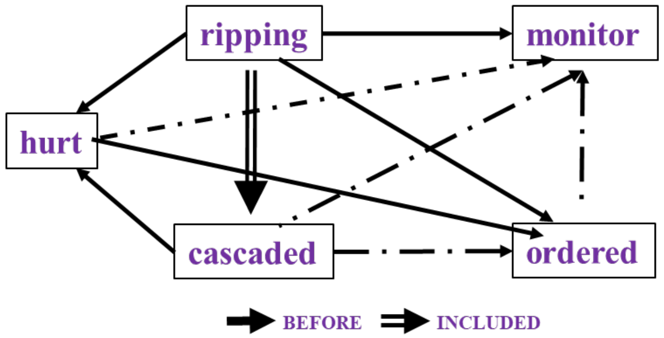
- Difficult task— even for human annotators (O(N^2) edges)

Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

- 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.



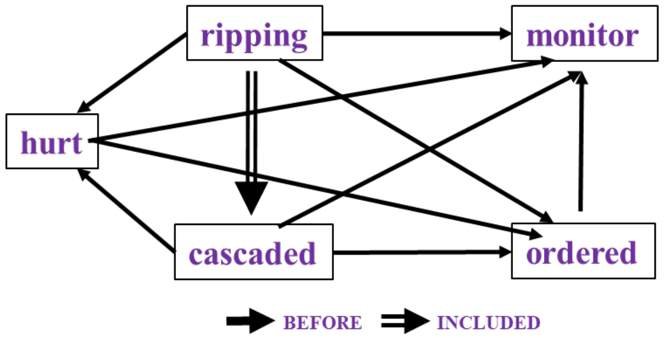


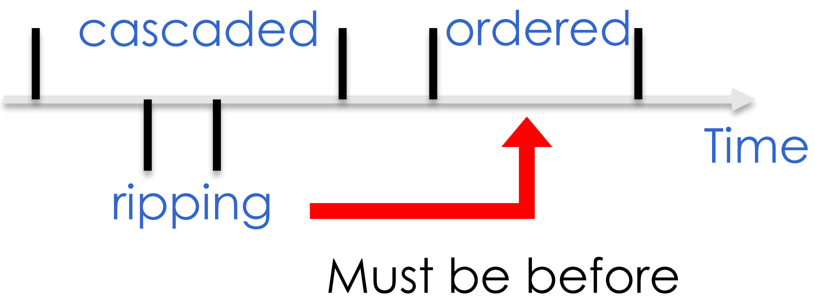
- Difficult task— even for human annotators (O(N^2) edges)

Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

- 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.



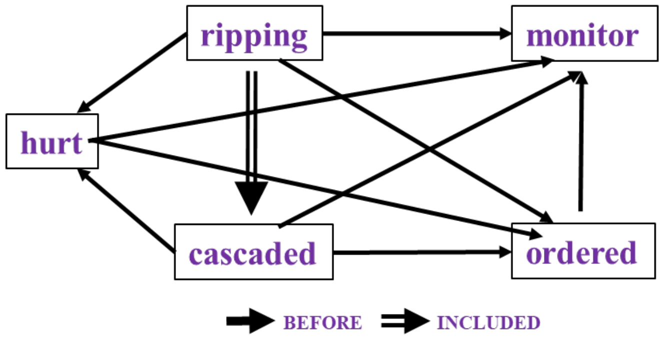


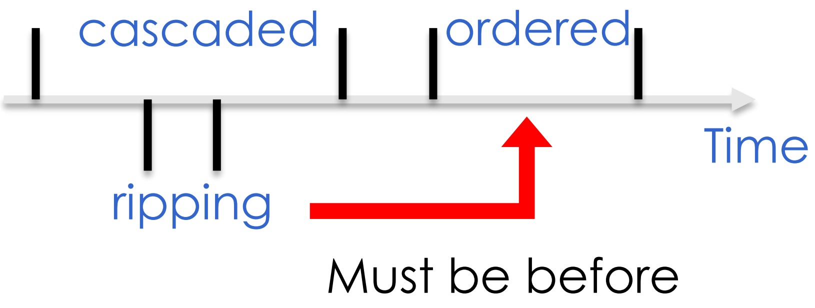
- Difficult task— even for human annotators (O(N^2) edges)
- Approaches exploit **strong expectations** from the output: Commonsense
  - Transitivity
  - Some events tend to precede others, or follow others

Temporal Relations

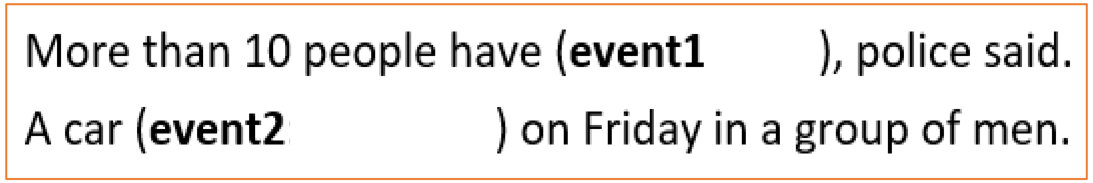
- The most commonly studied problem in temporal NLP is that of temporal relations

- 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.





- Difficult task— even for human annotators (O(N^2) edges)
- Approaches exploit **strong expectations** from the output: Commonsense
  - Transitivity
  - Some events tend to precede others, or follow others



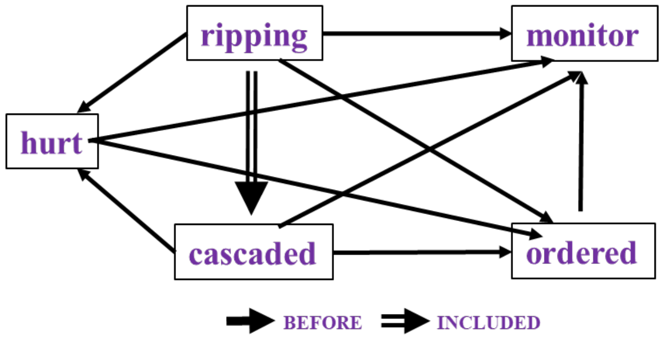
More than 10 people have (**event1**), police said.

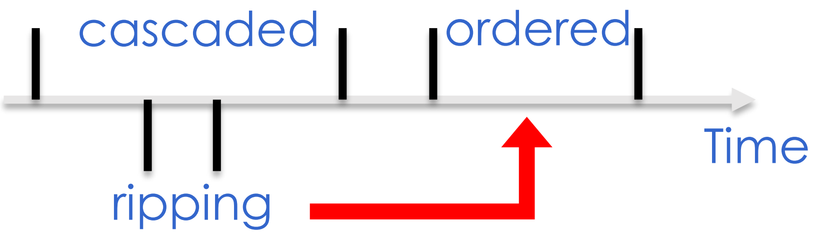
A car (**event2**) on Friday in a group of men.

Temporal Relations

- The most commonly studied problem in temporal NLP is that of temporal relations

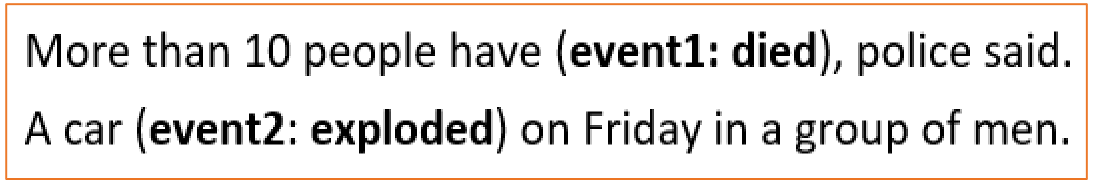
- In Los Angeles that lesson was brought home today when tons of earth **cascaded** down a hillside, **ripping** two houses from their foundations. No one was **hurt**, but firefighters **ordered** the evacuation of nearby homes and said they'll **monitor** the shifting ground until March 23rd.





Must be before

- Difficult task— even for human annotators (O(N^2) edges)
- Approaches exploit **strong expectations** from the output: Commonsense
  - □ Transitivity
  - □ Some events tend to precede others, or follow others



More than 10 people have (**event1: died**), police said.

A car (**event2**: **exploded**) on Friday in a group of men.

Commonsense: Temporal Relations

- **TemProb:** Temporal Relation Probabilistic Knowledge Base [Ning et al. NAACL'18]
- Run initial temporal relationssystem 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

Commonsense: Temporal Relations

- **TemProb:** Temporal Relation Probabilistic Knowledge Base [Ning et al. NAACL’18]
- Run initial temporal relationssystem 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

Commonsense: Temporal Relations

- **TemProb:** Temporal Relation Probabilistic Knowledge Base [Ning et al. NAACL’18]
- Run initial temporal relationssystem 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

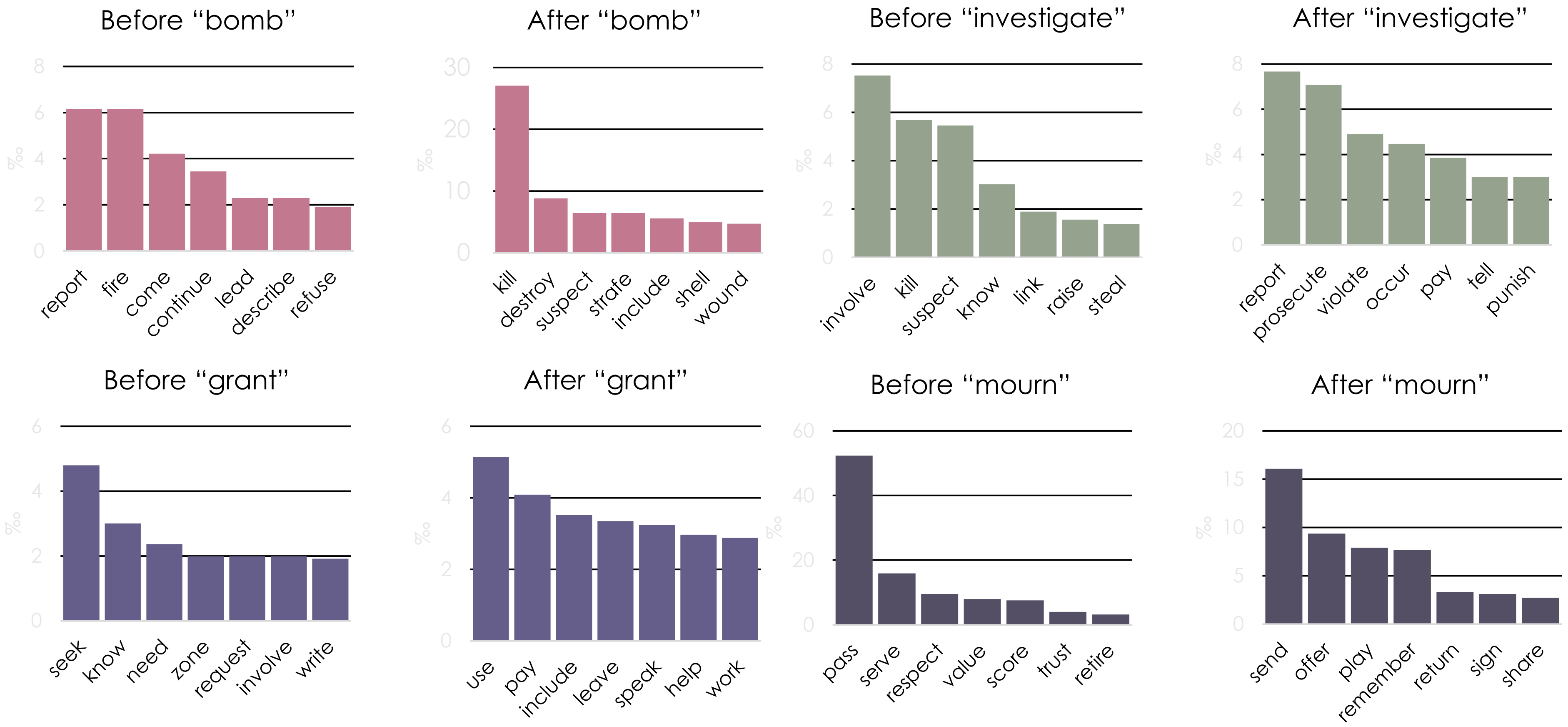
Commonsense: Temporal Relations

- **TemProb:** Temporal Relation Probabilistic Knowledge Base [Ning et al. NAACL'18]
- Run initial temporal relationssystem on New York Times 1987-2007, #Articles~1M
- Identify events; identify temporal order
- 80M temporal relations
- Noisy statistics is sufficient to give good priors.

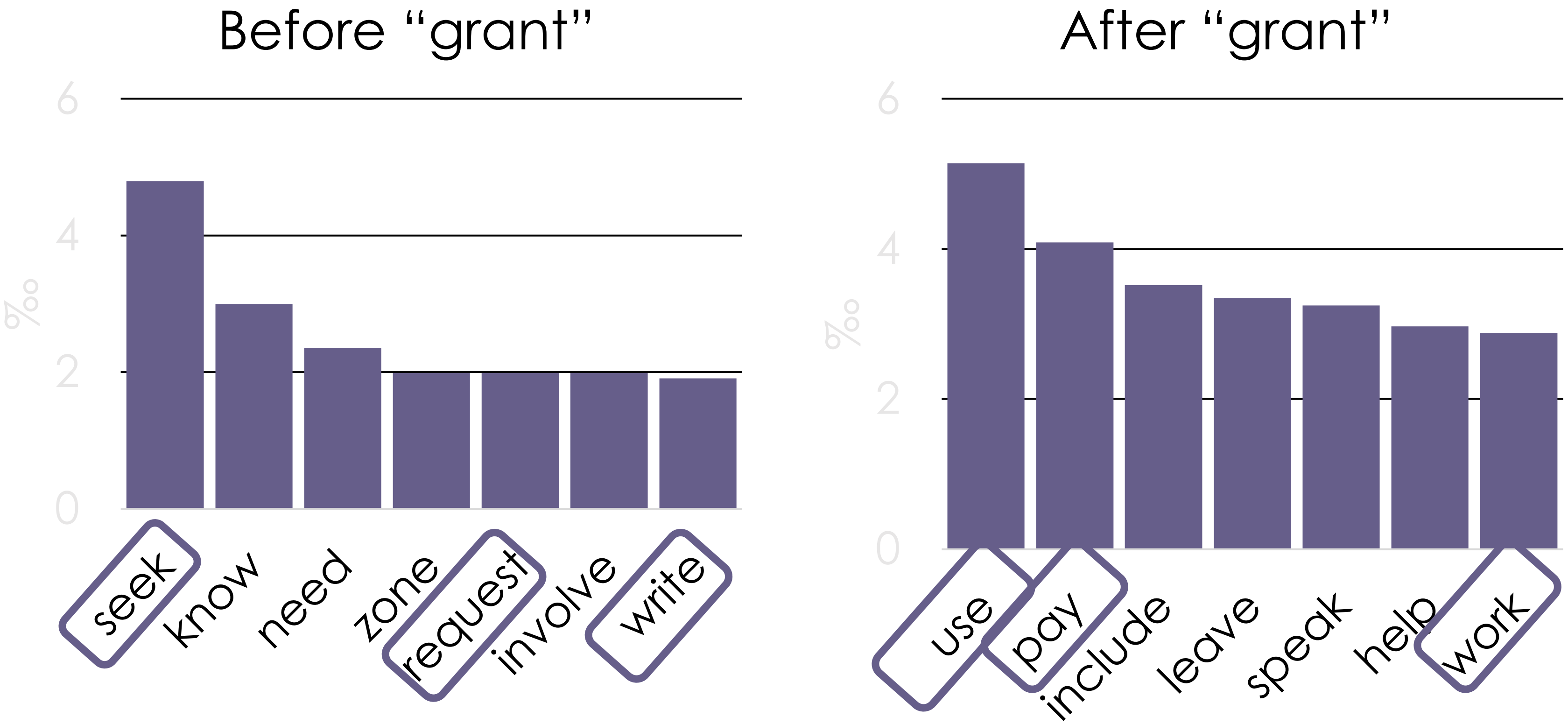
Example pairs		**Temporal Before (%)**	**Temporal After (%)**
Text Before	**Text After**		
Ask	Help	86	9
Attend	Schedule	1	82
Accept	Propose	10	77
Die	Explode	14	83

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

Event Order Distributions

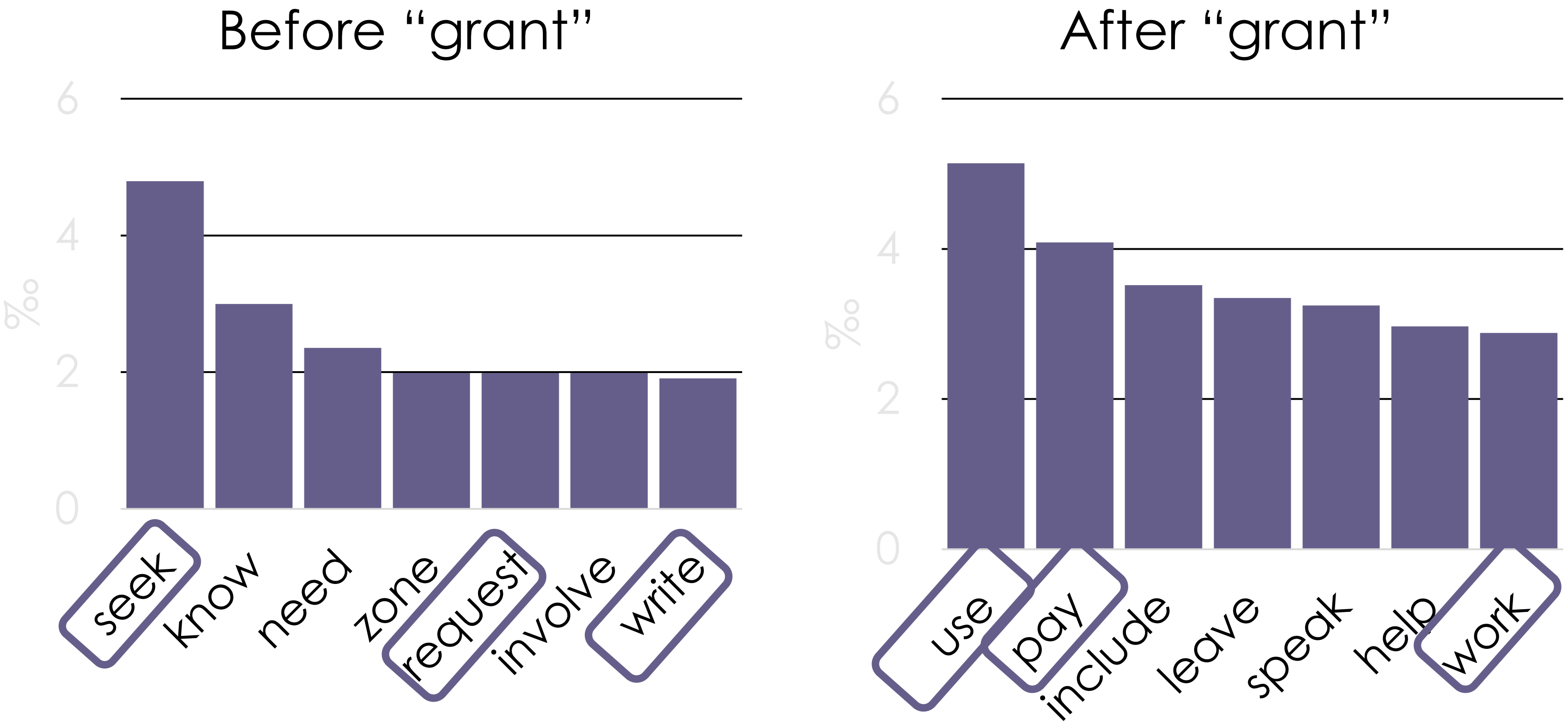


Event Order Distributions



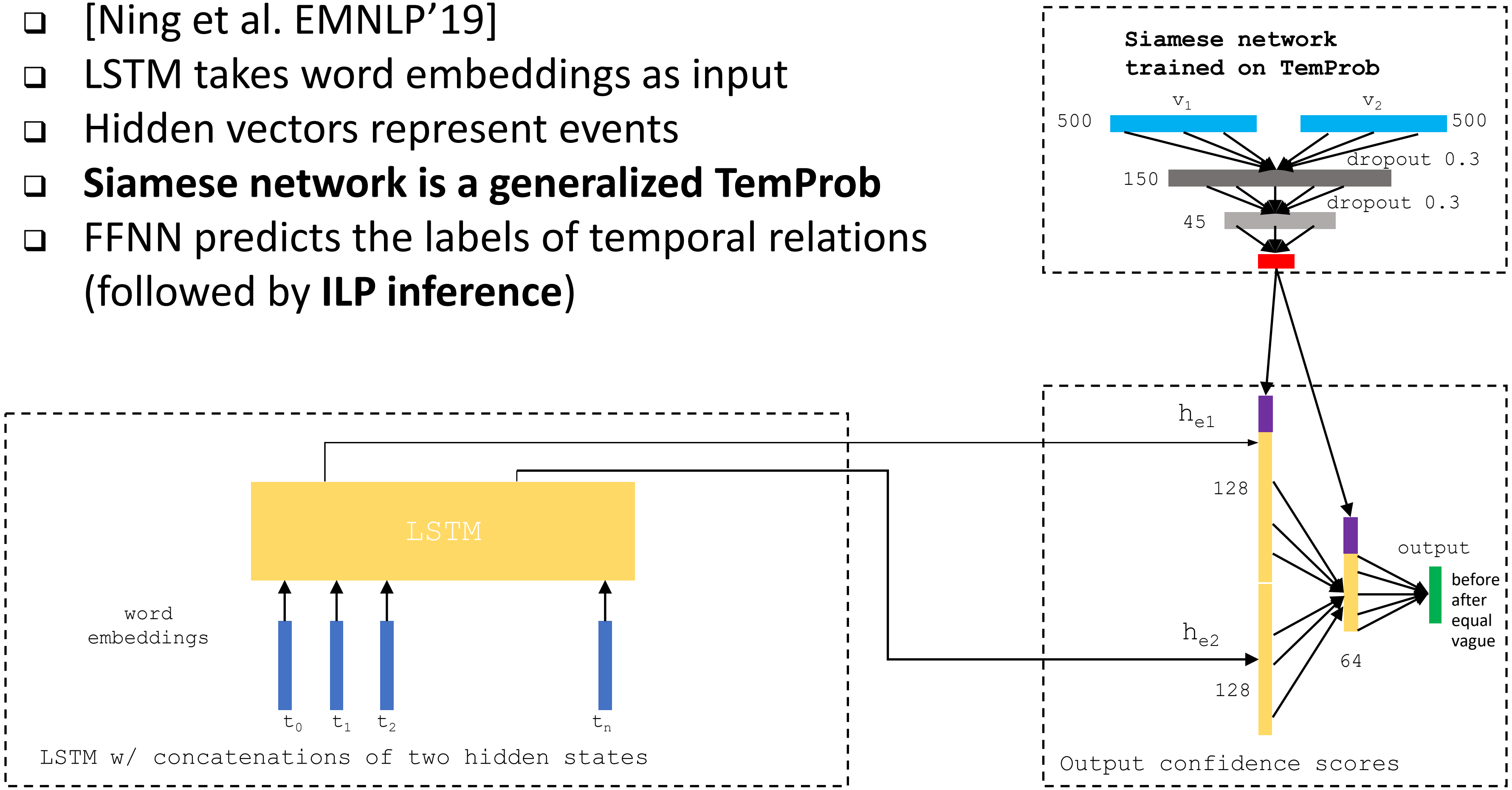
Event Order Distributions

- These statistical "symbolic" priors can be used as is, or within a neural architecture



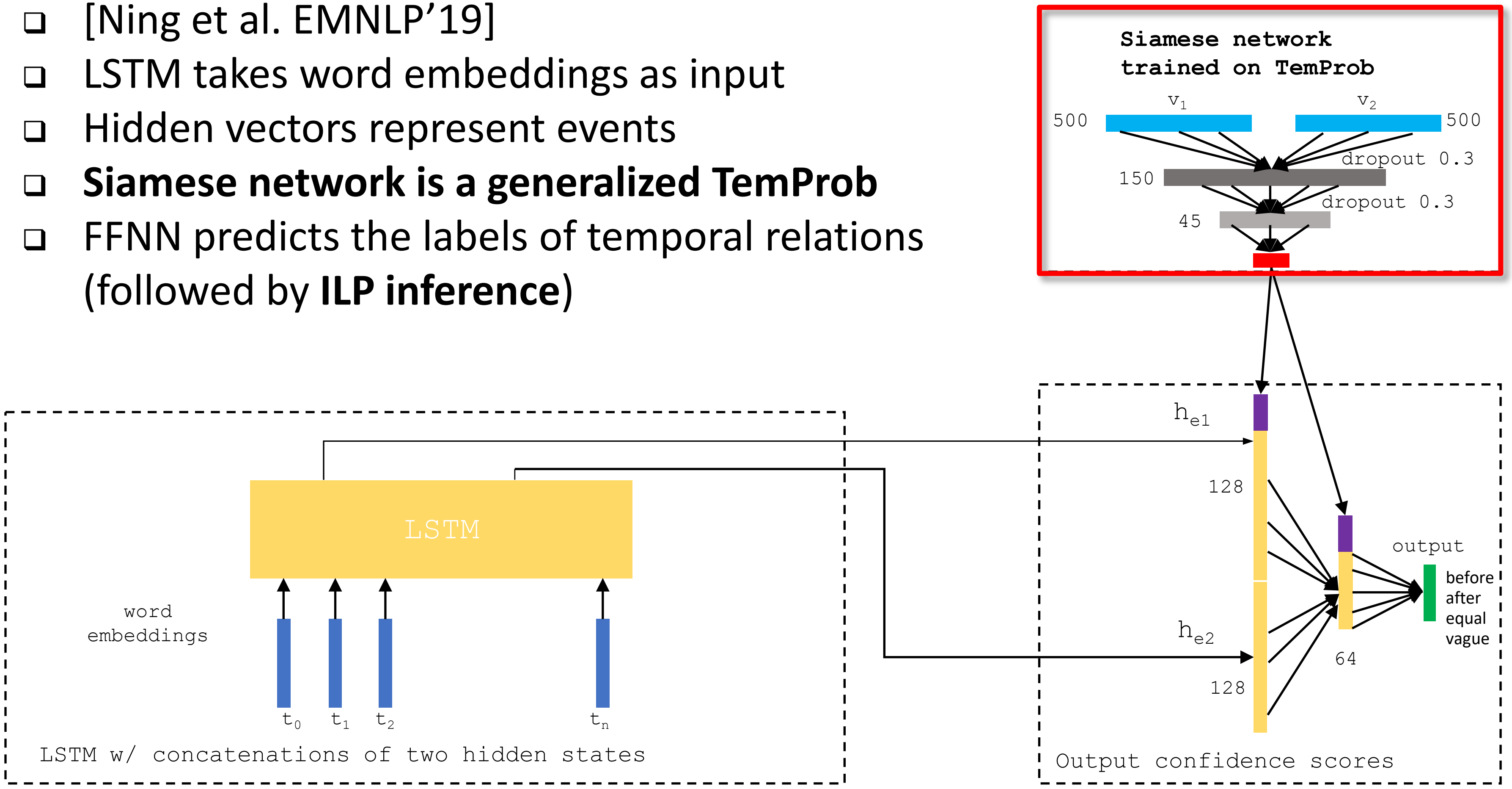
A Neural Architecture for Temporal Relations

- [Ning et al. EMNLP’19]
- LSTM takes word embeddings as input
- Hidden vectors represent events
- **Siamese network is a generalized TemProb**
- FFNN predicts the labels of temporal relations (followed by **ILP inference**)



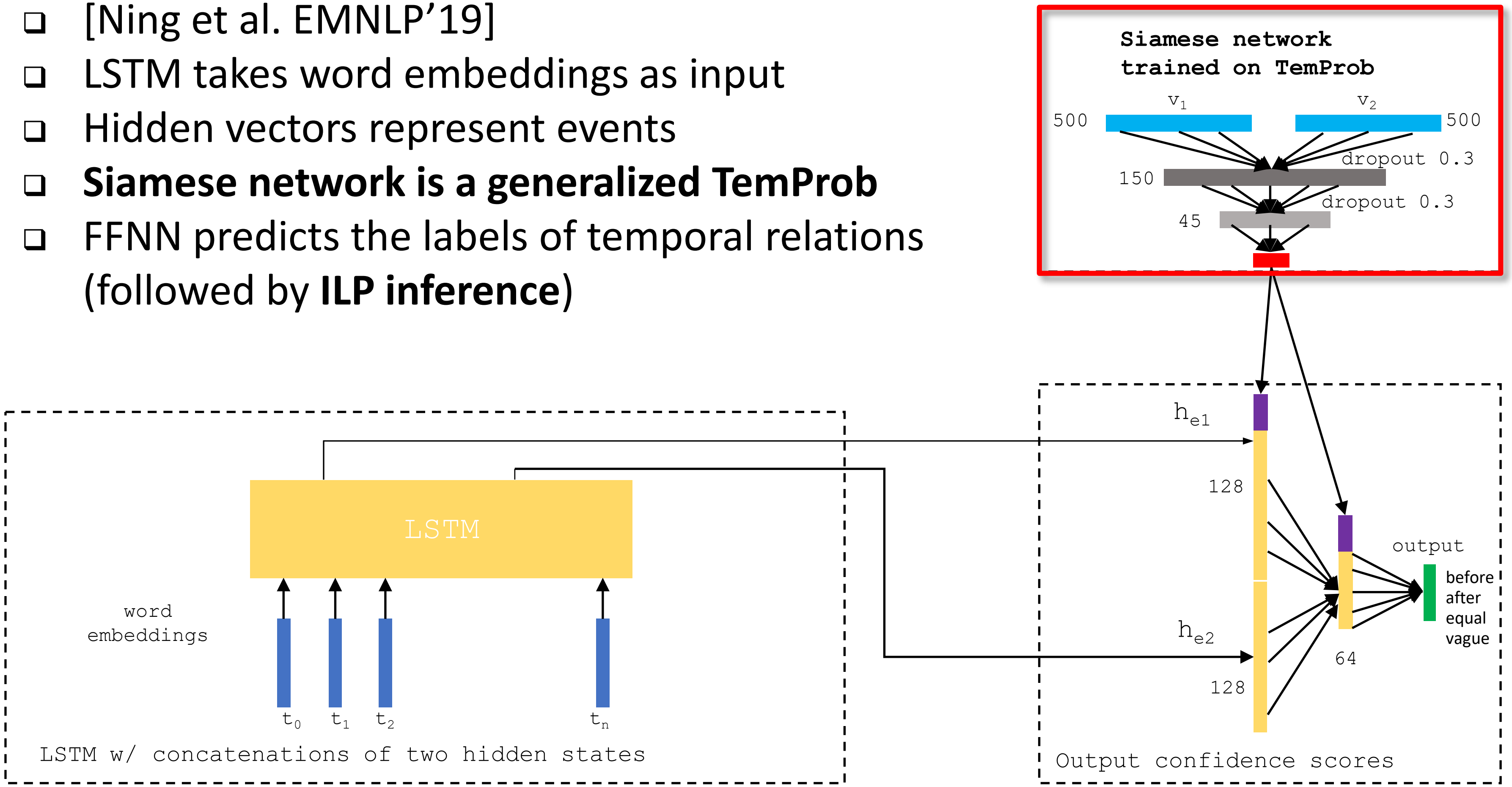
A Neural Architecture for Temporal Relations

- [Ning et al. EMNLP’19]
- LSTM takes word embeddings as input
- Hidden vectors represent events
- **Siamese network is a generalized TemProb**
- FFNN predicts the labels of temporal relations (followed by **ILP inference**)



A Neural Architecture for Temporal Relations

- [Ning et al. EMNLP’19]
- LSTM takes word embeddings as input
- Hidden vectors represent events
- **Siamese network is a generalized TemProb**
- FFNN predicts the labels of temporal relations (followed by **ILP inference**)



We should address additional aspects of temporal commonsense…

Temporal Commonsense

- *"will"* or *"will not"*?

Dr. Porter is **taking a vacation** and _______ be able to see you soon.

Dr. Porter is **taking a walk** and ___ be able to see you soon.

Temporal Commonsense

- *"will"* or *"will not"*?

Days

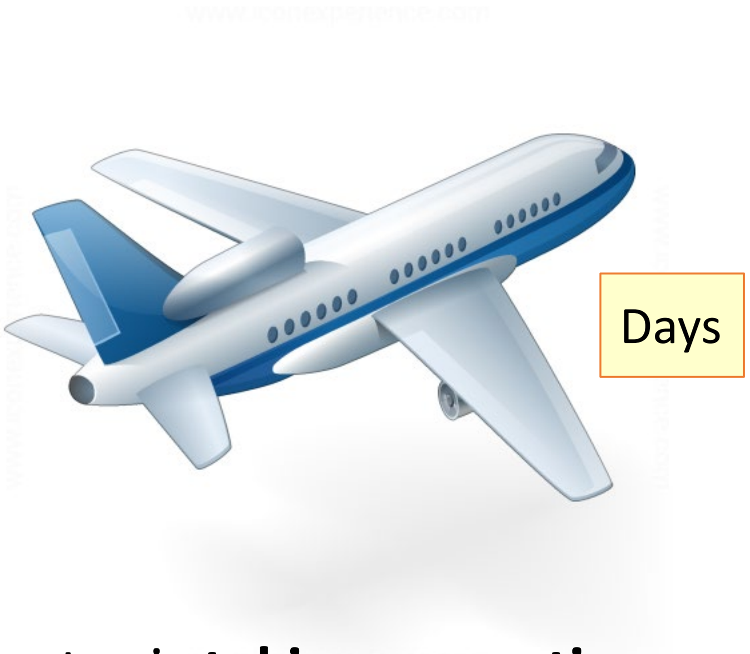
Dr. Porter is **taking a vacation** and _______ be able to see you soon.

Minutes

Dr. Porter is **taking a walk** and ___ be able to see you soon.

Temporal Commonsense

- “*will*” or “*will not*”?



Dr. Porter is **taking a vacation** and will not be able to see you soon.

Dr. Porter is **taking a walk** and will be able to see you soon.

Defining the Temporal Commonsense Challenge

- **Events** are associated with time
  - □ Beyond order – **Typical Time, Duration, Frequency**
- Most **attributes** and **relations** change over time
  - □ Employment, schooling, location, nationality, headquarters, president, event participation , etc.
- **Knowledge Bases** (knowledge Graphs) need to be qualified temporally

Defining the Temporal Commonsense Challenge

- **Events** are associated with time
  - □ Beyond order – **Typical Time, Duration, Frequency**
- Most **attributes** and **relations** change over time
  - □ Employment, schooling, location, nationality, headquarters, president, event participation , etc.
- **Knowledge Bases** (knowledge Graphs) need to be qualified temporally

Senator Obama & President Obama

Tom Cruise has three spouses

Defining the Temporal Commonsense Challenge

- **Events** are associated with time
  - □ Beyond order – **Typical Time, Duration, Frequency**
- Most **attributes** and **relations** change over time
  - □ Employment, schooling, location, nationality, headquarters, president, event participation , etc.
- **Knowledge Bases** (knowledge Graphs) need to be qualified temporally

- Goal: Represent a range of temporal aspects of conditions that change over time

Senator Obama & President Obama

Tom Cruise has three spouses

Defining the Temporal Commonsense Challenge

- **Events** are associated with time
  - □ Beyond order – **Typical Time, Duration, Frequency**
- Most **attributes** and **relations** change over time
  - □ Employment, schooling, location, nationality, headquarters, president, event participation , etc.
- **Knowledge Bases** (knowledge Graphs) need to be qualified temporally
- Goal: Represent a range of temporal aspects of conditions that change over time

Temporal information is often **implicit** in text

Senator Obama & President Obama

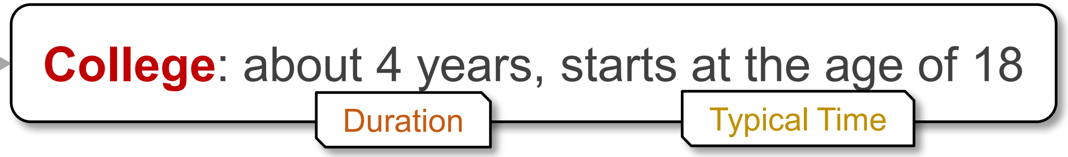
Tom Cruise has three spouses

The Temporal Commonsense Challenge

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.

The Temporal Commonsense Challenge

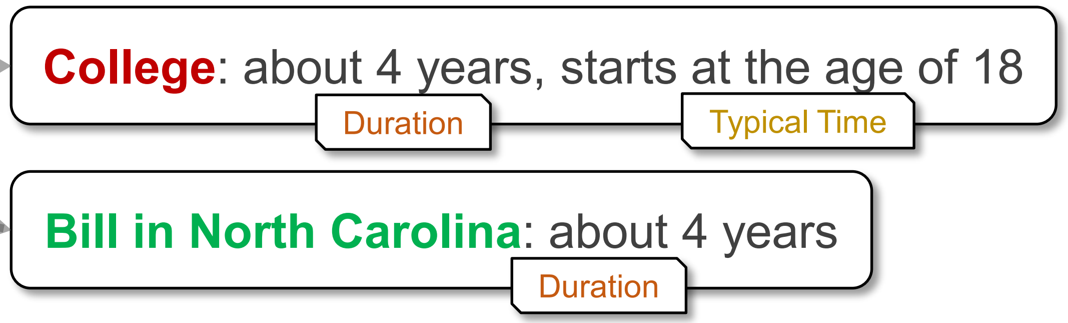
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.



College: about 4 years, starts at the age of 18

The Temporal Commonsense Challenge

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.



College: about 4 years, starts at the age of 18

Duration

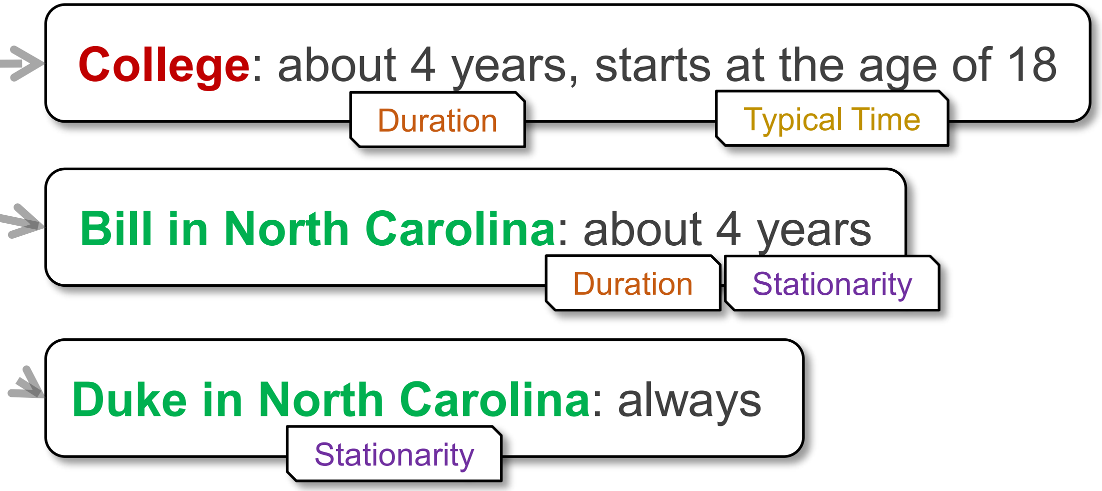
Typical Time

Bill in North Carolina: about 4 years

Duration

The Temporal Commonsense Challenge

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.



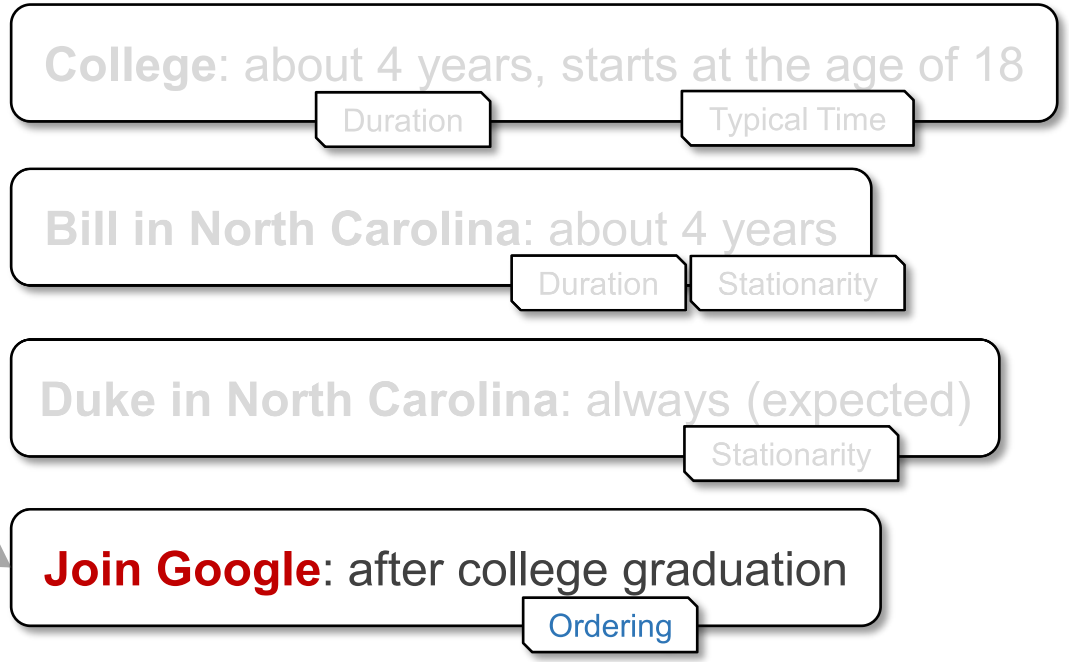
College: about 4 years, starts at the age of 18

Bill in North Carolina: about 4 years

Duke in North Carolina: always

The Temporal Commonsense Challenge

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.



College: about 4 years, starts at the age of 18

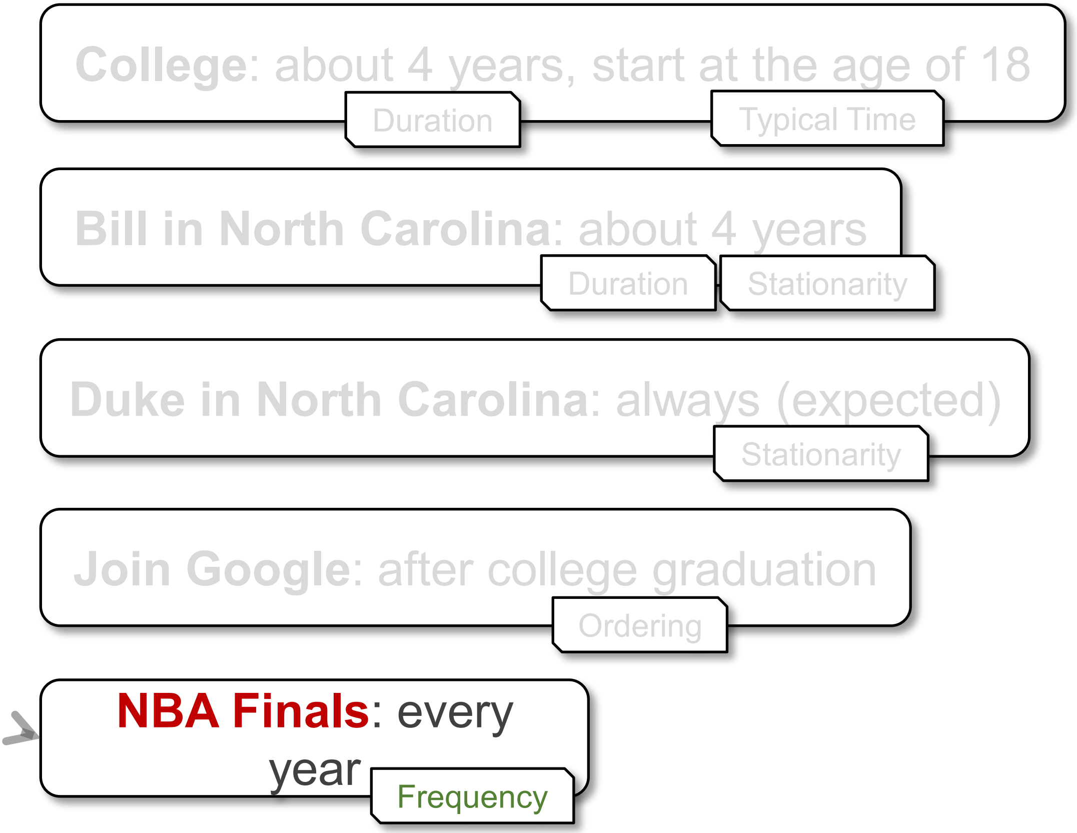
Bill in North Carolina: about 4 years

Duke in North Carolina: always (expected)

Join Google: after college graduation

The Temporal Commonsense Challenge

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.



College: about 4 years, start at the age of 18
Duration
Typical Time

Bill in North Carolina: about 4 years
Duration
Stationarity

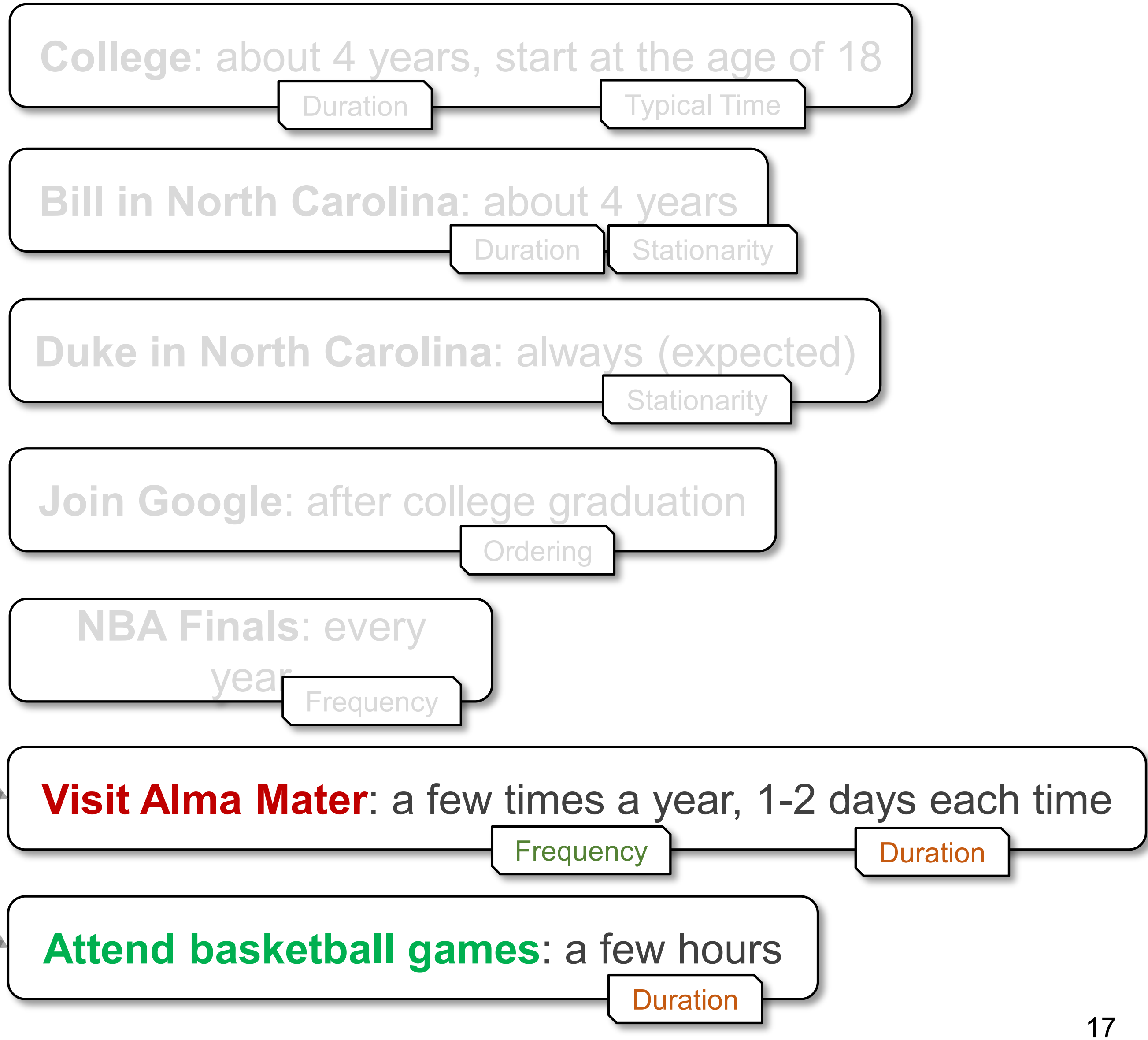
Duke in North Carolina: always (expected)
Stationarity

Join Google: after college graduation
Ordering

NBA Finals: every year
Frequency

The Temporal Commonsense Challenge

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.



College: about 4 years, start at the age of 18
Duration | Typical Time

Bill in North Carolina: about 4 years
Duration | Stationarity

Duke in North Carolina: always (expected)
Stationarity

Join Google: after college graduation
Ordering

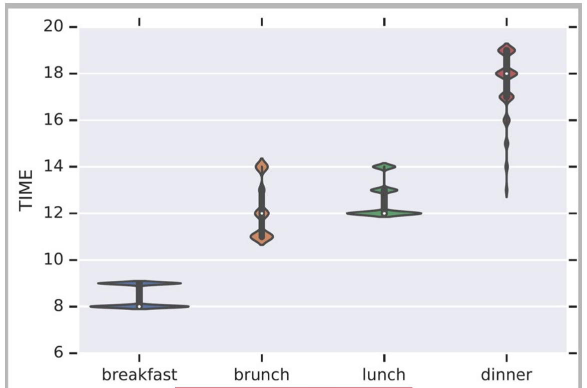
NBA Finals: every year
Frequency

Visit Alma Mater: a few times a year, 1-2 days each time
Frequency | Duration

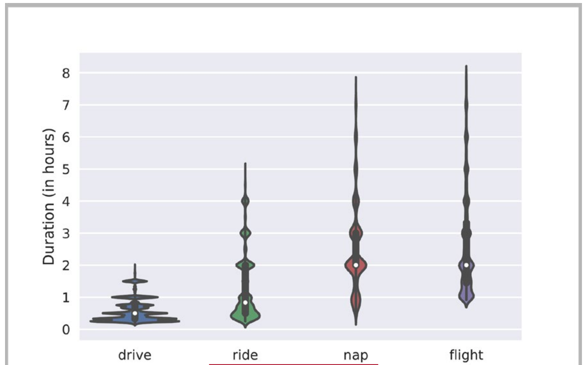
Attend basketball games: a few hours
Duration

Temporal Common Sense

- Two efforts:
  - □ A dataset MC-TACO [Zhou et al. EMNLP’19]
  - □ Acquisition + Representation [Zhou et al. ACL’20]: Duration, typical time, frequency.



Typical Time



Duration

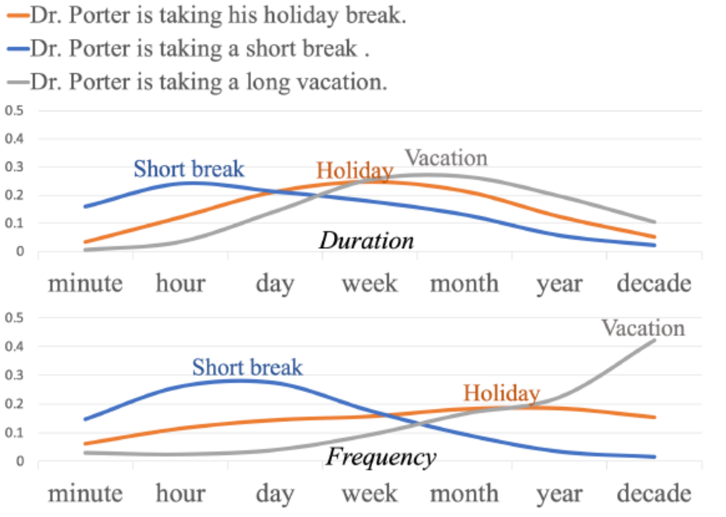
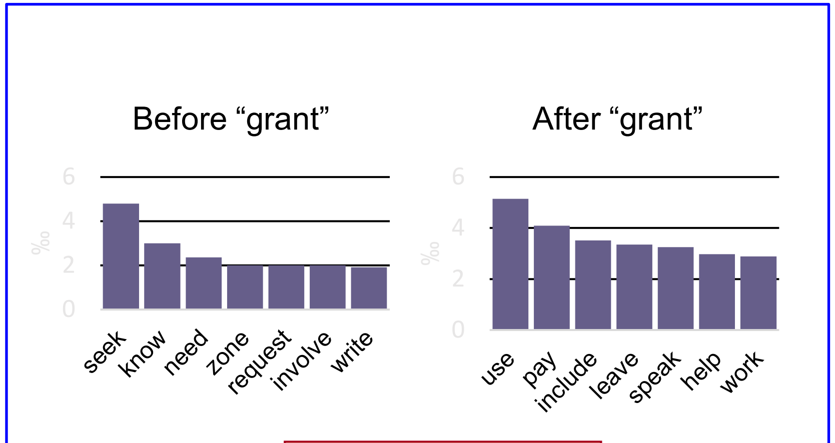


Figure 1: Our model’s predicted distributions about event **duration** and **frequency**. The model is able to distinguish fine-grained contexts and produce quality estimations.

Zhou et al. ACL’20]

[Elazar et al. ACL’19]



Typical Temporal Relations

Ning et al. NAACL’18

Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - □ **M**ultiple **C**hoice **T**empor**A**l **CO**mmon-sense
  - □ 1,893 questions; 13,225 question-answer pairs
  - □ Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

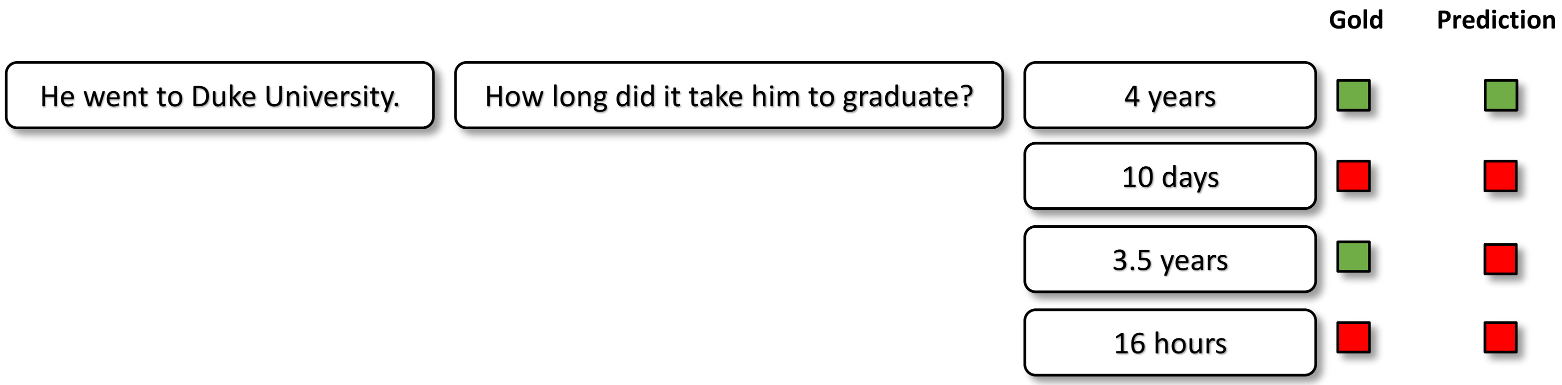
Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - **M**ultiple **C**hoice **T**empor**A**l **CO**mmon-sense
  - 1,893 questions; 13,225 question-answer pairs
  - Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

			Gold
He went to Duke University.	How long did it take him to graduate?	4 years	🟩
		10 days	🟥
		3.5 years	🟩
		16 hours	🟥

Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - **M**ultiple **C**hoice **T**empor**A**l **CO**mmon-sense
  - 1,893 questions; 13,225 question-answer pairs
  - Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering



			Gold	Prediction
He went to Duke University.	How long did it take him to graduate?	4 years	🟩	🟩
		10 days	🟥	🟥
		3.5 years	🟩	🟥
		16 hours	🟥	🟥

Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - □ **M**ultiple **C**hoice **T**empor**A**l **CO**mmon-sense
  - □ 1,893 questions; 13,225 question-answer pairs
  - □ Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

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

Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - □ **M**ultiple **C**hoice **T**empor**A**l **CO**mmon-sense
  - □ 1,893 questions; 13,225 question-answer pairs
  - □ Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

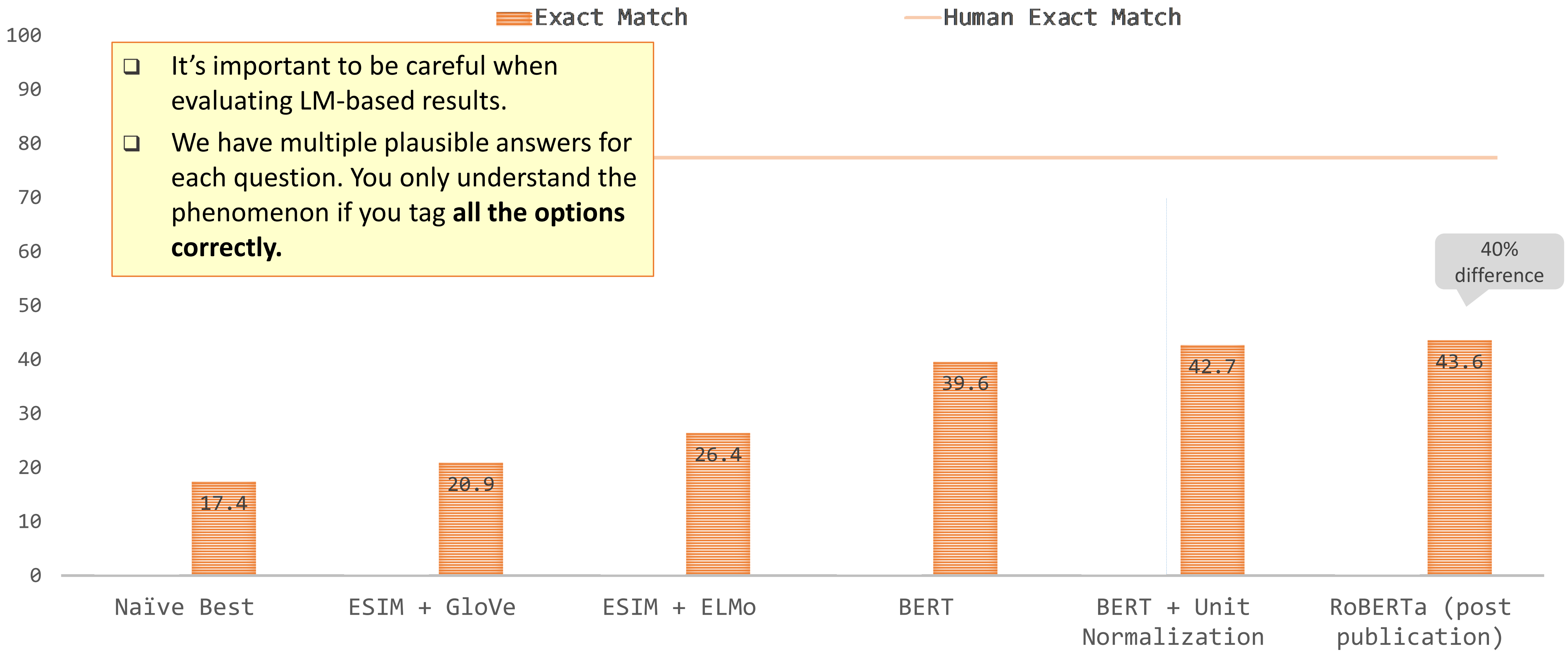
He went to Duke University.

How long did it take him to graduate?

	Gold	Prediction	
4 years	🟩	🟩	✓
10 days	🟥	🟥	✓
3.5 years	🟩	🟥	✗
16 hours	🟥	🟥	✓

- □ Exact Match: the percentage of questions of which **all** candidates are predicted correctly (here: 0.0)
- □ F1: Gives partial credit (credits “accidental” correct perditions (here: 66.7%)

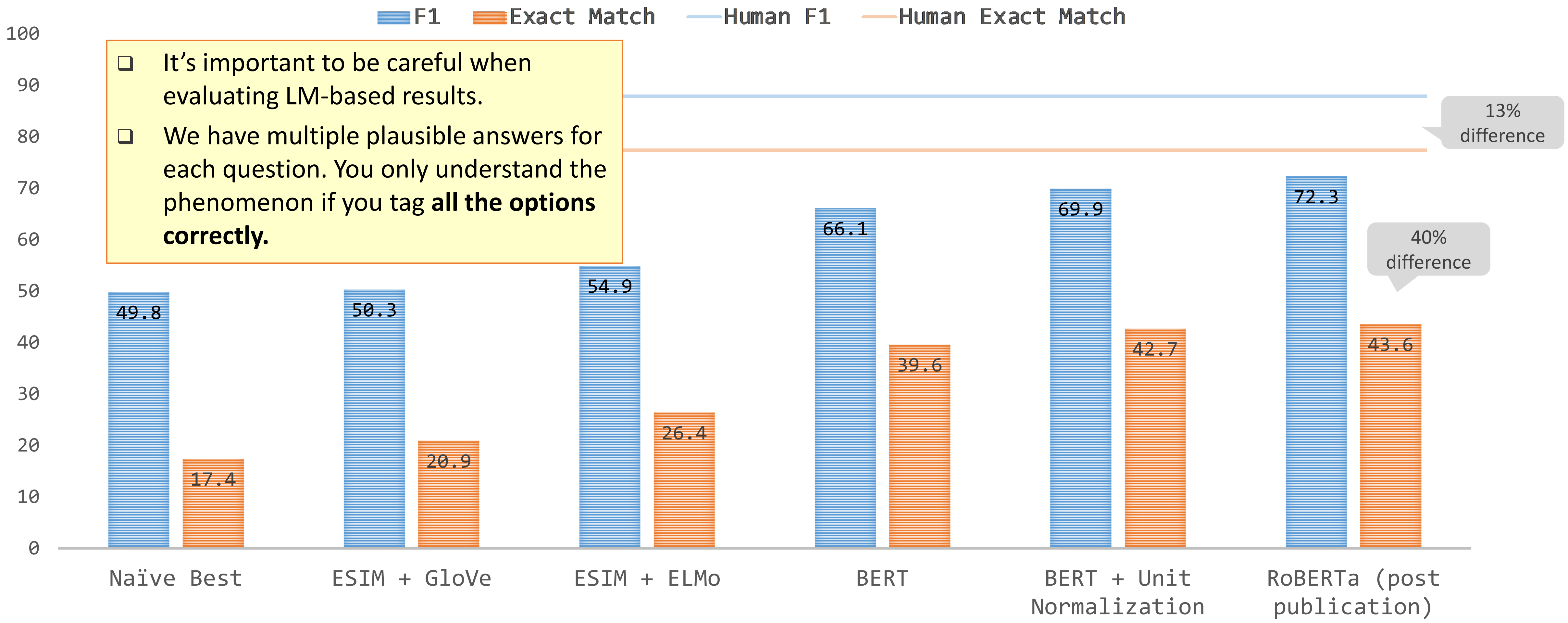
Results: We are Far (from where we want to be)



- ❑ It’s important to be careful when evaluating LM-based results.
- ❑ We have multiple plausible answers for each question. You only understand the phenomenon if you tag **all the options correctly.**

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)

Results: We are Far (from where we want to be)



- ❑ It's important to be careful when evaluating LM-based results.
- ❑ We have multiple plausible answers for each question. You only understand the phenomenon if you tag **all the options correctly.**

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)

MC-TACO🌮: A Temporal Commonsense Dataset [Zhou et al. EMNLP'19]

- Stationarity:
  - Paul Simon is in NYC. Let's go see him.
  - The Empire State Building is in NYC.

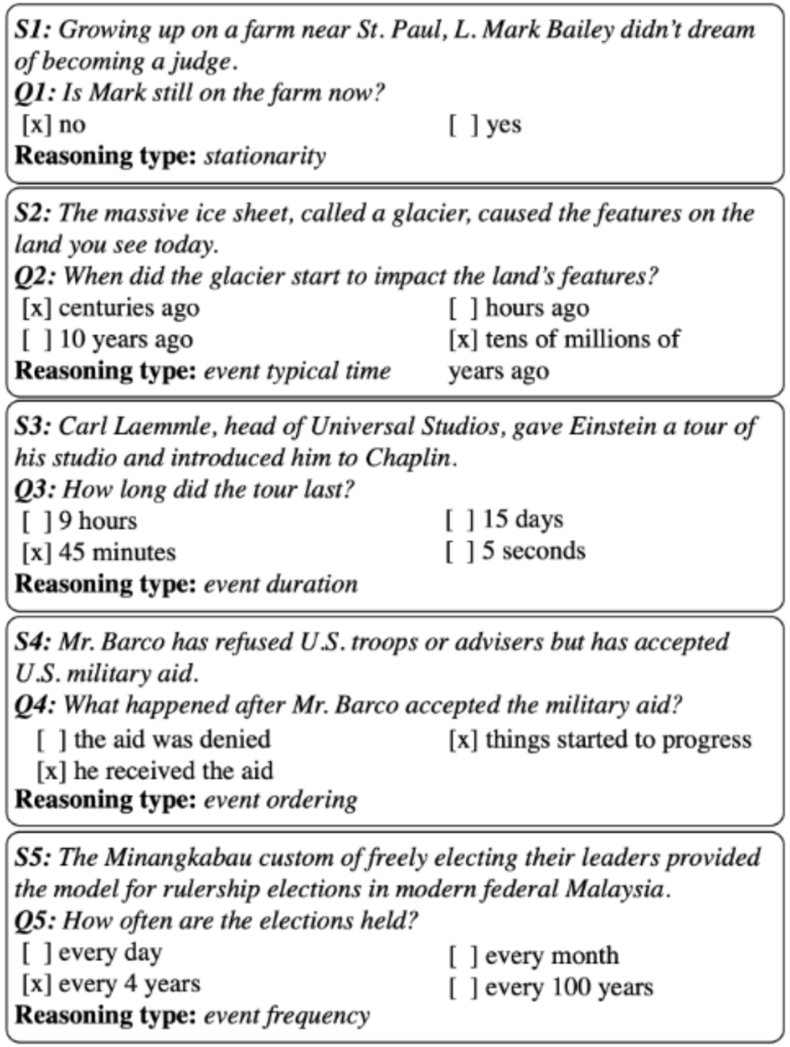
Stationarity

Typical Time

Duration

Temporal Ordering

Event Frequency



***S1:** Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.*
***Q1:** Is Mark still on the farm now?*
[x] no　　[ ] yes
Reasoning type: *stationarity*

***S2:** The massive ice sheet, called a glacier, caused the features on the land you see today.*
***Q2:** When did the glacier start to impact the land's features?*
[x] centuries ago　　[ ] hours ago
[ ] 10 years ago　　[x] tens of millions of years ago
Reasoning type: *event typical time*

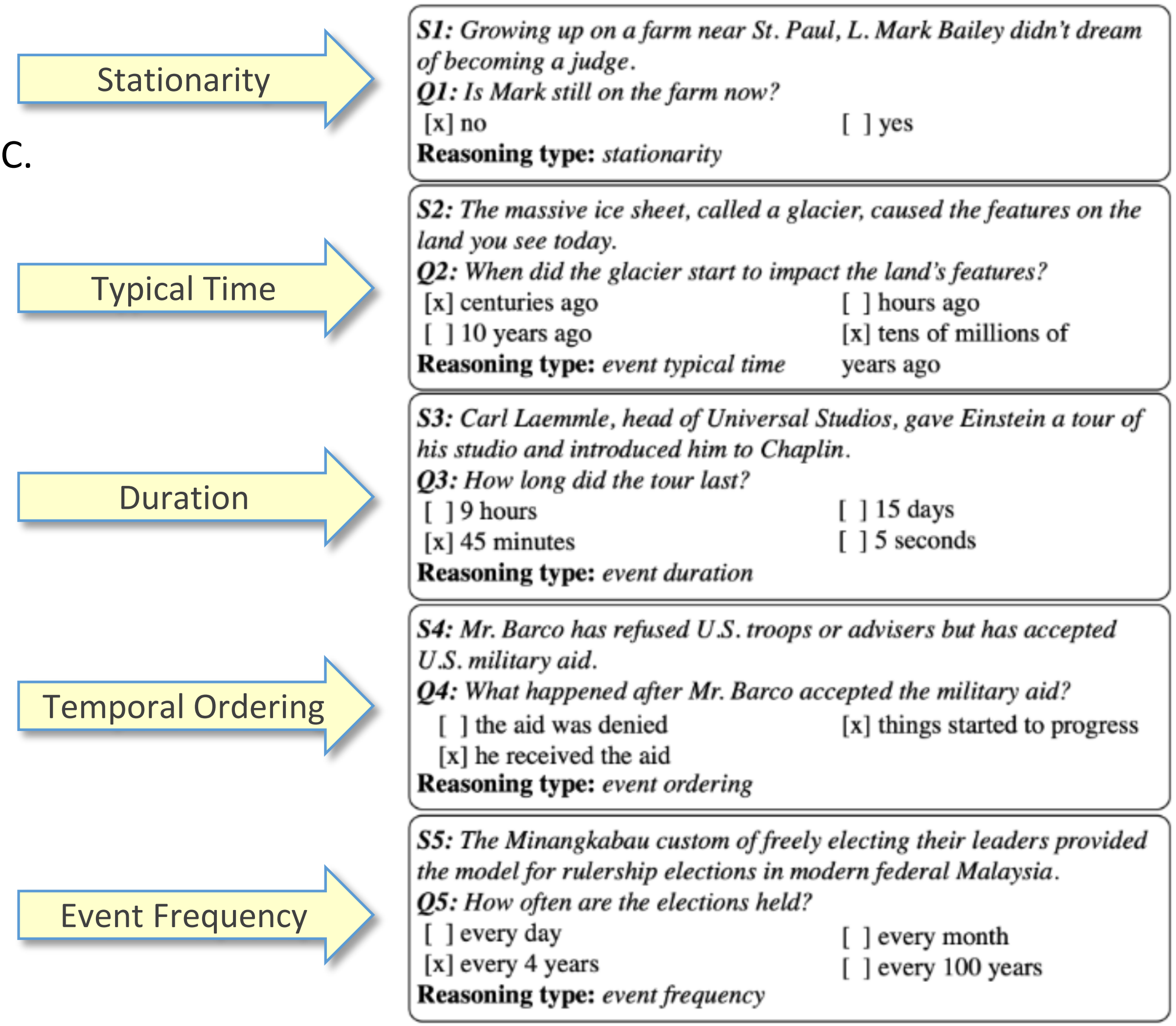
***S3:** Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.*
***Q3:** How long did the tour last?*
[ ] 9 hours　　[ ] 15 days
[x] 45 minutes　　[ ] 5 seconds
Reasoning type: *event duration*

***S4:** Mr. Barco has refused U.S. troops or advisers but has accepted U.S. military aid.*
***Q4:** What happened after Mr. Barco accepted the military aid?*
[ ] the aid was denied　　[x] things started to progress
[x] he received the aid
Reasoning type: *event ordering*

***S5:** The Minangkabau custom of freely electing their leaders provided the model for rulership elections in modern federal Malaysia.*
***Q5:** How often are the elections held?*
[ ] every day　　[ ] every month
[x] every 4 years　　[ ] every 100 years
Reasoning type: *event frequency*

MC-TACO🌮: A Temporal Commonsense Dataset [Zhou et al. EMNLP'19]

- Stationarity:
  - Paul Simon is in NYC. Let's go see him.
  - The Empire State Building is in NYC.



Stationarity

Typical Time

Duration

Temporal Ordering

Event Frequency

***S1:** Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.*
***Q1:** Is Mark still on the farm now?*
[x] no [ ] yes
Reasoning type: *stationarity*

***S2:** The massive ice sheet, called a glacier, caused the features on the land you see today.*
***Q2:** When did the glacier start to impact the land's features?*
[x] centuries ago [ ] hours ago
[ ] 10 years ago [x] tens of millions of years ago
Reasoning type: *event typical time*

***S3:** Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.*
***Q3:** How long did the tour last?*
[ ] 9 hours [ ] 15 days
[x] 45 minutes [ ] 5 seconds
Reasoning type: *event duration*

***S4:** Mr. Barco has refused U.S. troops or advisers but has accepted U.S. military aid.*
***Q4:** What happened after Mr. Barco accepted the military aid?*
[ ] the aid was denied [x] things started to progress
[x] he received the aid
Reasoning type: *event ordering*

***S5:** The Minangkabau custom of freely electing their leaders provided the model for rulership elections in modern federal Malaysia.*
***Q5:** How often are the elections held?*
[ ] every day [ ] every month
[x] every 4 years [ ] every 100 years
Reasoning type: *event frequency*

The results of a RoBERTa-based models are **very low**. Not surprising given the need to have **commonsense** to address these challenges.

Perhaps more importantly, it illustrates the need to **decompose**, and know how to **incorporate knowledge**.

MC-TACO🌮: A Temporal Commonsense Dataset [Zhou et al. EMNLP'19]

- Stationarity:
  - Paul Simon is in NYC. Let's go see him.
  - The Empire State Building is in NYC.

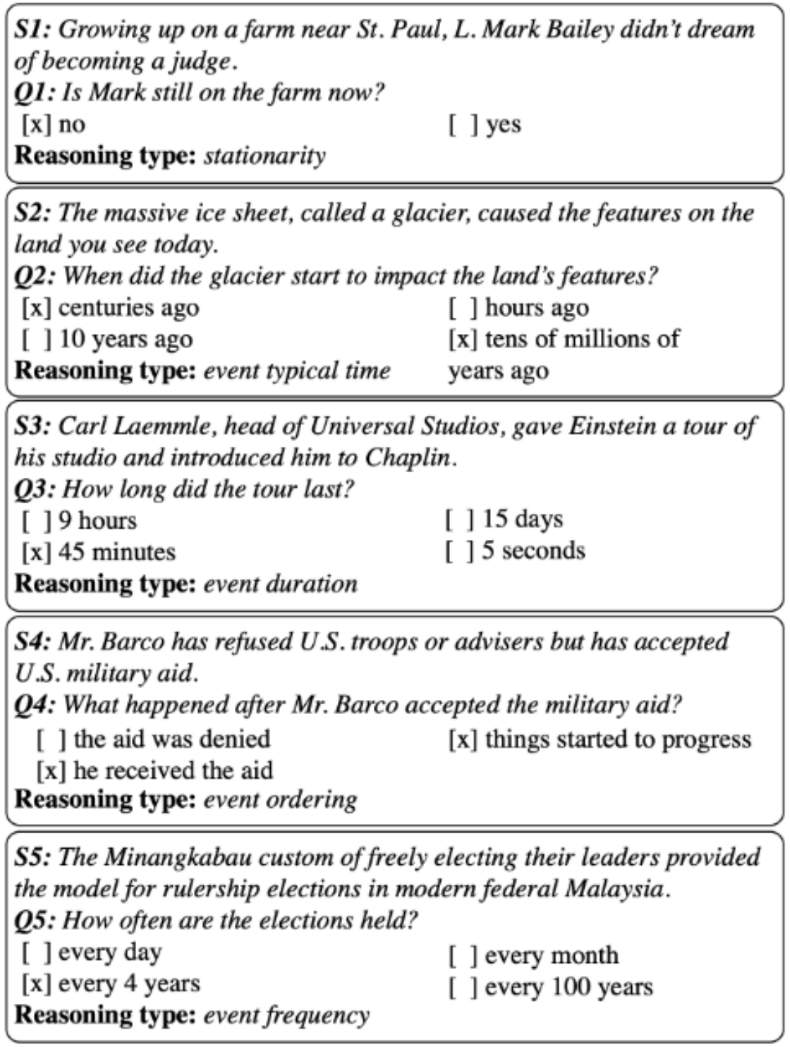
Stationarity

Typical Time

Duration

Temporal Ordering

Event Frequency



***S1:** Growing up on a farm near St. Paul, L. Mark Bailey didn't dream of becoming a judge.*
***Q1:** Is Mark still on the farm now?*
[x] no [ ] yes
Reasoning type: *stationarity*

***S2:** The massive ice sheet, called a glacier, caused the features on the land you see today.*
***Q2:** When did the glacier start to impact the land's features?*
[x] centuries ago [ ] hours ago
[ ] 10 years ago [x] tens of millions of years ago
Reasoning type: *event typical time*

***S3:** Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin.*
***Q3:** How long did the tour last?*
[ ] 9 hours [ ] 15 days
[x] 45 minutes [ ] 5 seconds
Reasoning type: *event duration*

***S4:** Mr. Barco has refused U.S. troops or advisers but has accepted U.S. military aid.*
***Q4:** What happened after Mr. Barco accepted the military aid?*
[ ] the aid was denied [x] things started to progress
[x] he received the aid
Reasoning type: *event ordering*

***S5:** The Minangkabau custom of freely electing their leaders provided the model for rulership elections in modern federal Malaysia.*
***Q5:** How often are the elections held?*
[ ] every day [ ] every month
[x] every 4 years [ ] every 100 years
Reasoning type: *event frequency*

The results of a RoBERTa-based models are **very low**. Not surprising given the need to have **commonsense** to address these challenges.

Perhaps more importantly, it illustrates the need to **decompose**, and know how to **incorporate knowledge**.

Will we make it to dinner before the movie?

TemporAl COmmonsense LM

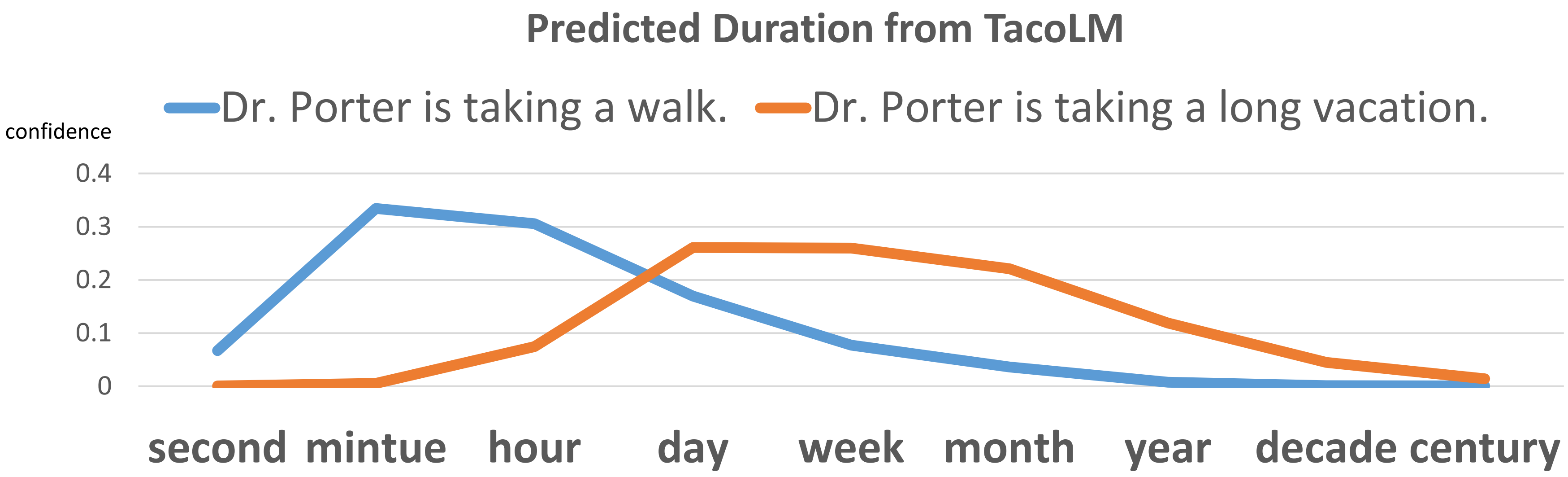
- TacoLM – A general LM that is aware of time and temporal common sense
  - □ Minimal Supervision
- Used to develop contextual estimation for Duration, Typical Time and Duration
  - □ Time is represented as a distribution over time units

TemporAI COmmonsense LM

- TacoLM – A general LM that is aware of time and temporal common sense
  - □ Minimal Supervision
- Used to develop contextual estimation for Duration, Typical Time and Duration
  - □ Time is represented as a distribution over time units

TemporAl COmmonsense LM

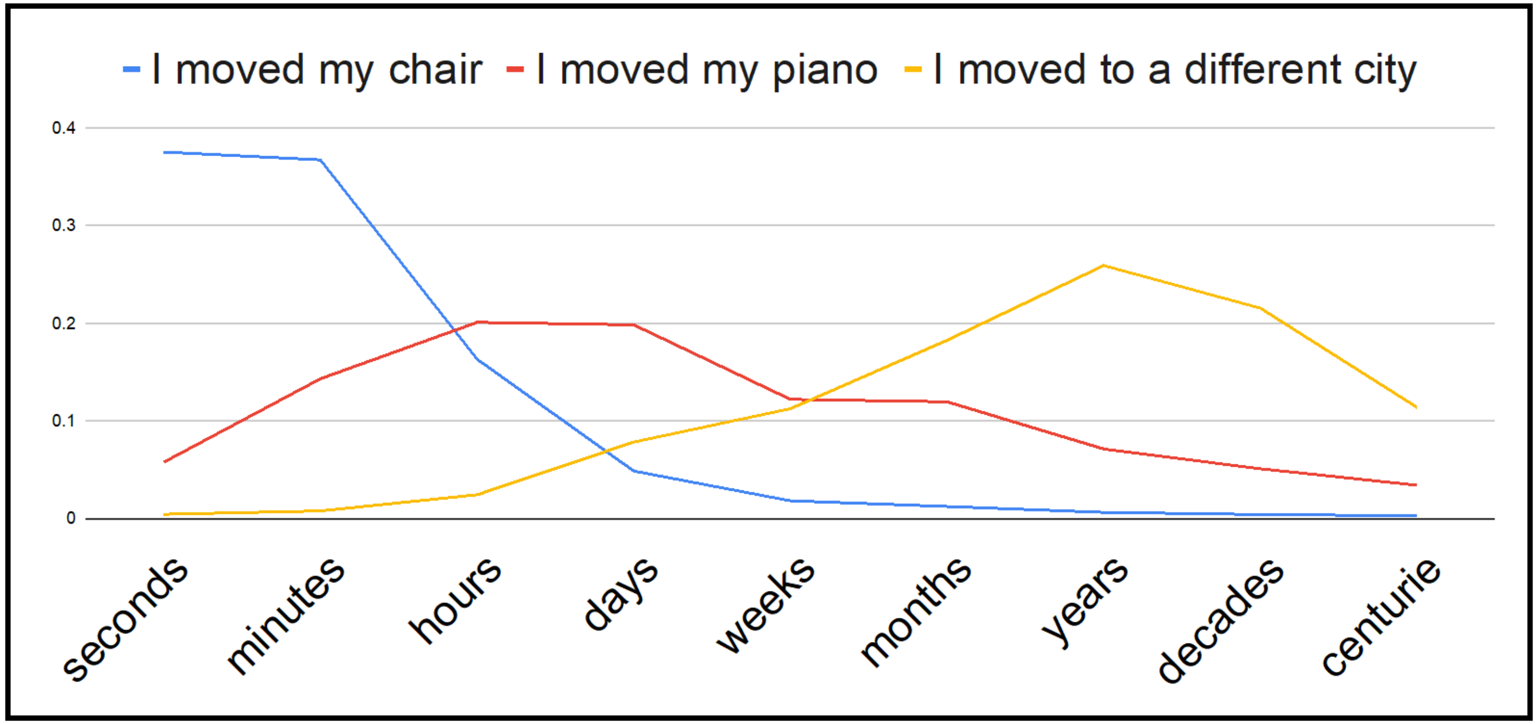
- TacoLM – A general LM that is aware of time and temporal common sense
  - □ Minimal Supervision
- Used to develop contextual estimation for Duration, Typical Time and Duration
  - □ Time is represented as a distribution over time units



Modeling Temporal Common Sense

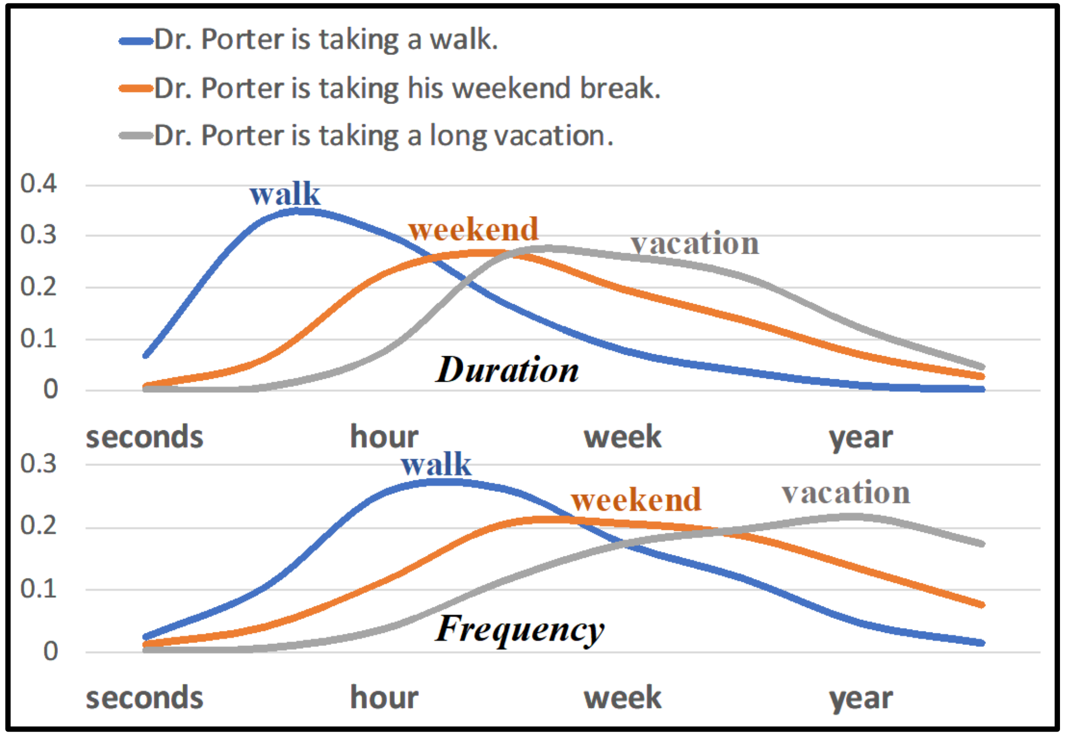
- Context
  - □ How long does “move” take?
    - Highly contextual: Move a chair? Move a piano?
    - Needs more than direct event arguments

- Joint Modeling
  - □ Do people often write how long they brushed their teeth in text?
    - But they’ll say: I brushed my teeth in the morning; I brushed it in the shower
  - □ (Partly) addresses reporting bias



Technical Highlights

- Unsupervised collection of auxiliary signals
  - □ Using patterns from free text
  - □ Extract complete events – predicate and arguments



- Joint model across interrelated dimensions
  - □ Assume no signal on the duration of “brushing teeth”, we can still get upper bounds from “brush teeth in the morning” or “brush teeth every day” or “brush teeth during shower”
  - □ Natural constraints: duration <= 1/frequency

TacoLM – the Big Picture

Goal: build a general time-aware LM with minimal supervision

TacoLM – the Big Picture

Goal: build a general time-aware LM with minimal supervision

Step 2: Joint Masked Language Model

TacoLM – the Big Picture

Goal: build a general time-aware LM with minimal supervision

Step 1: Information Extraction

- □ Using high-precision patterns to acquire temporal information
  - ■ Unsupervised automatic extraction
- □ Overcomes reporting biases with a large amount of natural text

Step 2: Joint Masked Language Model

TacoLM – the Big Picture

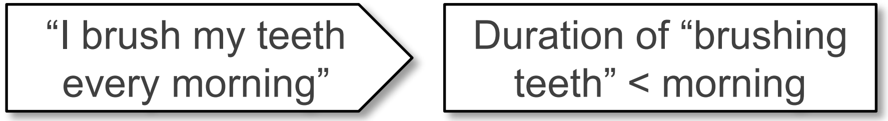
Goal: build a general time-aware LM with minimal supervision

Step 1: Information Extraction

- Using high-precision patterns to acquire temporal information
  - Unsupervised automatic extraction
- Overcomes reporting biases with a large amount of natural text

Step 2: Joint Masked Language Model

- Multiple temporal dimensions
  - Duration ~ 1 / Frequency



“I brush my teeth every morning”

Duration of “brushing teeth” < morning

  - Further generalization to combat reporting biases

TacoLM – the Big Picture

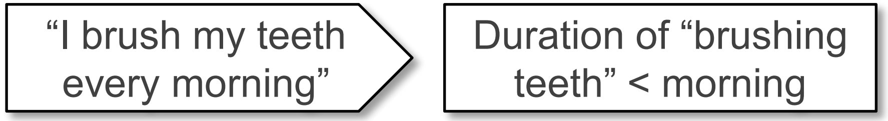
Goal: build a general time-aware LM with minimal supervision

Step 1: Information Extraction

- □ Using high-precision patterns to acquire temporal information
  - ■ Unsupervised automatic extraction
- □ Overcomes reporting biases with a large amount of natural text

Step 2: Joint Masked Language Model

- □ Multiple temporal dimensions
  - ■ Duration ~ 1 / Frequency



"I brush my teeth every morning"

Duration of "brushing teeth" < morning

  - ■ Further generalization to combat reporting biases

Output: TacoLM- a time-aware general BERT

Information Extraction

- Use high-precision patterns based on SRL
  - Duration
  - Frequency
  - Typical Time
  - Duration Upper bound
  - Hierarchy

Information Extraction

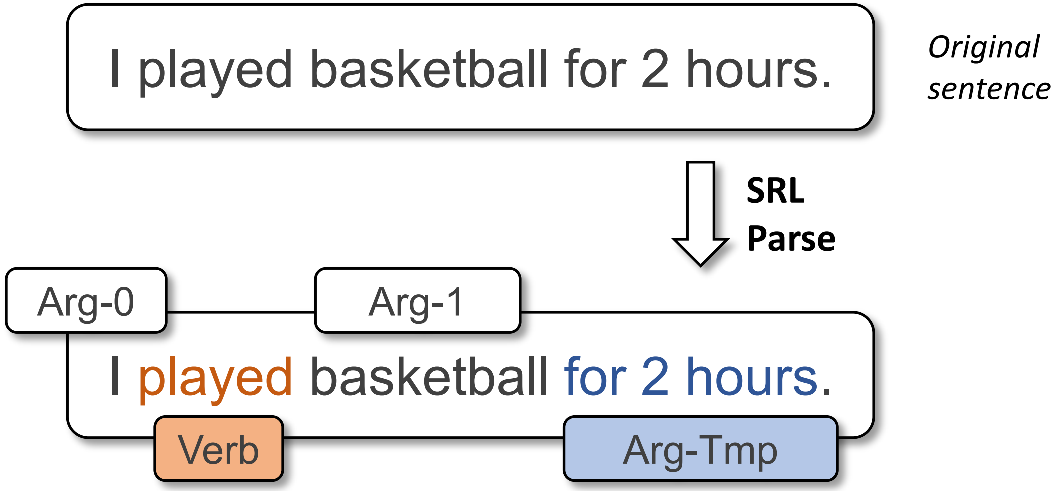
- Use high-precision patterns based on SRL
  - Duration
  - Frequency
  - Typical Time
  - Duration Upper bound
  - Hierarchy

I played basketball for 2 hours.

Original sentence

Information Extraction

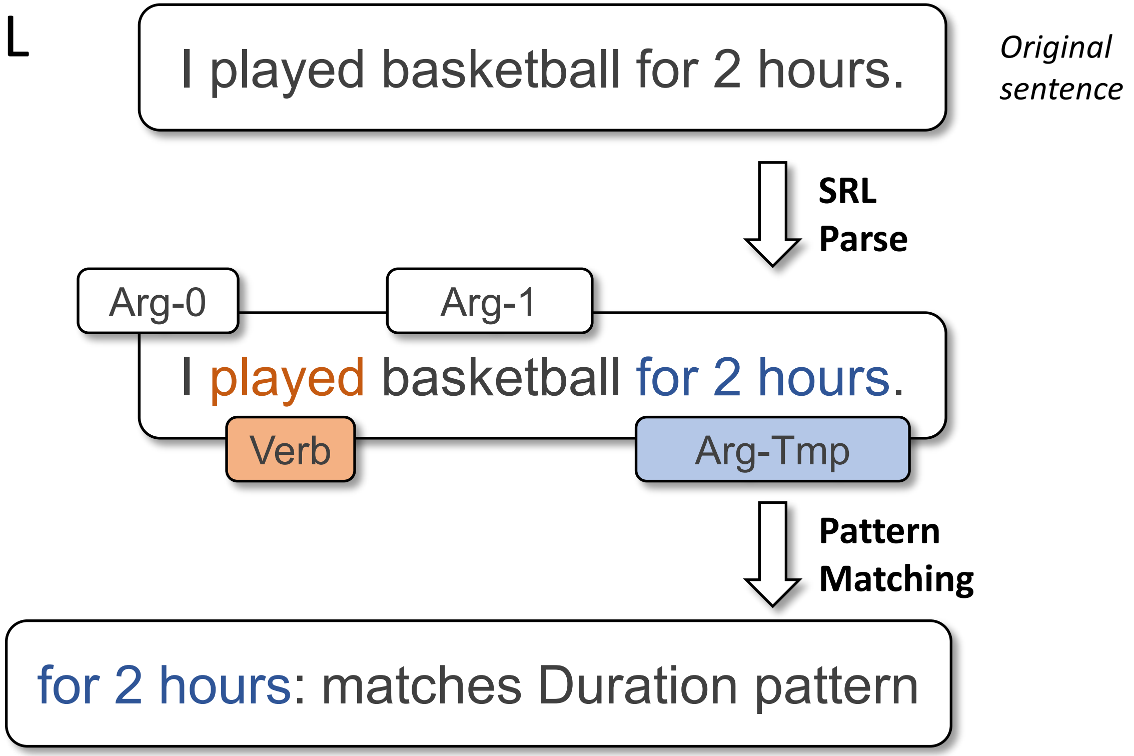
- Use high-precision patterns based on SRL
  - □ Duration
  - □ Frequency
  - □ Typical Time
  - □ Duration Upper bound
  - □ Hierarchy



I played basketball for 2 hours.
Original sentence
SRL Parse
Arg-0
Arg-1
I played basketball for 2 hours.
Verb
Arg-Tmp

Information Extraction

- Use high-precision patterns based on SRL
  - □ Duration
  - □ Frequency
  - □ Typical Time
  - □ Duration Upper bound
  - □ Hierarchy



I played basketball for 2 hours.

Original sentence

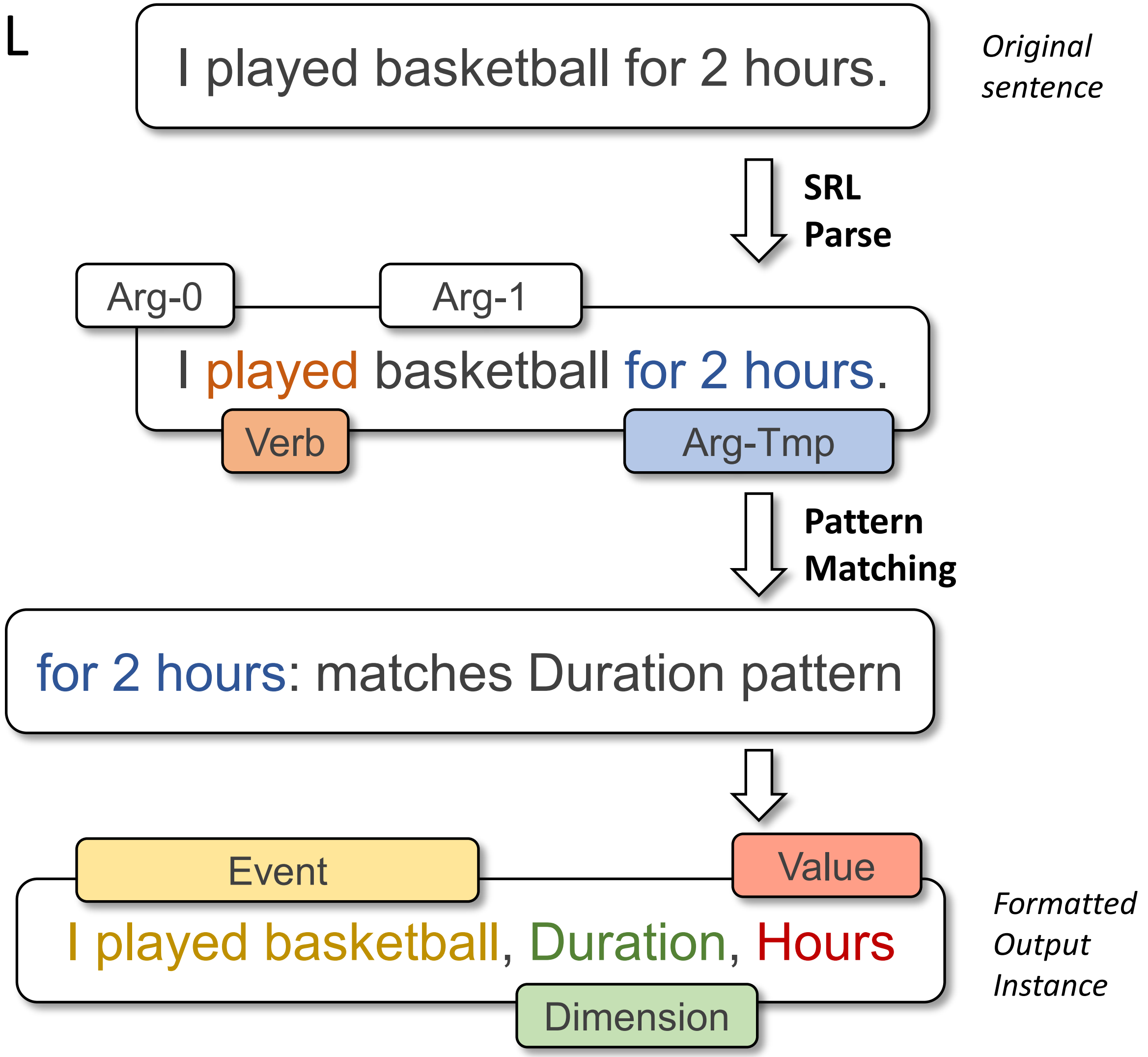
SRL Parse

Pattern Matching

for 2 hours: matches Duration pattern

Information Extraction

- Use high-precision patterns based on SRL
  - □ Duration
  - □ Frequency
  - □ Typical Time
  - □ Duration Upper bound
  - □ Hierarchy



I played basketball for 2 hours.

Original sentence

SRL Parse

Arg-0 Arg-1

I played basketball for 2 hours.

Verb Arg-Tmp

Pattern Matching

for 2 hours: matches Duration pattern

Event Value

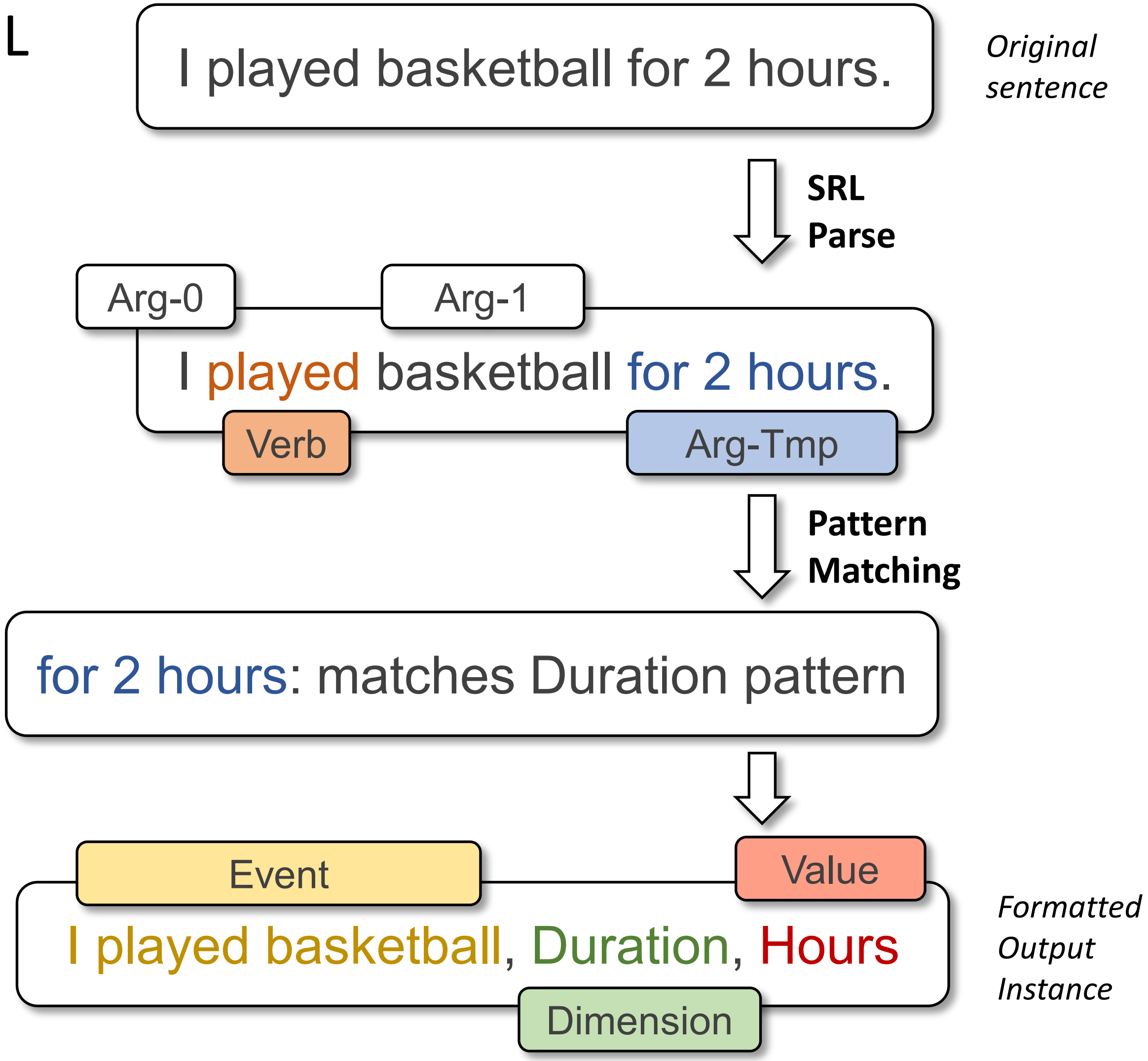
I played basketball, Duration, Hours

Dimension

Formatted Output Instance

Information Extraction

- Use high-precision patterns based on SRL
  - □ Duration
  - □ Frequency
  - □ Typical Time
  - □ Duration Upper bound
  - □ Hierarchy
- Labels
  - □ Units (seconds, … centuries)
  - □ Temporal keywords (Monday, January, …)
- Output
  - □ 4.3M instances of (event, dimension, value) tuple

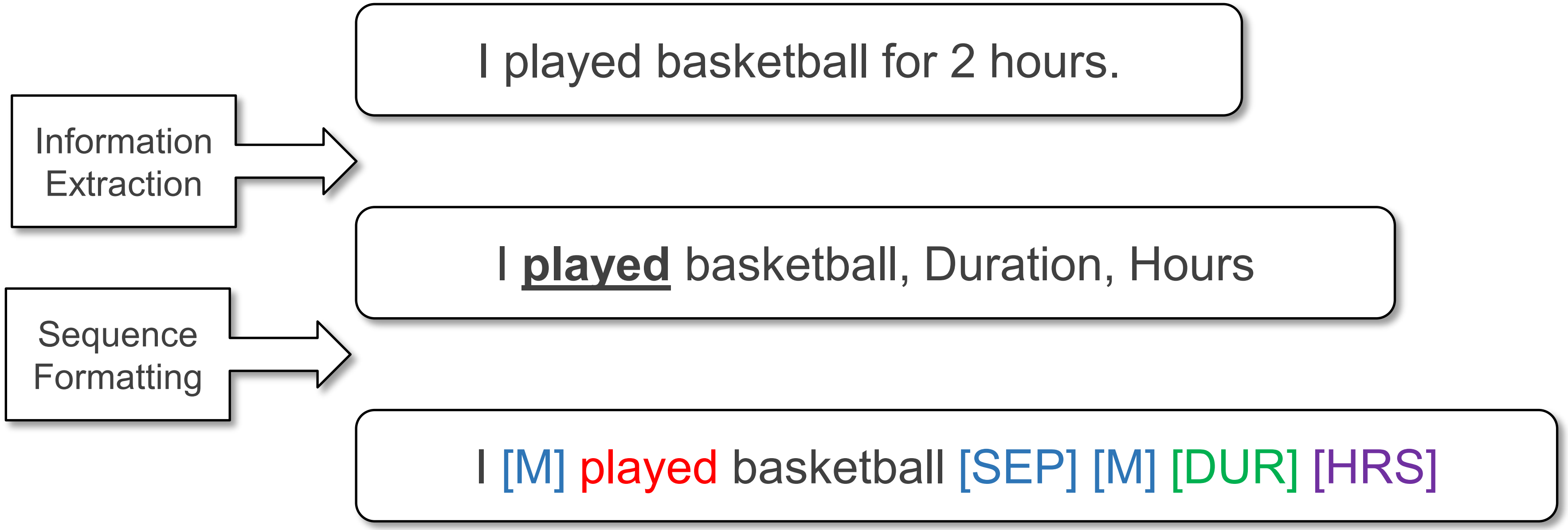

I played basketball for 2 hours.
Original sentence
SRL Parse
Arg-0
Arg-1
I played basketball for 2 hours.
Verb
Arg-Tmp
Pattern Matching
for 2 hours: matches Duration pattern
Event
Value
I played basketball, Duration, Hours
Dimension
Formatted Output Instance

Sequence Formatting

- Consider [Event] [Dimension] [Value] tuples in each instance
- [E1, E2, … M, ET … En, SEP, M, Dim, Val]
  - □ M is a special marker, same across all dimension/value
  - □ Dim is a marker for each dimension, Val is a marker for the value of the dimension

Sequence Formatting

- Consider [Event] [Dimension] [Value] tuples in each instance
- [E1, E2, … M, ET … En, SEP, M, Dim, Val]
  - □ M is a special marker, same across all dimension/value
  - □ Dim is a marker for each dimension, Val is a marker for the value of the dimension
- An example:



I played basketball for 2 hours.

Information Extraction

I **played** basketball, Duration, Hours

Sequence Formatting

I [M] played basketball [SEP] [M] [DUR] [HRS]

Joint Model with Masked LM

I [M] played basketball [SEP] [M] [DUR] [HRS]

- Baseline Model: Pre-trained BERT-base
- Main objective: mask some tokens and recover them

Joint Model with Masked LM

I [M] played basketball [SEP] [M] [DUR] [HRS]

- Baseline Model: Pre-trained BERT-base
- Main objective: mask some tokens and recover them
- How we mask:
  - □ With some probability, mask temporal value while keeping others

    I [M] played basketball [SEP] [M] [DUR] **[MASK]**
  - □ Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged

    I [M] **[MASK] [MASK]** [SEP] [M] [DUR] **[HRS]**
  - □ Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability

Joint Model with Masked LM

I [M] played basketball [SEP] [M] [DUR] [HRS]

- Baseline Model: Pre-trained BERT-base
- Main objective: mask some tokens and recover them
- How we mask:
  - □ With some probability, mask temporal value while keeping others

    I [M] played basketball [SEP] [M] [DUR] **[MASK]**

  - □ Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged

    I [M] **[MASK] [MASK]** [SEP] [M] [DUR] **[HRS]**

  - □ Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - □ Jointly learns **one** transformer for **all** dimensions
  - □ Labels play a role in the transformer
  - □ One event may contain more than one (Dim + Val), so the model learns dimension relationships

Joint Model with Masked LM

I [M] played basketball [SEP] [M] [DUR] [HRS]

- 1: Soft cross entropy for recovering Val
  - □ If gold label is “hours”, the label vector **y** for “minutes, hours, days” will be [0.16, 0.47, 0.25]

$$\hat{\mathbf{x}} = \log(\mathrm{softmax}(\mathbf{x}))$$

$$\mathrm{loss} = -\hat{\mathbf{x}}^{\top}\mathbf{y}$$

- 2: Label weight adjustment
  - □ Instances with “seconds” have higher loss than those with “years”

Joint Model with Masked LM

I [M] played basketball [SEP] [M] [DUR] [HRS]

- 1: Soft cross entropy for recovering Val
  - □ If gold label is “hours”, the label vector **y** for “minutes, hours, days” will be [0.16, 0.47, 0.25]

$$\hat{\mathbf{x}} = \log(\text{softmax}(\mathbf{x}))$$

$$\text{loss} = -\hat{\mathbf{x}}^\top \mathbf{y}$$

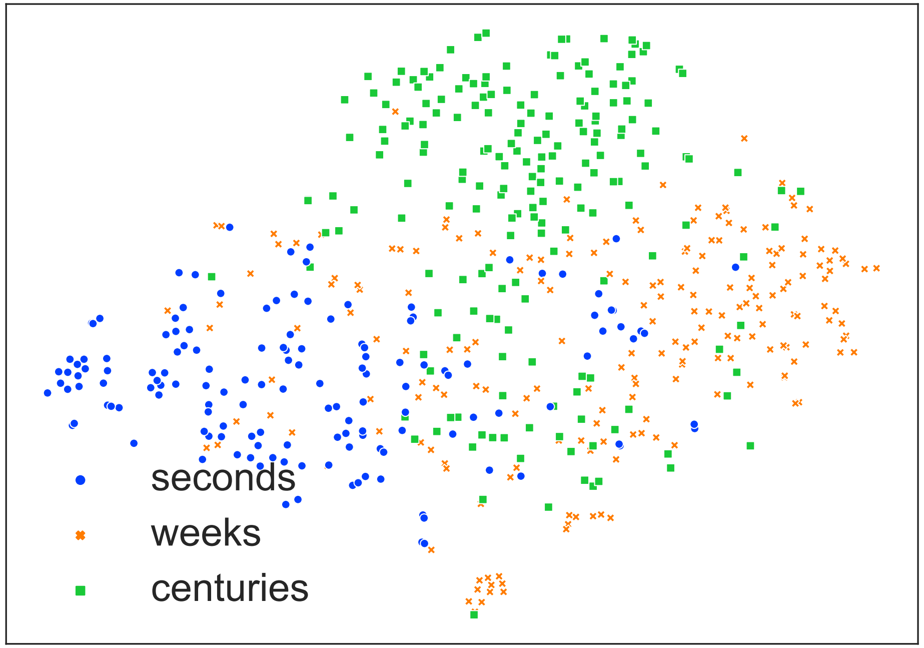
- 2: Label weight adjustment
  - □ Instances with “seconds” have higher loss than those with “years”
- 3: Full event masking
  - □ Instead of 15% used by BERT, we use 60% when masking E1, … En to reduce biases

I [M] had **a cup of** **[MASK]** [SEP] [M] [TYP] [Evening] -> MASK = coffee, because “cup of”

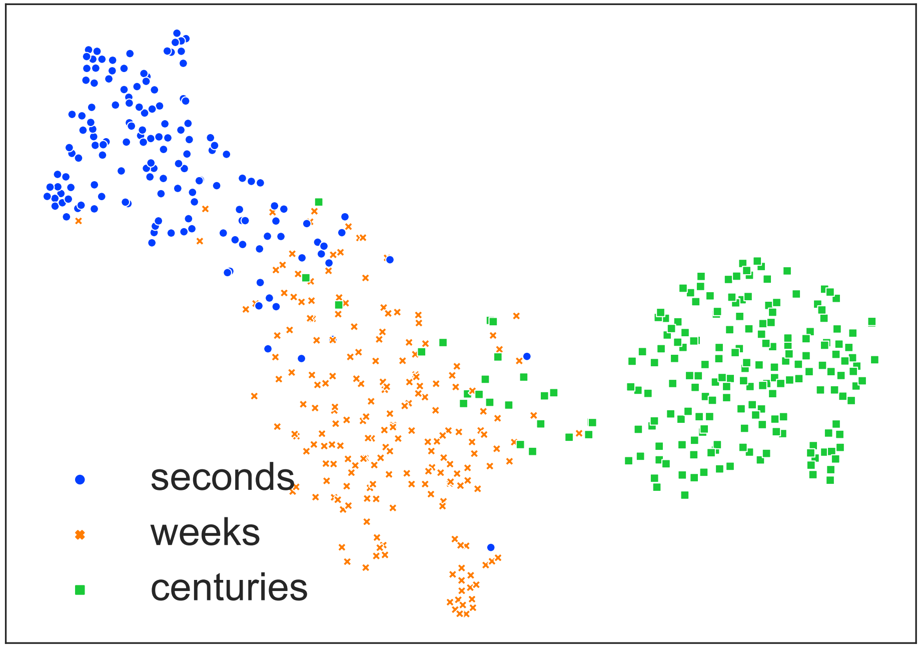
I [M] had **[MASK]** **[MASK]** of **[MASK]** [SEP] [M] [TYP] [Evening]

Evaluation: (Embedding space)

- A collection of events with duration of “seconds,” “weeks” or “centuries” (three extremes)
- BERT (left), Ours (right) representation on the event’s trigger
  - □ PCA + t-SNE to 2D visualization
- Our model separates the events much better (➔ our model is aware of time)



BERT



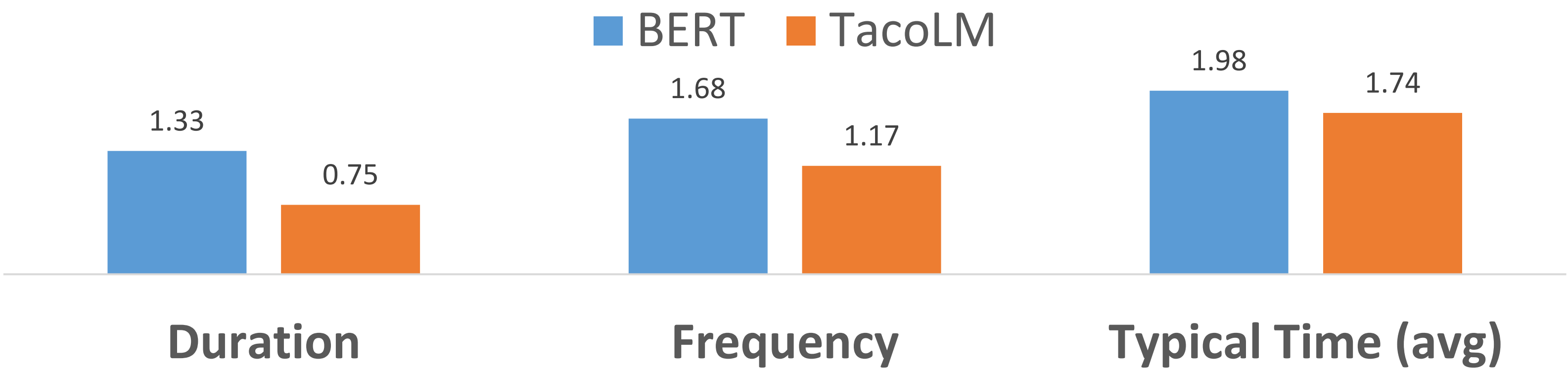
TacoLM

Quantitative Evaluation:

- Metric: Distance to gold label
  - □ Dist (seconds, hours)=2, Dist (minutes, hours)=1
  - □ **Lower the better**

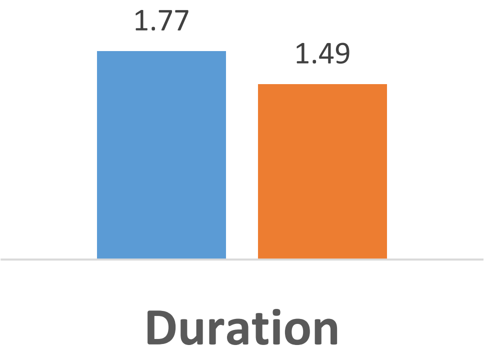
Quantitative Evaluation:

- Metric: Distance to gold label
  - □ Dist (seconds, hours)=2, Dist (minutes, hours)=1
  - □ **Lower the better**
- RealNews [Zellers et al. 2019]: no document overlap
  - □ Raw corpus + MTurk annotation



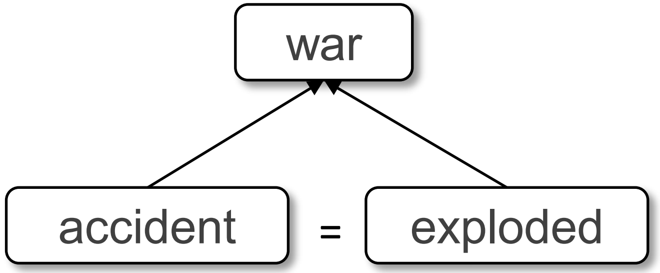
19% average improvement

- UDS-T [Vashishtha et al. 2019]: duration only



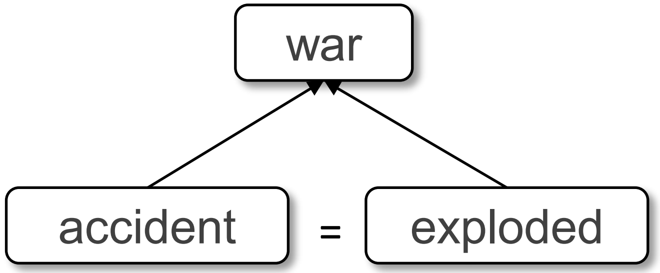
Evaluation: Event-Event Relations

- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
  - □ HiEVE [Glavas et al. 2014]
  - □ Child-Parent / Parent-Child / Coreference
    - A bomb exploded. This is the sixth accident since the war started.



Evaluation: Event-Event Relations

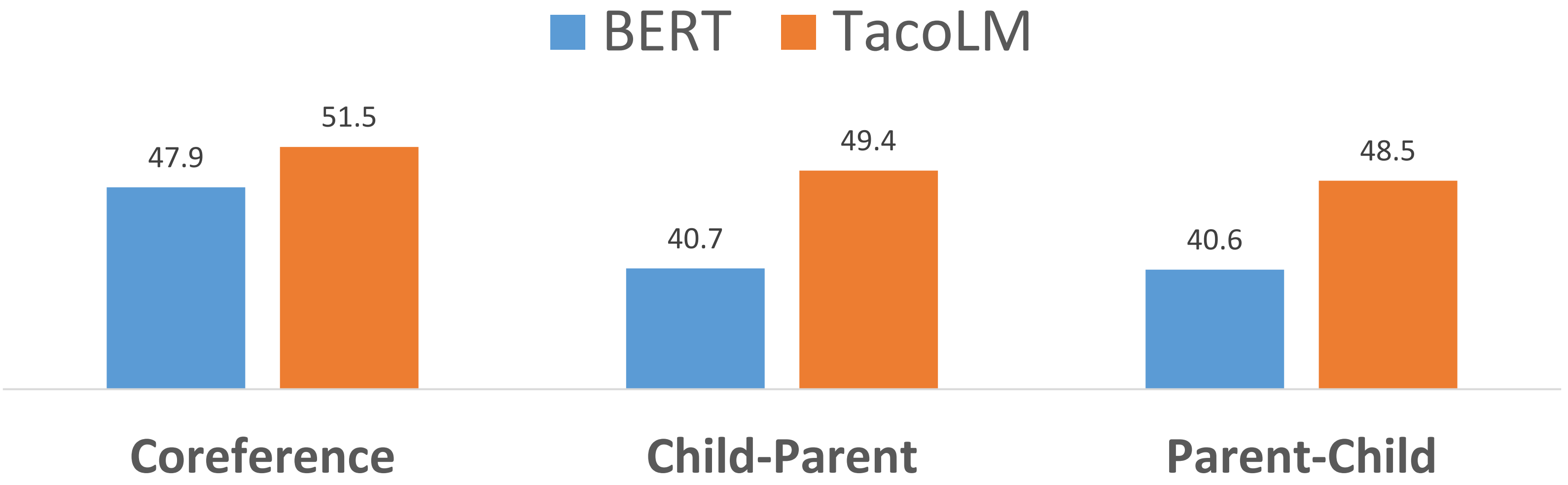
- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
    - HiEVE [Glavas et al. 2014]
    - Child-Parent / Parent-Child / Coreference
        - A bomb exploded. This is the sixth accident since the war started.



war

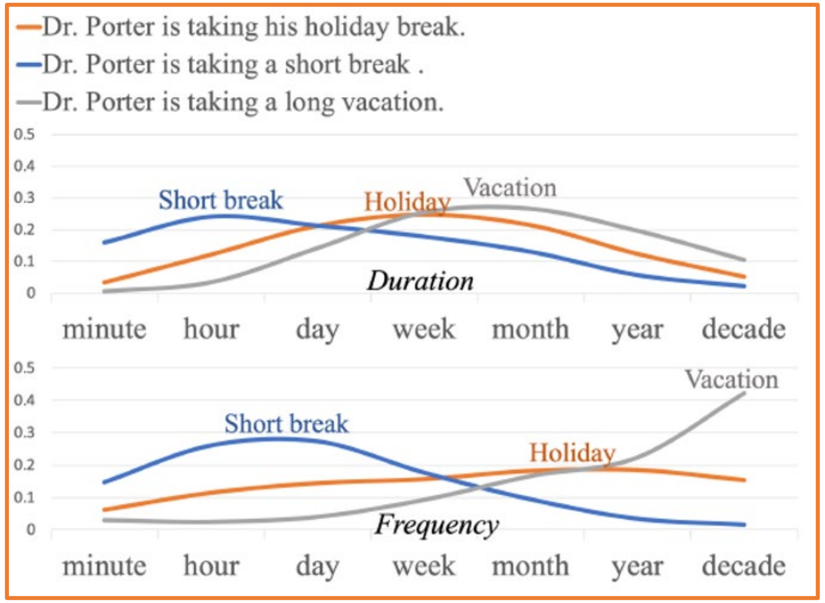
accident = exploded

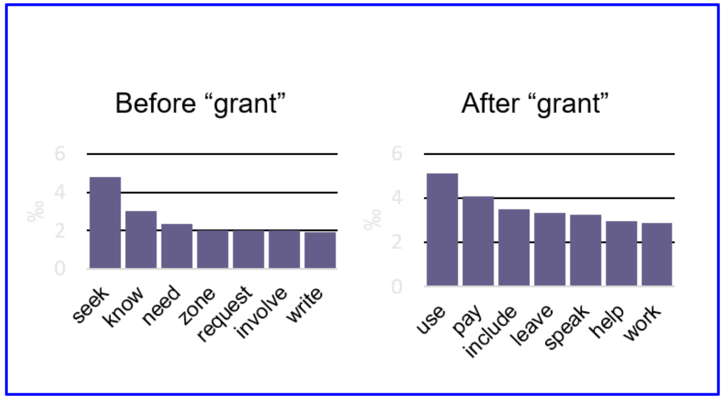
- Model (finetuned):
    - Sentence pair classification
- Results (F1, **higher is better**)



Conclusion – Temporal Commonsense

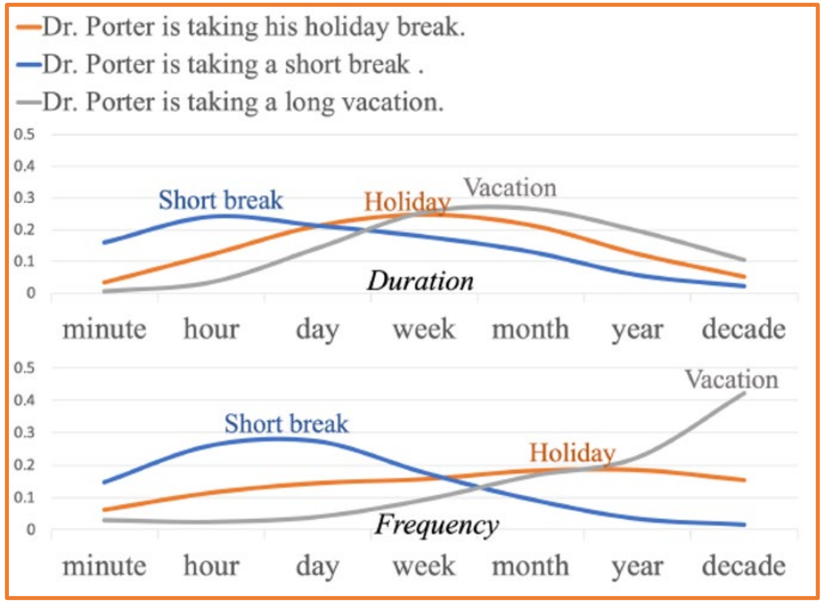
- A range of natural language understanding tasks require that we "understand" time
  - □ And many of these require that we have commonsense
    - E.g., time is transitive; how long things take; typical order, etc.
- Time is interesting for many reasons
  - □ In particular, since natural language rarely provides explicit temporal information
    - How long does it take to open a window?
    - What "things" change with time (and what do not)?
  - □ In most cases – temporal knowledge is distributional

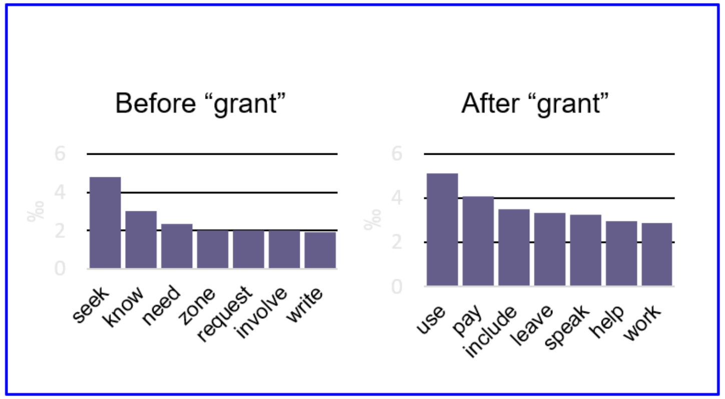




Conclusion – Temporal Commonsense

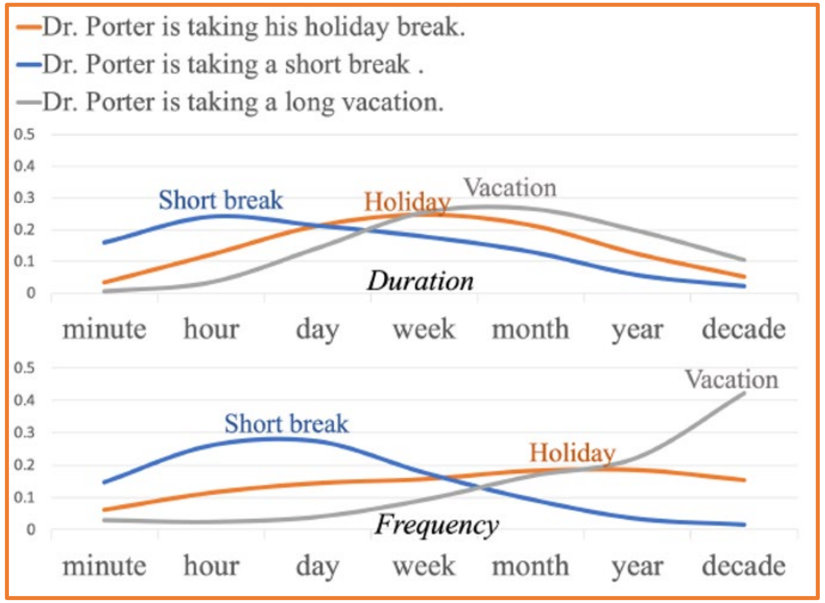
- A range of natural language understanding tasks require that we “understand” time
  - □ And many of these require that we have commonsense
    - E.g., time is transitive; how long things take; typical order, etc.
- Time is interesting for many reasons
  - □ In particular, since natural language rarely provides explicit temporal information
    - How long does it take to open a window?
    - What “things” change with time (and what do not)?
  - □ In most cases – temporal knowledge is distributional

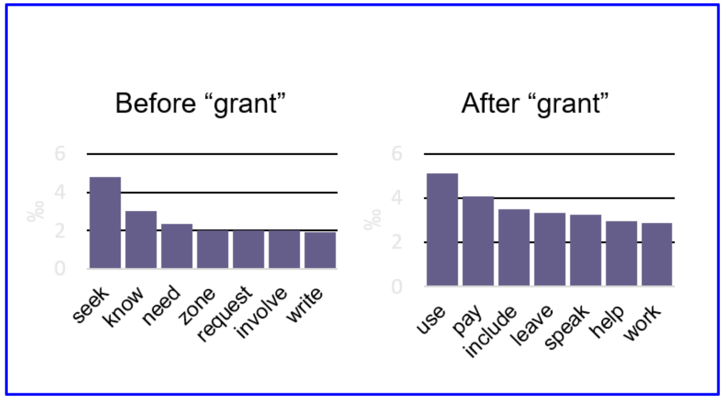




Conclusion – Temporal Commonsense

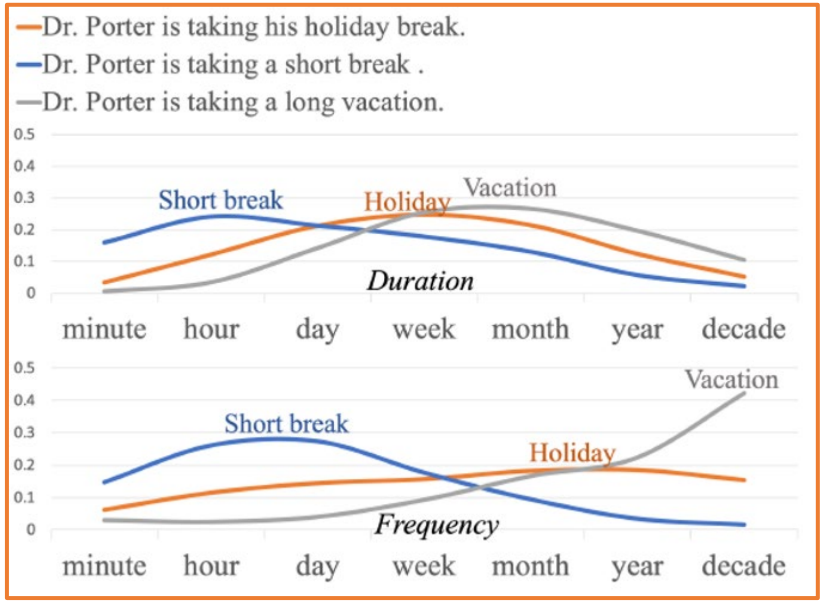
- A range of natural language understanding tasks require that we "understand" time
  - □ And many of these require that we have commonsense
    - E.g., time is transitive; how long things take; typical order, etc.
- Time is interesting for many reasons
  - □ In particular, since natural language rarely provides explicit temporal information
    - How long does it take to open a window?
    - What "things" change with time (and what do not)?
  - □ In most cases – temporal knowledge is distributional

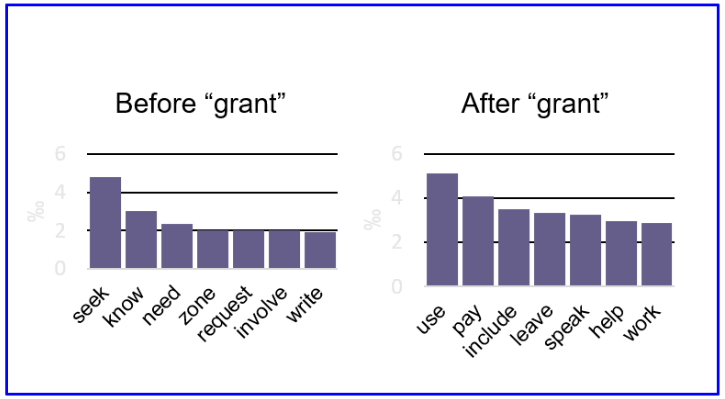




Conclusion – Temporal Commonsense

- A range of natural language understanding tasks require that we “understand” time
  - □ And many of these require that we have commonsense
    - E.g., time is transitive; how long things take; typical order, etc.
- Time is interesting for many reasons
  - □ In particular, since natural language rarely provides explicit temporal information
    - How long does it take to open a window?
    - What “things” change with time (and what do not)?
  - □ In most cases – temporal knowledge is distributional
- Presented MC-TACO data set
  - □ A challenge QA dataset for temporal commonsense
- Discussed TACO-LM
  - □ A time-aware Contextual Language Model
  - □ Duration, typical time, frequency
- There is a lot more to do!





Tutorial Conclusion

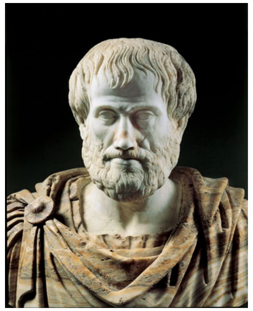
- Ways to acquire, represent and distill commonsense knowledge
  - □ Along multiple dimensions: Physical, Social, Temporal
  - □ Some require crowdsourcing, some can be extracted directly from te
- Ways to integrate it into models
  - □ The CoMET paradigm; ERNIE-style integration; Temporally-aware contextual LM
- Ways to measure commonsense abilities
  - □ Extending commonsense probes
  - □ Creating robust benchmarks & evaluation setups

Tutorial Conclusion

- Ways to acquire, represent and distill commonsense knowledge
  - Along multiple dimensions: Physical, Social, Temporal
  - Some require crowdsourcing, some can be extracted directly from te
- Ways to integrate it into models
  - The CoMET paradigm; ERNIE-style integration; Temporally-aware contextual LM
- Ways to measure commonsense abilities
  - Extending commonsense probes
  - Creating robust benchmarks & evaluation setups
- None of these is “solved”,
  - Extensions – acquisition within and across dimensions
  - Multi-modal commonsense knowledge acquisition
  - Commonsense “reasoning”
  - ….

Tutorial Conclusion

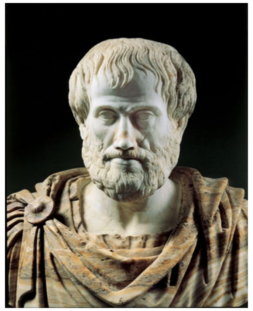
- Ways to acquire, represent and distill commonsense knowledge
  - Along multiple dimensions: Physical, Social, Temporal
  - Some require crowdsourcing, some can be extracted directly from te
- Ways to integrate it into models
  - The CoMET paradigm; ERNIE-style integration; Temporally-aware contextual LM
- Ways to measure commonsense abilities
  - Extending commonsense probes
  - Creating robust benchmarks & evaluation setups
- None of these is “solved”,
  - Extensions – acquisition within and across dimensions
  - Multi-modal commonsense knowledge acquisition
  - Commonsense “reasoning”
  - ….



Tutorial Conclusion

- Ways to acquire, represent and distill commonsense knowledge
  - Along multiple dimensions: Physical, Social, Temporal
  - Some require crowdsourcing, some can be extracted directly from te
- Ways to integrate it into models
  - The CoMET paradigm; ERNIE-style integration; Temporally-aware contextual LM
- Ways to measure commonsense abilities
  - Extending commonsense probes
  - Creating robust benchmarks & evaluation setups
- None of these is “solved”,
  - Extensions – acquisition within and across dimensions
  - Multi-modal commonsense knowledge acquisition
  - Commonsense “reasoning”
  - ….

So, did Aristotle have a laptop?



Tutorial Conclusion

Thank you!

- Ways to acquire, represent and distill commonsense knowledge
  - □ Along multiple dimensions: Physical, Social, Temporal
  - □ Some require crowdsourcing, some can be extracted directly from te
- Ways to integrate it into models
  - □ The CoMET paradigm; ERNIE-style integration; Temporally-aware contextual LM
- Ways to measure commonsense abilities
  - □ Extending commonsense probes
  - □ Creating robust benchmarks & evaluation setups
- None of these is “solved”,
  - □ Extensions – acquisition within and across dimensions
  - □ Multi-modal commonsense knowledge acquisition
  - □ Commonsense “reasoning”
  - □ ….

So, did Aristotle have a laptop?