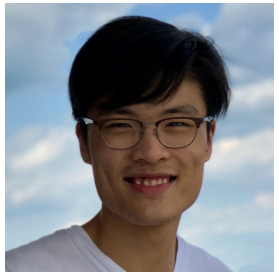
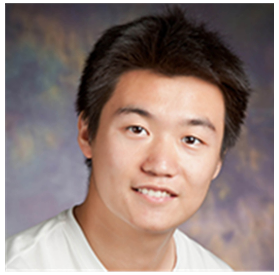
Penn
COGNITIVE COMPUTATION GROUP

Temporal Common Sense Acquisition with Minimal Supervision

Ben Zhou   Qiang Ning*   Daniel Khashabi*   Dan Roth

* Ai2

Time and Common Sense

- Choose from "*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.

Time and Common Sense

- Choose from "*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.

Time and Common Sense

- Choose from "*will*" or "*will not*"

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

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

Time and Common Sense

- Choose from "*will*" or "*will not*"

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

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

Time and Common Sense

- Choose from “*will*” or “*will not*”

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

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

Time and Common Sense

- Choose from "*will*" or "*will not*"

Time:
- An important component for reading comprehension
- Commonsense-level understanding is required

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

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

This work

This work

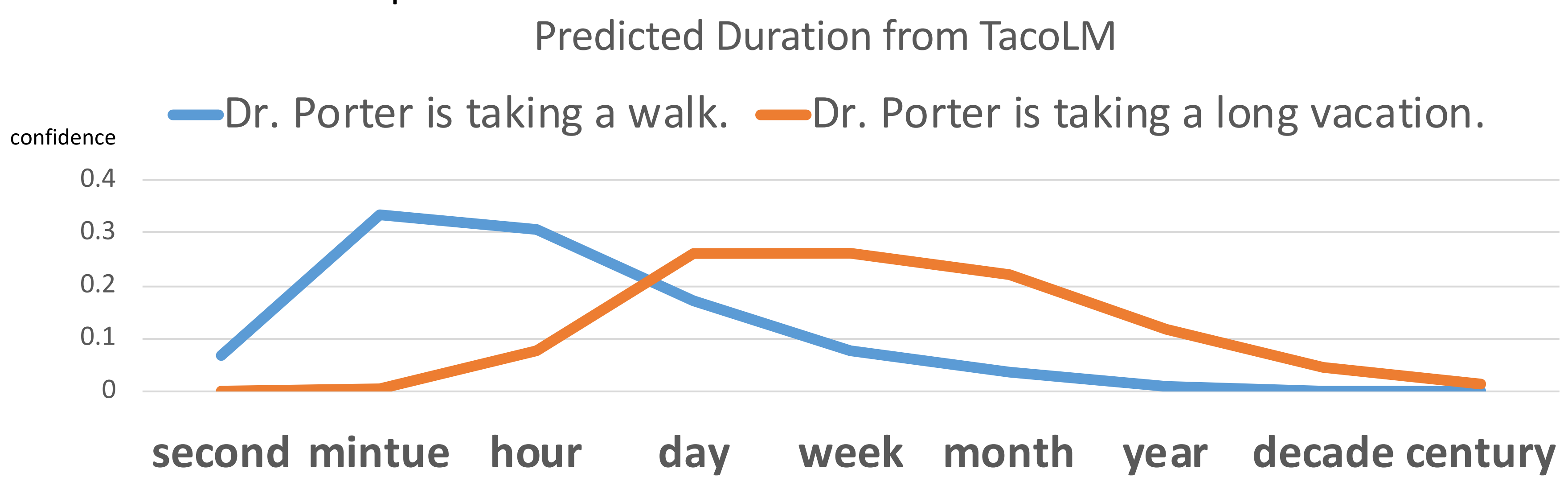
- Time
  - An important component for reading comprehension
  - Commonsense-level understanding is required
- In this work
  - TacoLM – A general LM that is aware of time and temporal common sense
    - Minimal Supervision

This work

- Time
  - An important component for reading comprehension
  - Commonsense-level understanding is required
- In this work
  - TacoLM – A general LM that is aware of time and temporal common sense
    - Minimal Supervision

This work

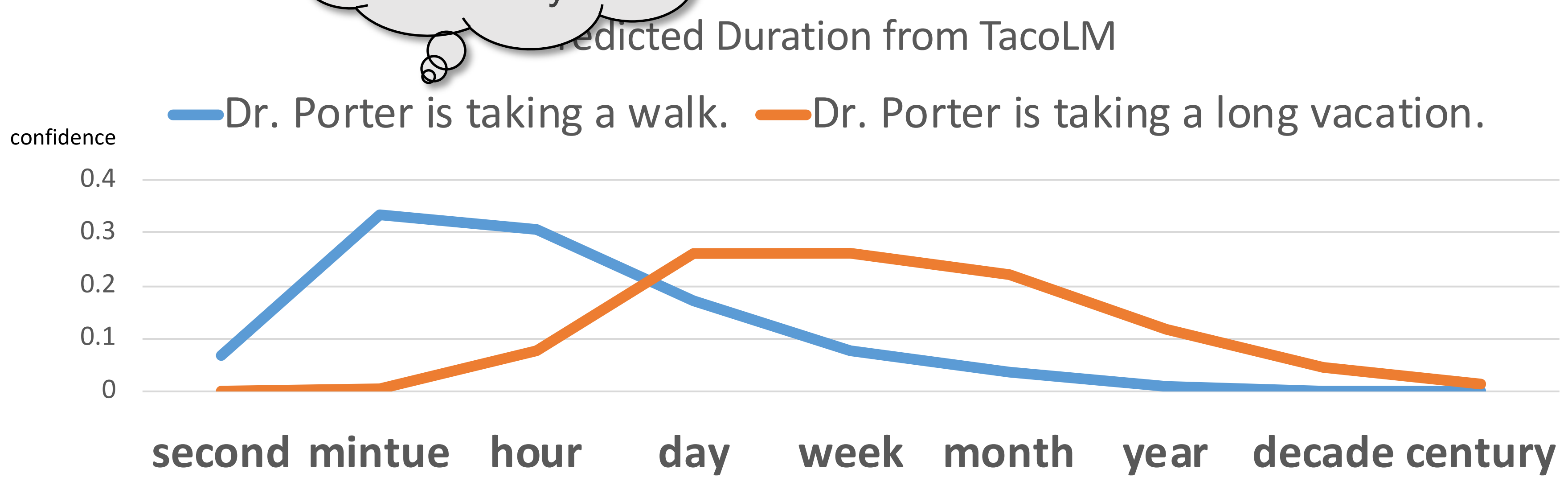
- Time
  - An important component for reading comprehension
  - Commonsense-level understanding is required
- In this work
  - TacoLM – A general LM that is aware of time and temporal common sense
    - Minimal Supervision



Time and Common Sense

- Time
  - An important component for reading comprehension
  - Commonsense-level understanding is required
- In this work
  - TacoLM – vare of time and temporal common sense
    - Minim

Dr. Porter is coming back shortly.



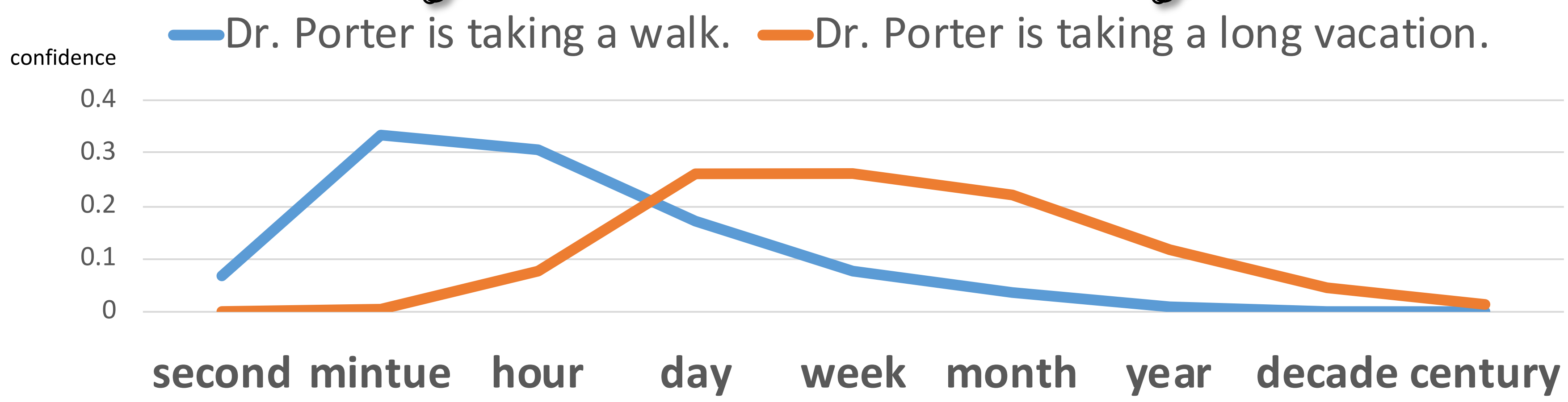
Time and Common Sense

- Time
  - An important component for reading comprehension
  - Commonsense-level understanding is required
- In this work
  - TacoLM - ... ware of time a... ense
    - Minim...

Dr. Porter is coming back shortly.

She may not be back for days.

...edicted Duration from TacoLM



Acquiring Temporal Common Sense

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient “~~*It took me 2 seconds to move my chair*~~”
    - Sometimes highlight rarities “*It took me an hour to move my chair*”
  - Highly Contextual:
    - The duration of “Move” depends on the object’s weight/size.

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "~~*It took me 2 seconds to move my chair*~~"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "~~*It took me 2 seconds to move my chair*~~"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "~~*It took me 2 seconds to move my chair*~~"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.

Acquiring Temporal Common Sense

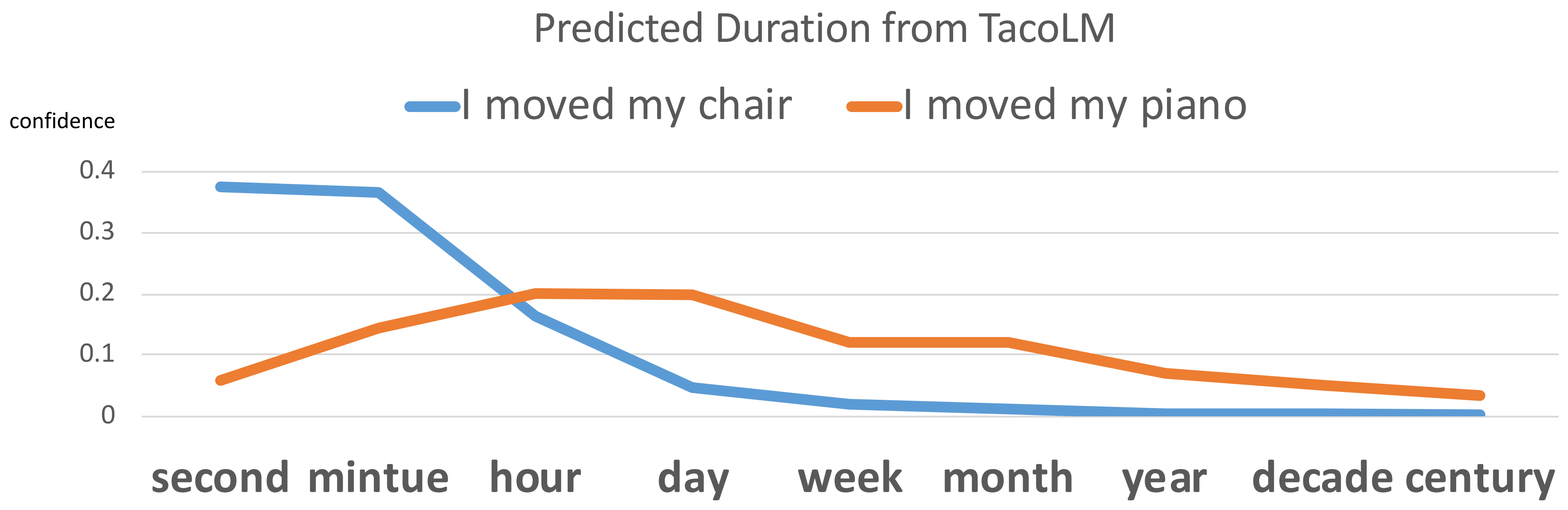
- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "~~*It took me 2 seconds to move my chair*~~"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "*~~It took me 2 seconds to move my chair~~*"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.

Acquiring Temporal Common Sense

- Challenging
  - Reporting Biases:
    - people rarely mention the common sense to be efficient "~~*It took me 2 seconds to move my chair*~~"
    - Sometimes highlight rarities "*It took me an hour to move my chair*"
  - Highly Contextual:
    - The duration of "Move" depends on the object's weight/size.



Time and Common Sense

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required
- Example: Choose from "*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from “*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from “*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from "*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from "*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from "*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required

- Example: Choose from "*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.

Time and Common Sense

- Time
  - An important component for reading comprehension
    - Temporal order
    - Event duration / frequency
    - Typical events and their occurring time
    - …
  - Explicit textual cues (before, after, at the same time) are rare
  - Commonsense-level understanding is required
- Example: Choose from "*will*" or "*will not*"
  - Dr. Porter is taking a vacation and <u>will not</u> be able to see you soon.
  - Dr. Porter is taking a walk and <u>will</u> be able to see you soon.

Temporal Common Sense

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model distributions of temporal properties of events in fine grained contexts

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments’ relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments’ relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments’ relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments’ relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

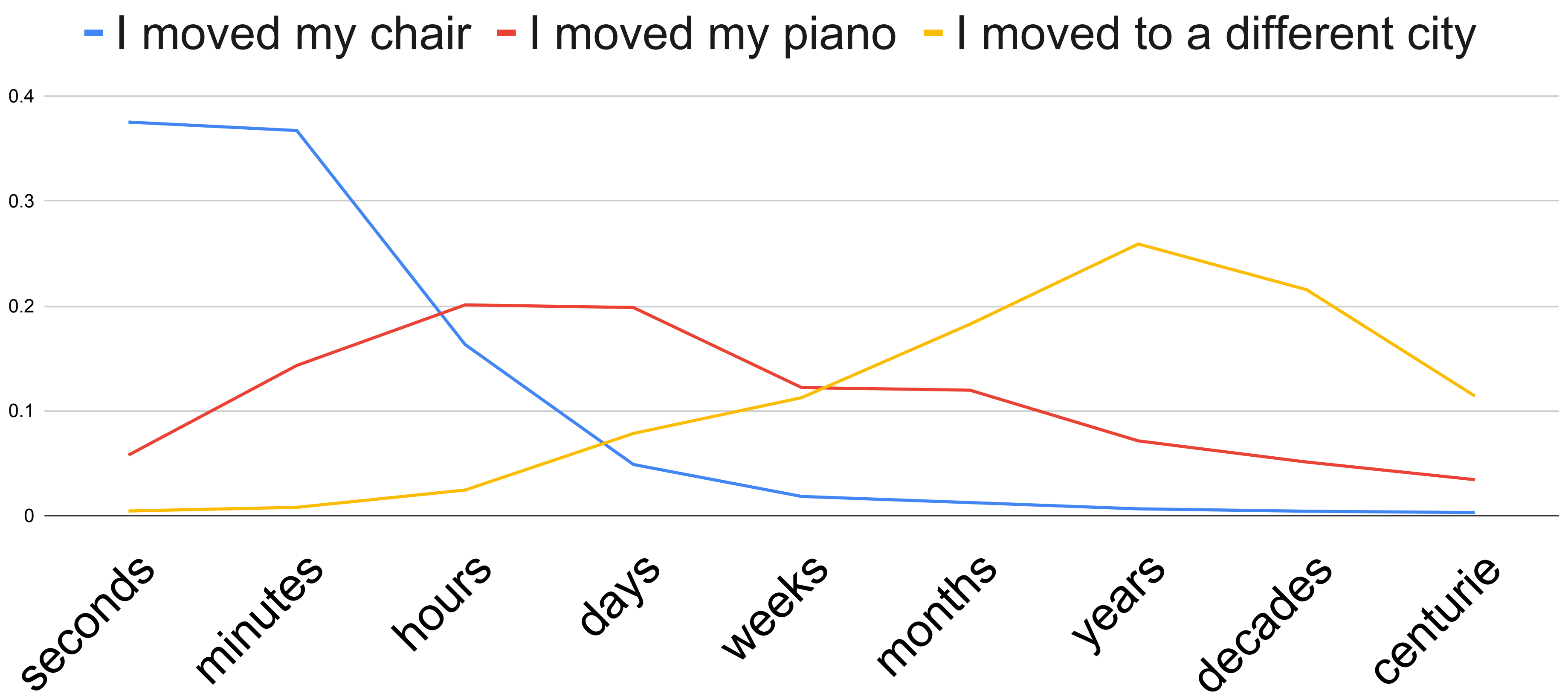
- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

Temporal Common Sense

- This work: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal Supervision
- It is challenging:
  - Highly contextual
  - Hard to understand event arguments' relation to its duration/frequency
    - Duration: I move a chair < I move a piano (weight)
    - Duration: I build a chair < I build a piano (complexity)
  - Reporting Biases
    - Rare to see people describing how long they brushed their teeth
- Our view: model <u>distributions</u> of temporal properties of events in <u>fine grained contexts</u>

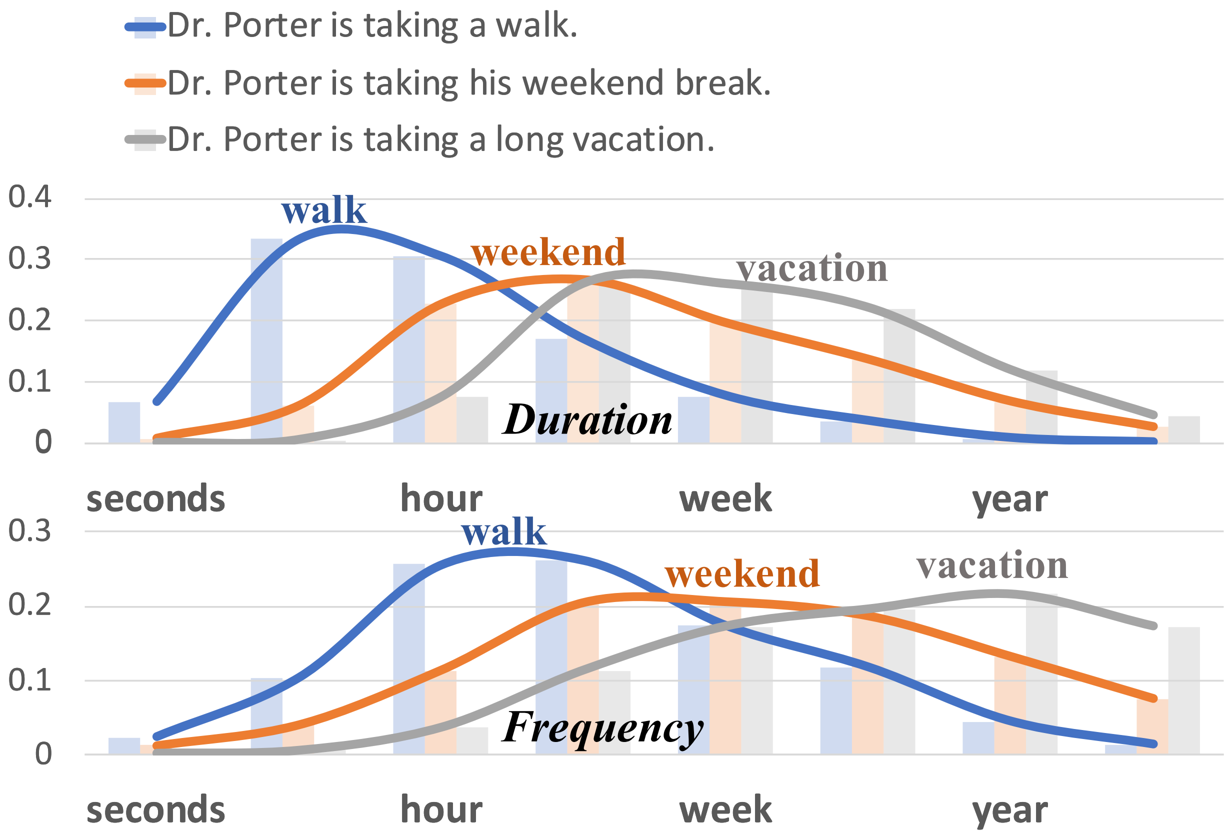
This Work

- TacoLM
  - a general time-aware language model that distincts temporal properties in fine grained contexts.



This Work

- Example: Choose from "*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.



TacoLM – the Big Picture

TacoLM – the Big Picture

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

TacoLM – the Big Picture

Step 1: Information Extraction

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 1: Information Extraction

- ☐ Use high-precision patterns to acquire temporal information
  - Unsupervised automatic extraction
- ☐ Overcomes reporting biases with a large amount of natural text
- ☐ Multiple temporal dimensions
  - Duration ~ 1 / Frequency
  - Further generalization to combat reporting biases

TacoLM – the Big Picture

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

Step 1: Information Extraction

- ☐ Use high-precision patterns to acquire temporal information
  - ■ Unsupervised automatic extraction
- ☐ Overcomes reporting biases with a large amount of natural text
- ☐ Multiple temporal dimensions
  - ■ Duration ~ 1 / Frequency
  - ■ Further generalization to combat reporting biases

TacoLM – the Big Picture

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

Step 1: Information Extraction

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

Step 2: Joint Language Model Pre-training

- ☐ Multiple temporal dimensions
  - ■ Duration ~ 1 / Frequency
  - ■ Further generalization to combat reporting biases

TacoLM – the Big Picture

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

Step 1: Information Extraction

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

Step 2: Joint Language Model Pre-training

- ☐ Multiple temporal dimensions
  - ■ Duration ~ 1 / Frequency
  - ■ Further generalization to combat reporting biases

TacoLM – the Big Picture

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

Step 1: Information Extraction

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

Step 2: Joint Language Model Pre-training

- Multiple temporal dimensions
  - Duration ~ 1 / Frequency
  - Further generalization to combat reporting biases

TacoLM – the Big Picture

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

Step 1: Information Extraction

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

Step 2: Joint Language Model Pre-training

- ☐ 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

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

Step 2: Joint Language Model Pre-training

- 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

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

Step 2: Joint Language Model Pre-training

- Multiple temporal dimensions
  - Duration ~ 1 / Frequency

“I brush my teeth every morning” → Duration of “brushing teeth” < morning

  - Further generalization to combat reporting biases

Output: <u>TacoLM</u>- a time-aware general BERT

Step 1: Information Extraction

Step 1: Information Extraction

Step 2: Joint Language Model Pre-training

Output: TacoLM- a time-aware general BERT

Joint learning from free text

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way
- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇨ Joint Pretraining ⇨ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - ☐ Unsupervised
  - ☐ Multiple Dimensions (duration, frequency, auxiliaries…)
  - ☐ Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - ☐ Use soft cross entropy that assumes a bell-shaped distribution across values
  - ☐ Also allows for circular relationships like day of weeks
  - ☐ Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇒ Joint Pretraining ⇒ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Joint learning from free text

- In general: we trained a BERT that is aware of time in a more unbiased way

Pattern Extraction ⇨ Joint Pretraining ⇨ General LM

- Pattern Extraction:
  - Unsupervised
  - Multiple Dimensions (duration, frequency, auxiliaries…)
  - Natural constraints: duration <= 1/frequency
- Joint Pretraining
  - Use soft cross entropy that assumes a bell-shaped distribution across values
  - Also allows for circular relationships like day of weeks
  - Use full event masking and label adjustment to combat reporting biases further
- General LM: with the off-the shelf capability of predicting temporal properties

Information Extraction

Information Extraction

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

Information Extraction

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

Information Extraction

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

Information Extraction

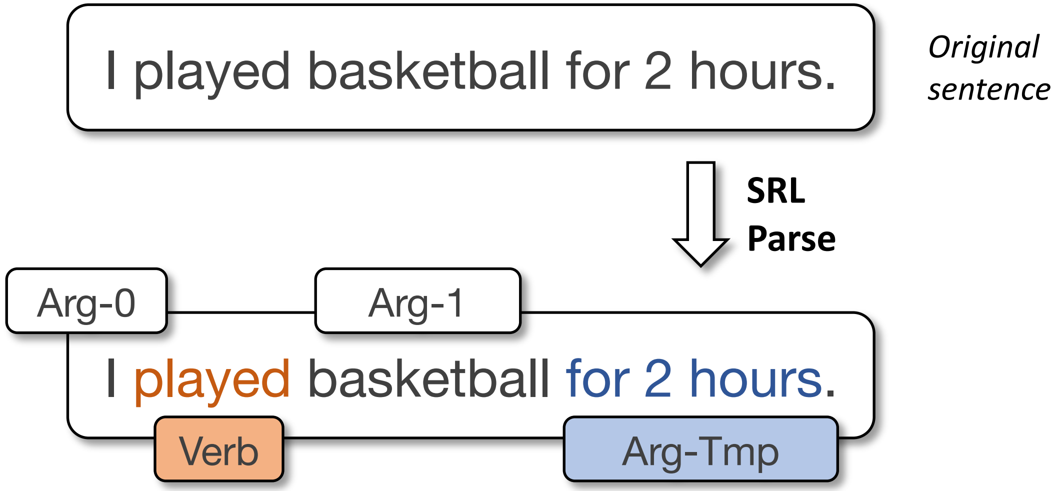
- Use high-precision patterns based on SRL
  - Duration
  - Frequency
  - Typical Time
  - Duration Upperbound
  - Hierarchy
- Labels
  - Units (seconds, … centuries)
  - Temporal keywords (Monday, January, …)
- Output
  - 4.3M instances of
    (event, dimension, value) tuple

I played basketball for 2 hours.

Original sentence

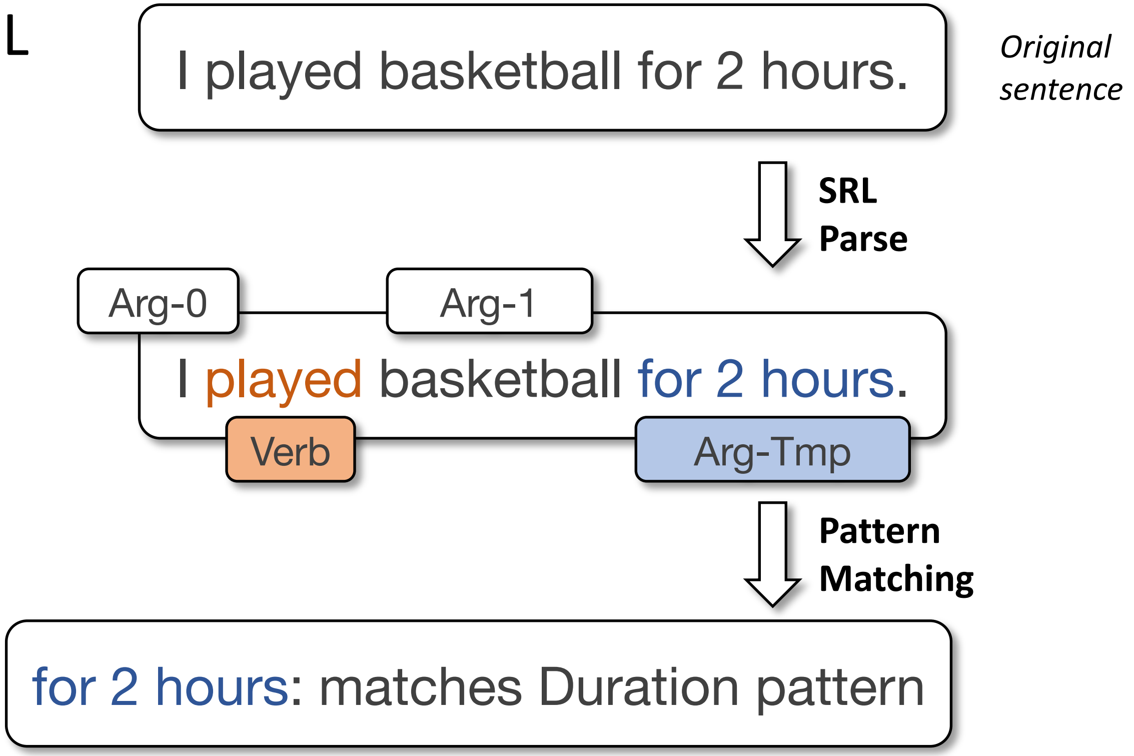
Information Extraction

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



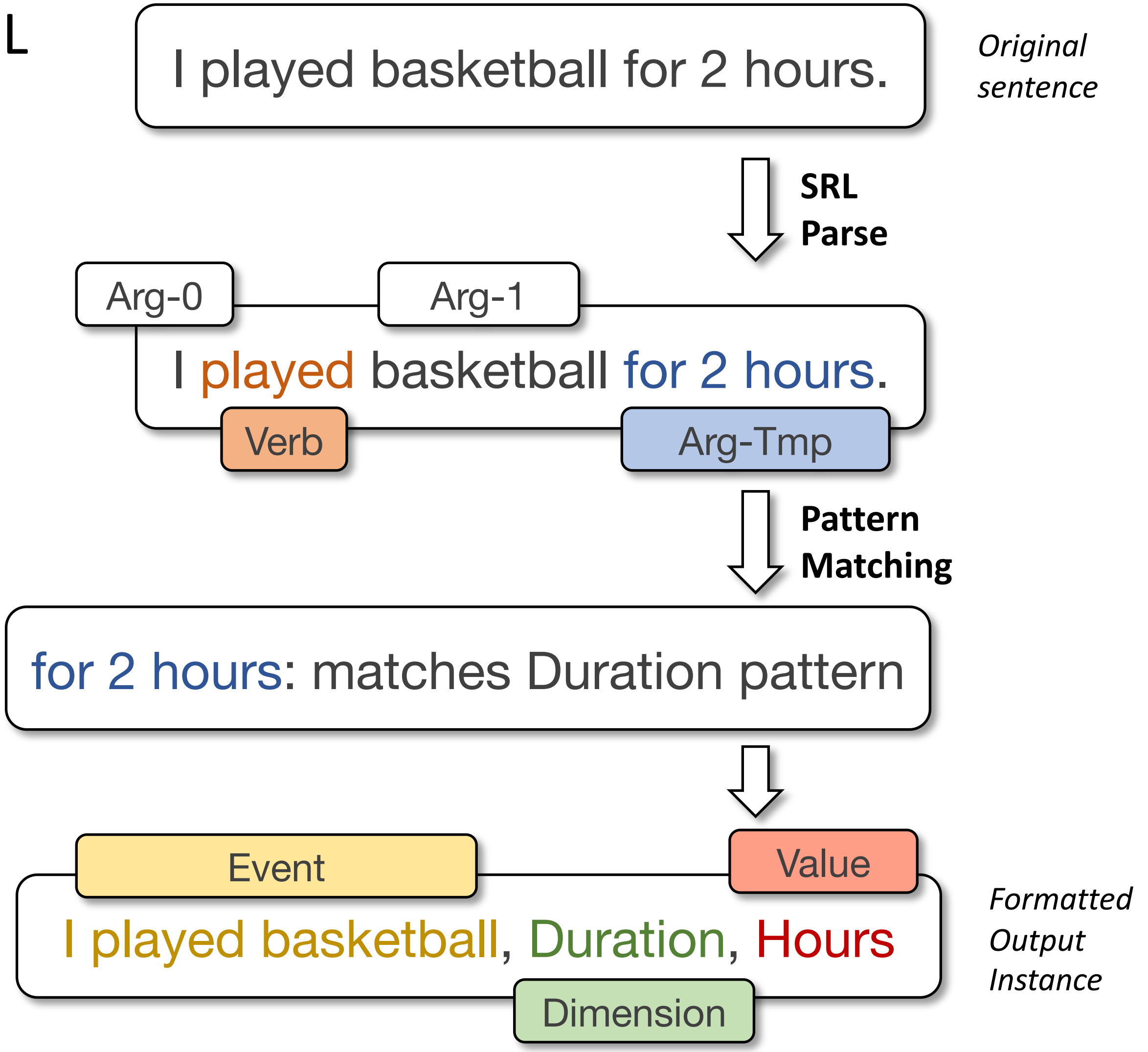
Information Extraction

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



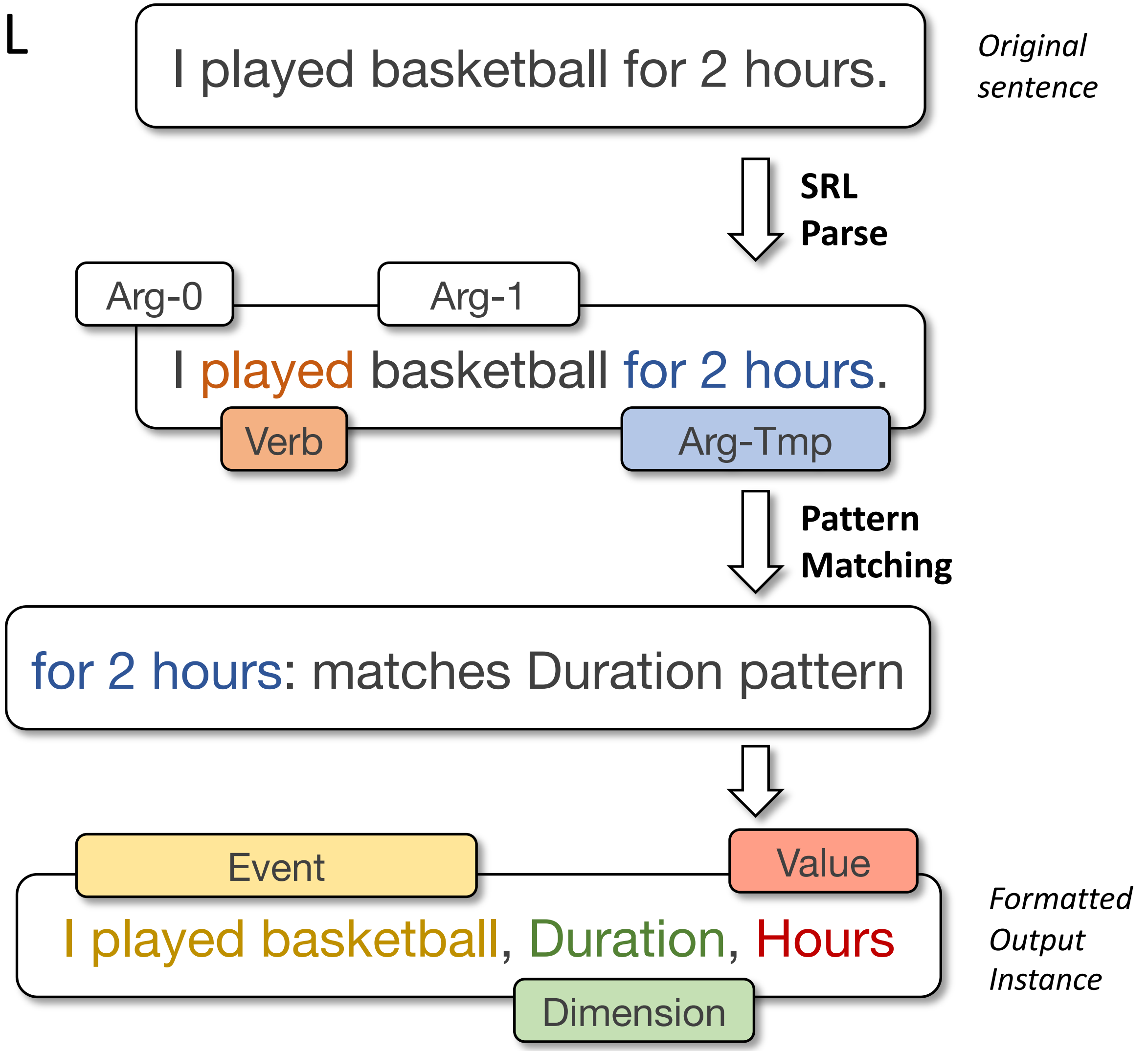
Information Extraction

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



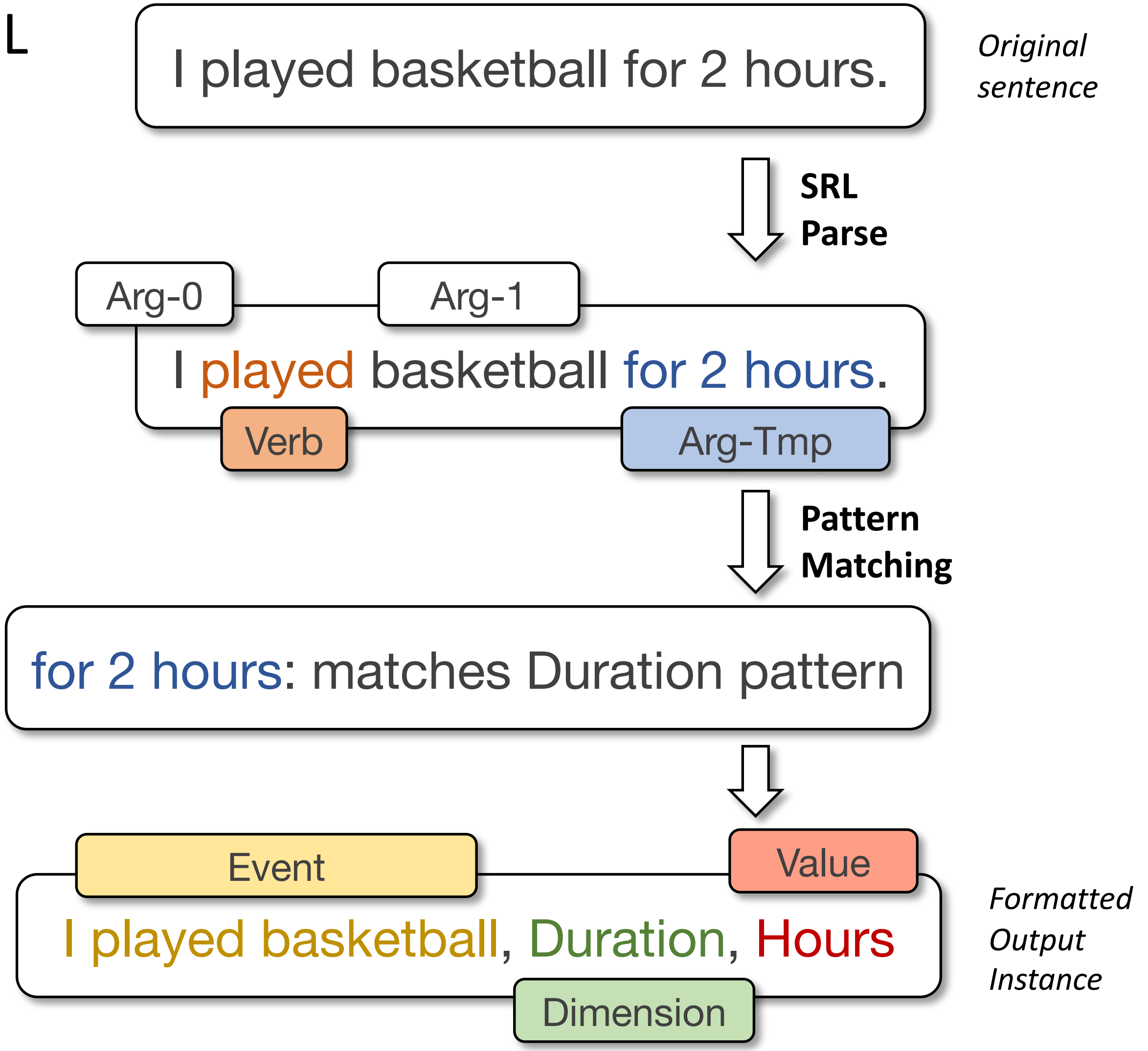
Information Extraction

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



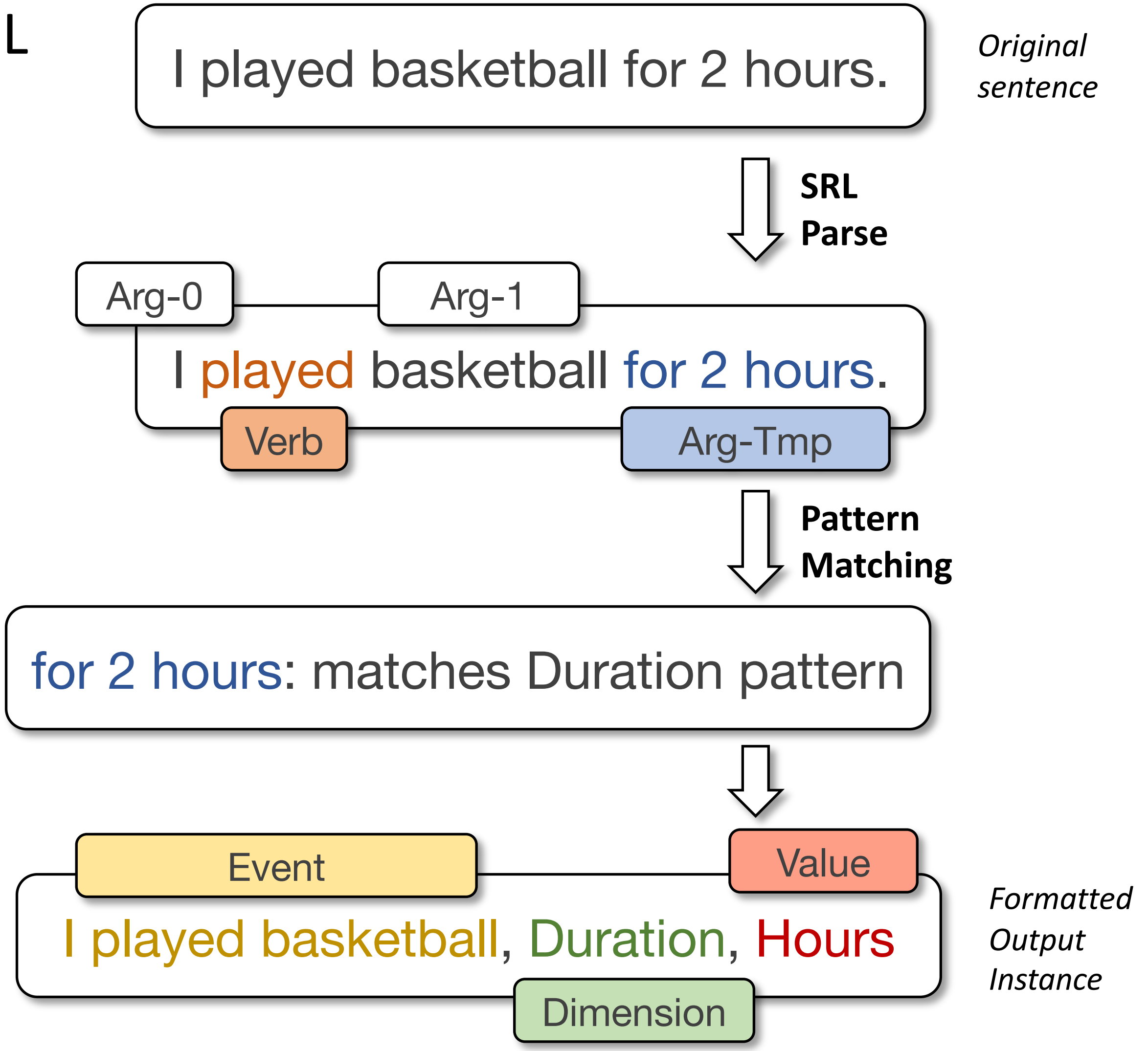
Information Extraction

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



Information Extraction

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



Step 2: Language Model Pre-training

Step 1: Information Extraction

⇩

Step 2: Joint Language Model Pre-training

Output: TacoLM- a time-aware general BERT

Sequence Classification

Sequence Classification

- 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
- With an example:

Sequence Classification

- 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
- With an example:

Sequence Classification

- 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
- With an example:

Sequence Classification

- 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
- With an example:

I played basketball for 2 hours.

Sequence Classification

- 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
- With an example:

Information Extraction

I played basketball for 2 hours.

Sequence Classification

- 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
- With an example:

Sequence Classification

- 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
- With an example:

I played basketball for 2 hours.

Information Extraction

I **played** basketball, Duration, Hours

Sequence Formatting

Sequence Classification

- 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
- With 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]

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
  - Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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
  - Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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
  - Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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
  - Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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
  - Max (P(Event|Dim,Val) + P(Val|Event,Dim)); Preserving original LM capability
- Benefits:
  - Jointly learn **one** transformer towards **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]

- 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 learn **one** transformer towards **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]

- 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 learn **one** transformer towards **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]

- 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 learn **one** transformer towards **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]

- 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 learn **one** transformer towards **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]

- 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 learn **one** transformer towards **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]

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]
- 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

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]
- 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

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"
- 3: Full event masking
  - Instead of 15% used by BERT, we use 60% when masking E1, … En to reduce biases

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"
- 3: Full event masking
  - Instead of 15% used by BERT, we use 60% when masking E1, … En to reduce biases

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"
- 3: Full event masking
  - Instead of 15% used by BERT, we use 60% when masking E1, … En to reduce biases

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”
- 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]

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"
- 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"

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

Step 1: Information Extraction

↓

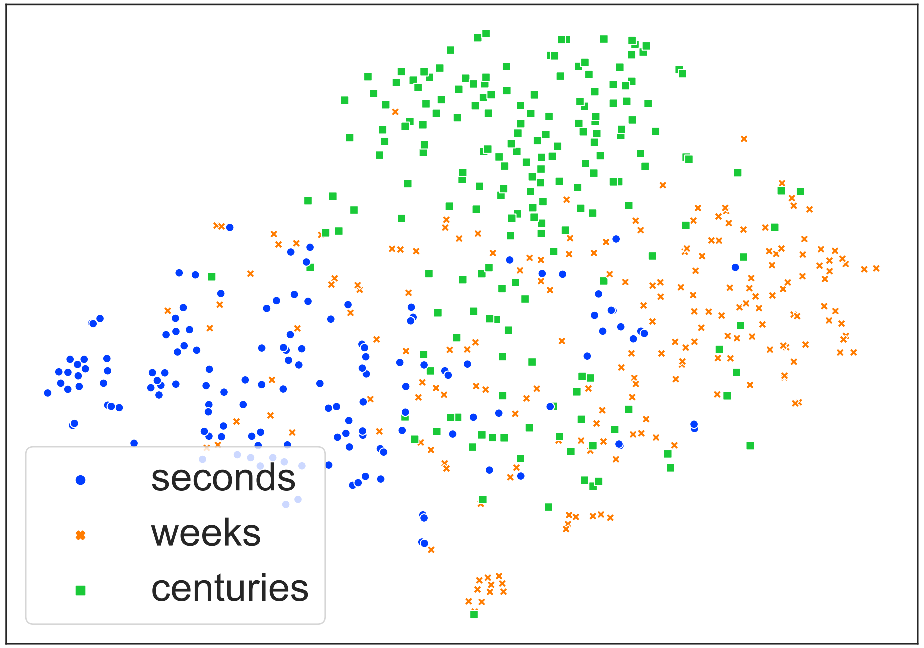
Step 2: Joint Language Model Pre-training

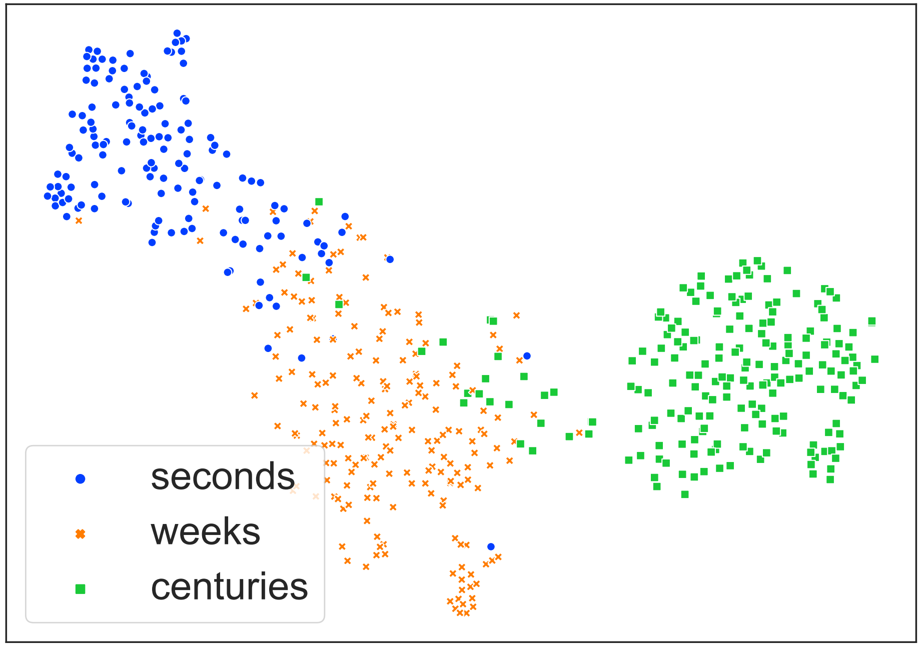
↓

Output: TacoLM- a time-aware general BERT

Evaluation: Intrinsic (Embedding space)

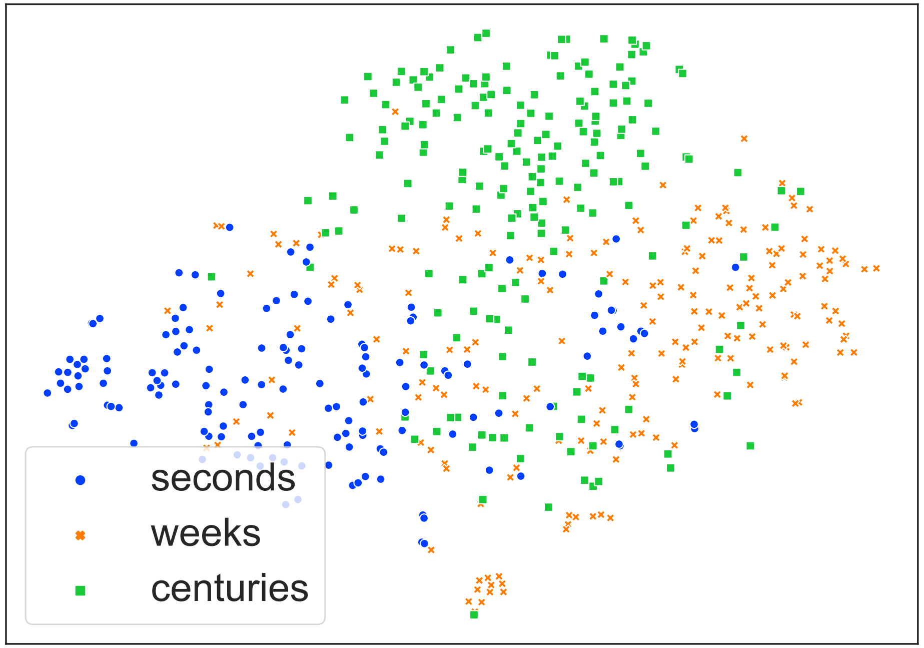
- A collection of events with duration of “seconds,” “weeks” or “centuries” (three extremes)
- BERT (left), Ours (right) representation on the event’s trigger
  - PCA + t-SNE to 2D visualization
- Our model separates the events much better (➔ our model is aware of time)

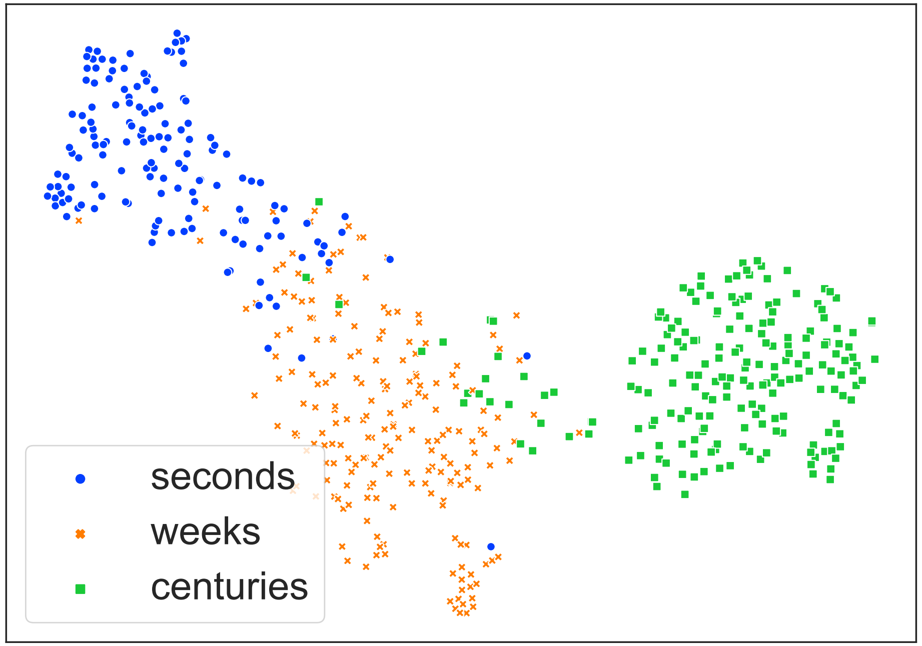




Evaluation: Intrinsic (Embedding space)

- 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)

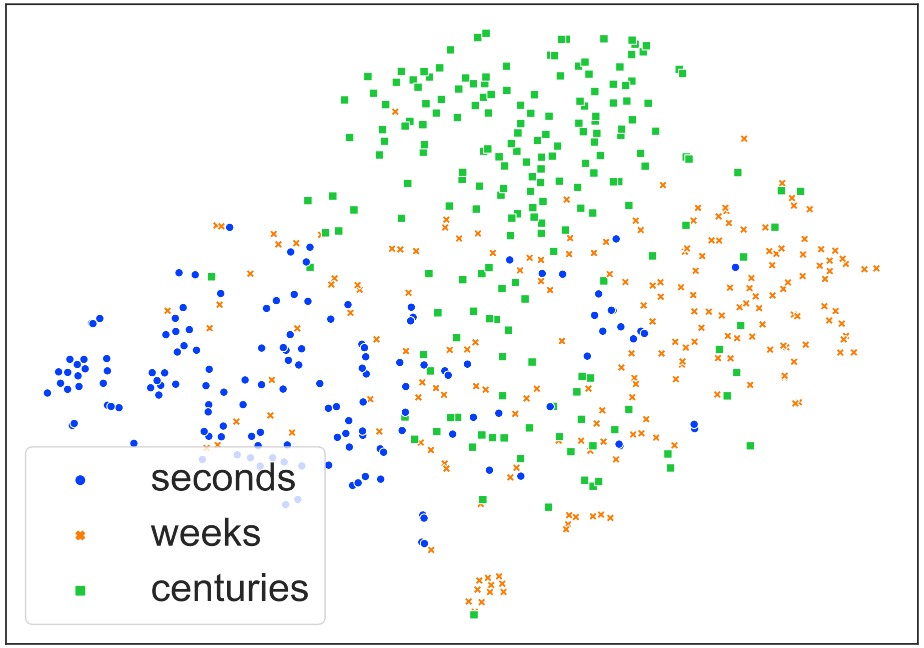




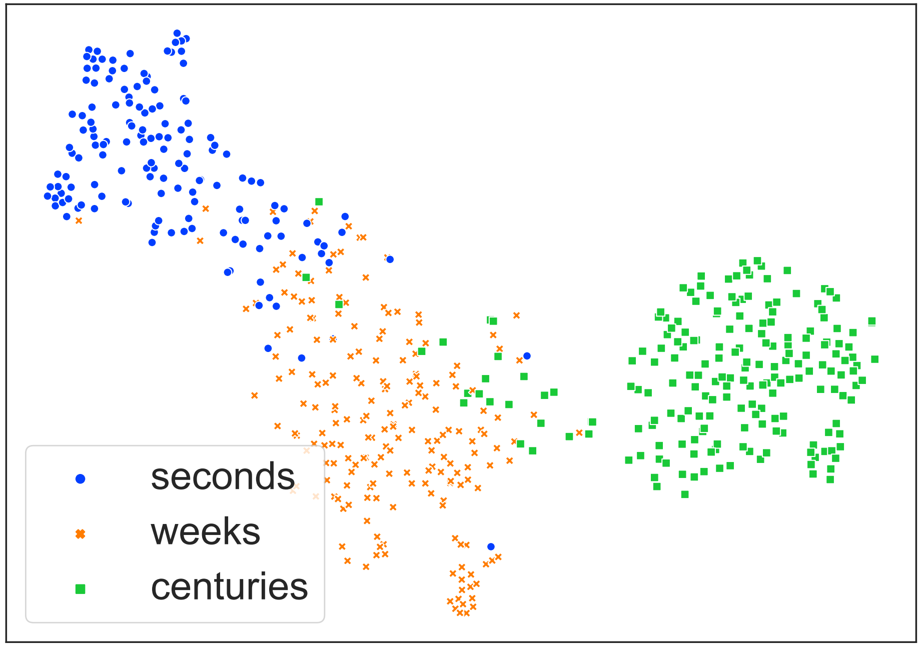
BERT

Evaluation: Intrinsic (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

Evaluation: Intrinsic (Quantitatively)

Evaluation: Intrinsic (Quantitatively)

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

Evaluation: Intrinsic (Quantitatively)

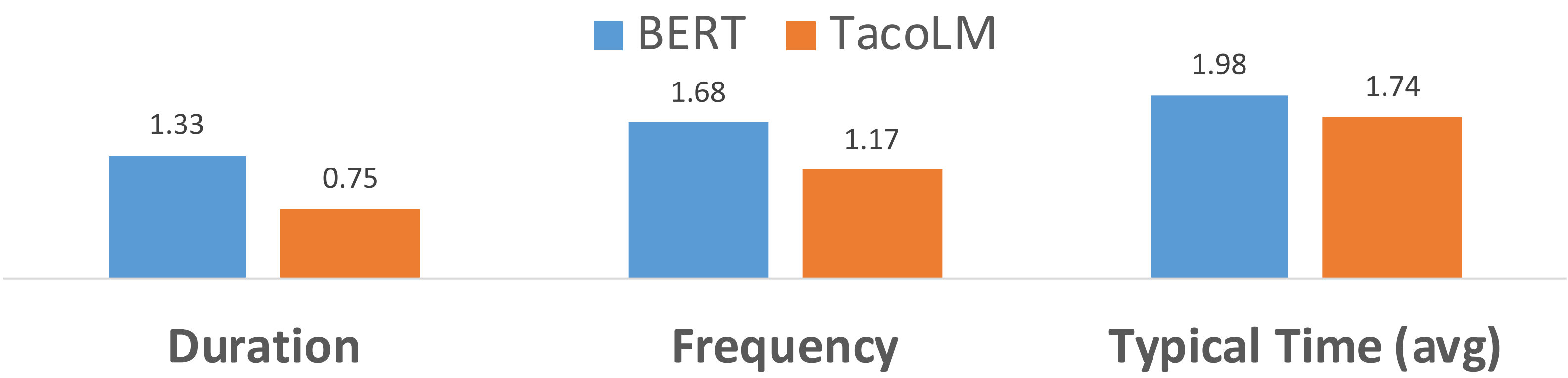
- Metric: Distance to gold label
  - Dist (seconds, hours)=2, Dist (minutes, hours)=1
  - **Lower the better**
- RealNews [Zellers et al. 2019]: no document overlap
  - Raw corpus + MTurk annotation
- UDS-T [Vashishtha et al. 2019]: duration only

Evaluation: Intrinsic (Quantitatively)

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

Evaluation: Intrinsic (Quantitatively)

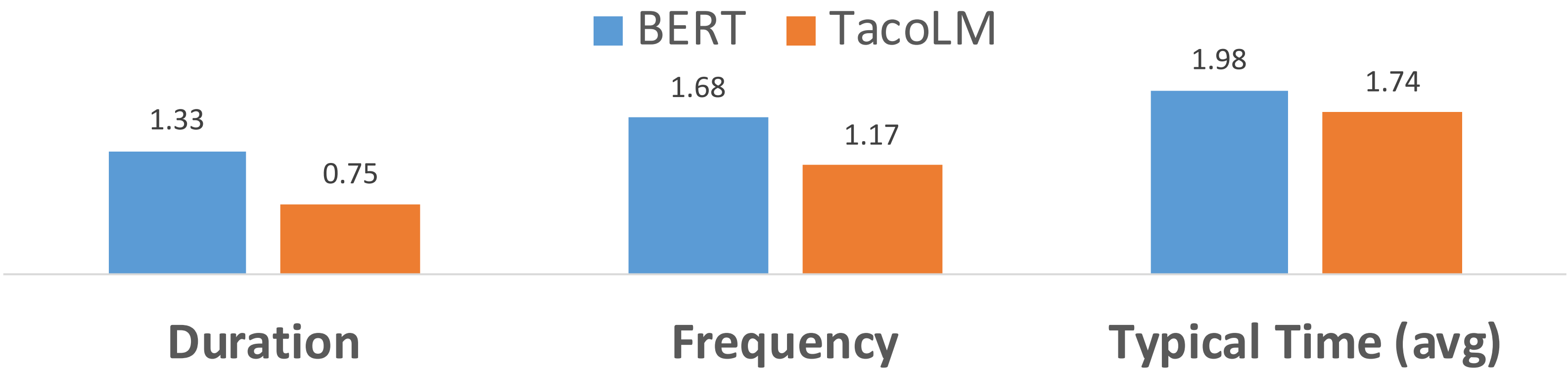
- Metric: Distance to gold label
  - Dist (seconds, hours)=2, Dist (minutes, hours)=1
  - **Lower the better**
- RealNews [Zellers et al. 2019]: no document overlap
  - Raw corpus + MTurk annotation



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

Evaluation: Intrinsic (Quantitatively)

- 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

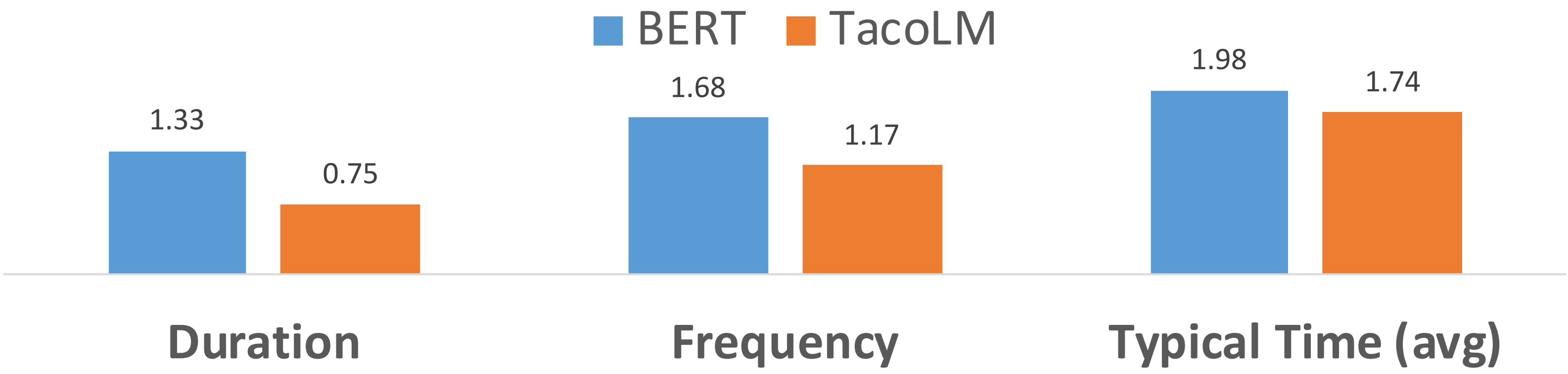


19% average improvement

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

Evaluation: Intrinsic (Quantitatively)

- 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

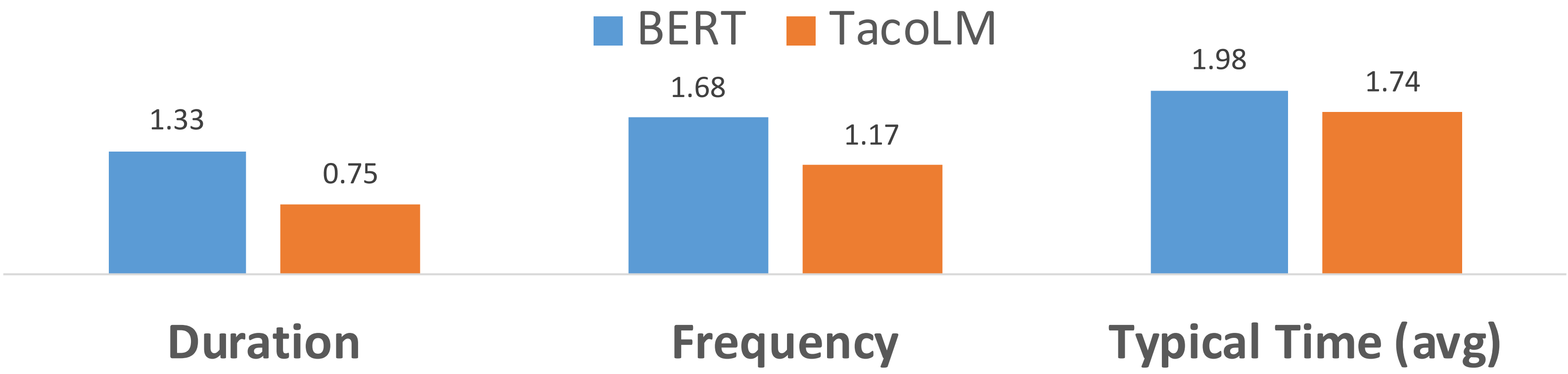


19% average improvement

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

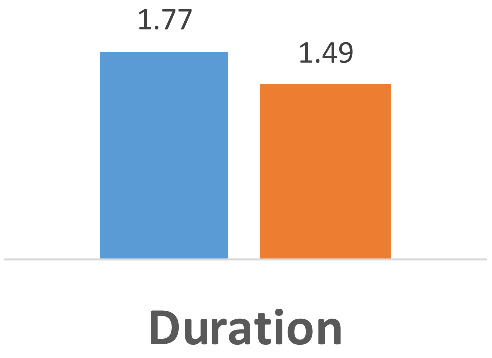
Evaluation: Intrinsic (Quantitatively)

- 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



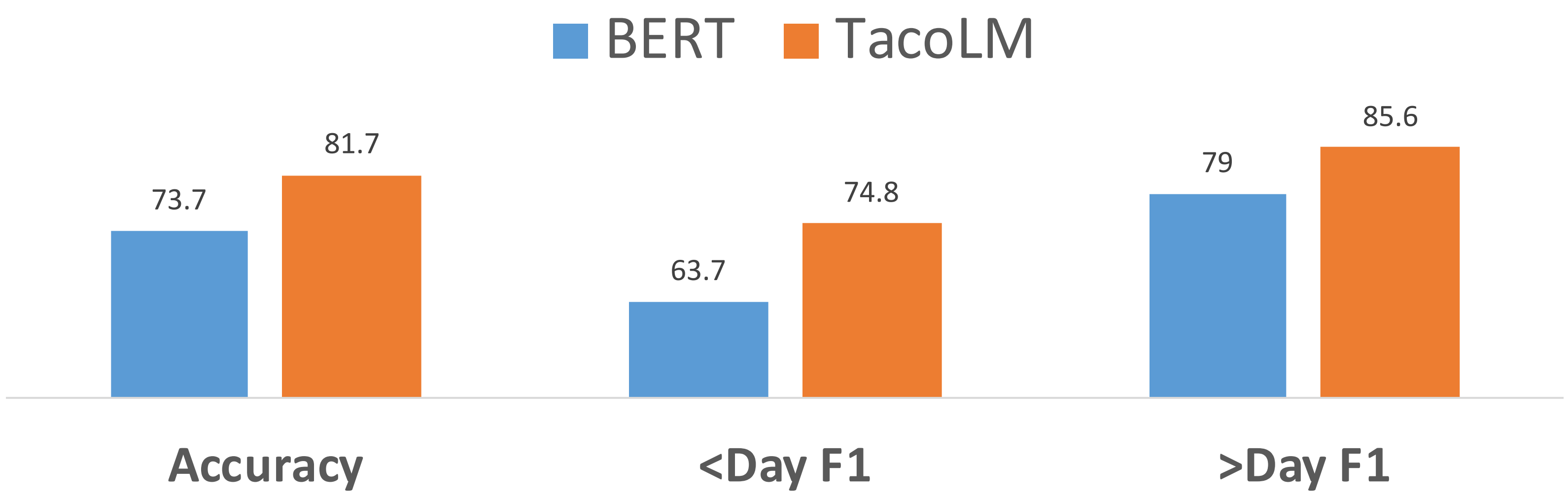
19% average improvement

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



Evaluation: Extrinsic (TimeBank)

- Task: Identify if an event's duration is longer than a day or shorter
- Model (finetuned):
  - Demonstrate the model as a general purpose LM
  - Pre-trained duration prediction layer is not used
- Results



Evaluation: Extrinsic

Evaluation: Extrinsic

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

Evaluation: Extrinsic

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

Evaluation: Extrinsic

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

Evaluation: Extrinsic

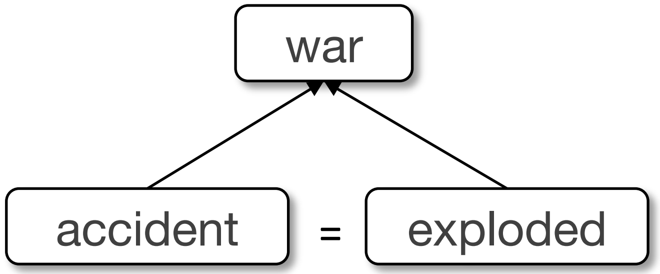
- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
  - HiEVE [Glavas et al. 2014]
  - Child-Parent / Parent-Child / Coreference
    - A bomb exploded. This is the sixth accident since the war started.
- Model (finetuned):
  - Sentence pair classification
- Results (F1, **higher the better**)

Evaluation: Extrinsic

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

Evaluation: Extrinsic

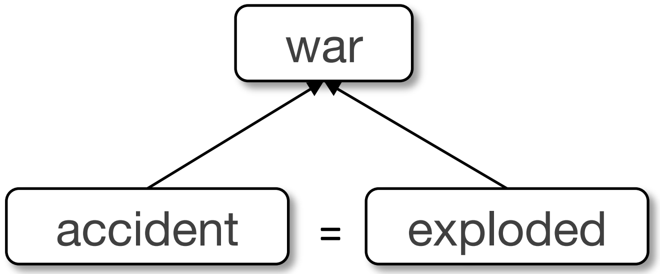
- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
  - HiEVE [Glavas et al. 2014]
  - Child-Parent / Parent-Child / Coreference
    - A bomb exploded. This is the sixth accident since the war started.



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

Evaluation: Extrinsic

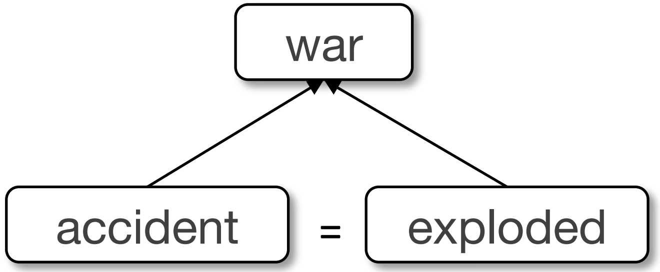
- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
  - HiEVE [Glavas et al. 2014]
  - Child-Parent / Parent-Child / Coreference
    - A bomb exploded. This is the sixth accident since the war started.



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

Evaluation: Extrinsic

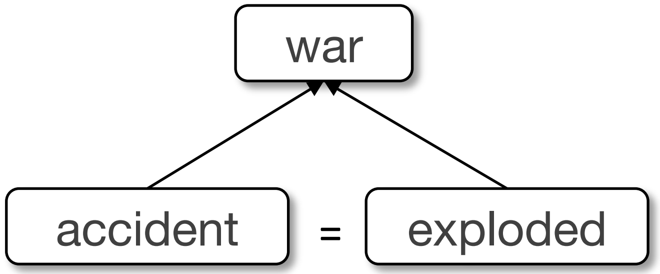
- Use as a general language model with finetuning
- Task: Identify event-event hierarchical relations
  - HiEVE [Glavas et al. 2014]
  - Child-Parent / Parent-Child / Coreference
    - A bomb exploded. This is the sixth accident since the war started.



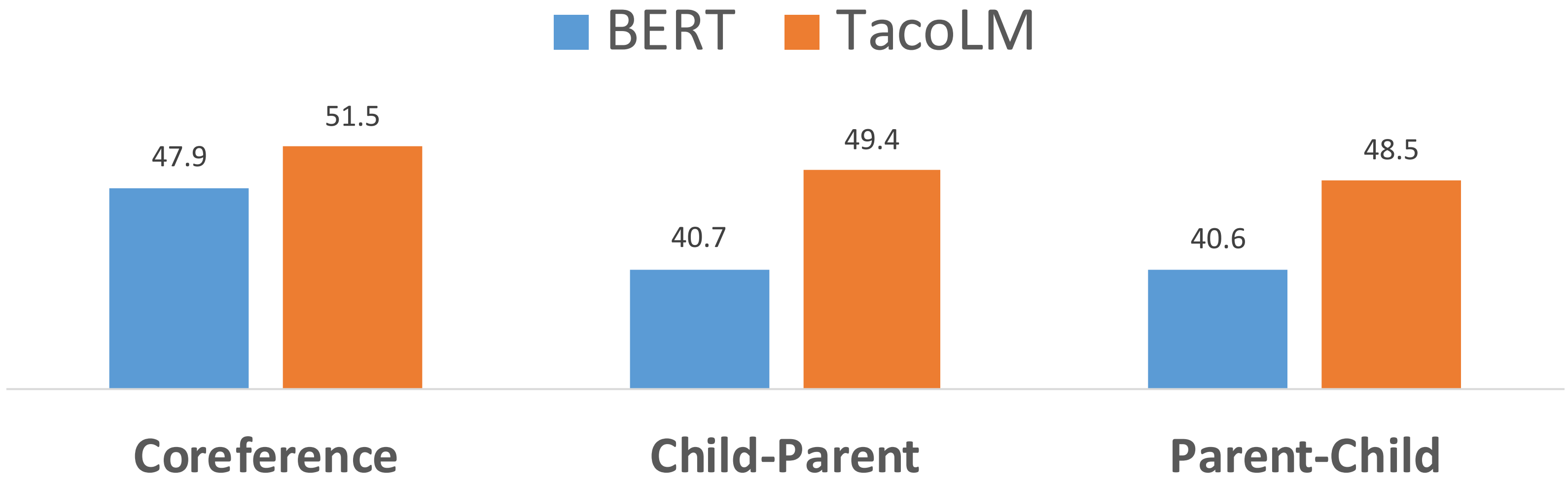
- Model (finetuned):
  - Sentence pair classification
- Results (F1, **higher the better**)

Evaluation: Extrinsic

- 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.

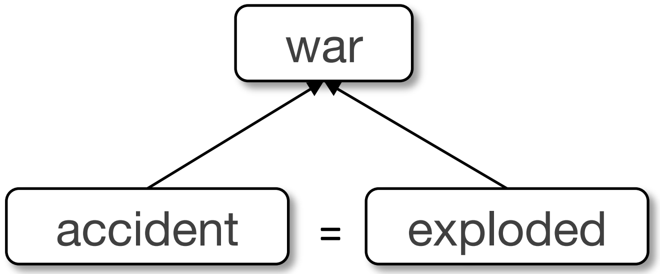


- Model (finetuned):
  - Sentence pair classification
- Results (F1, **higher the better**)

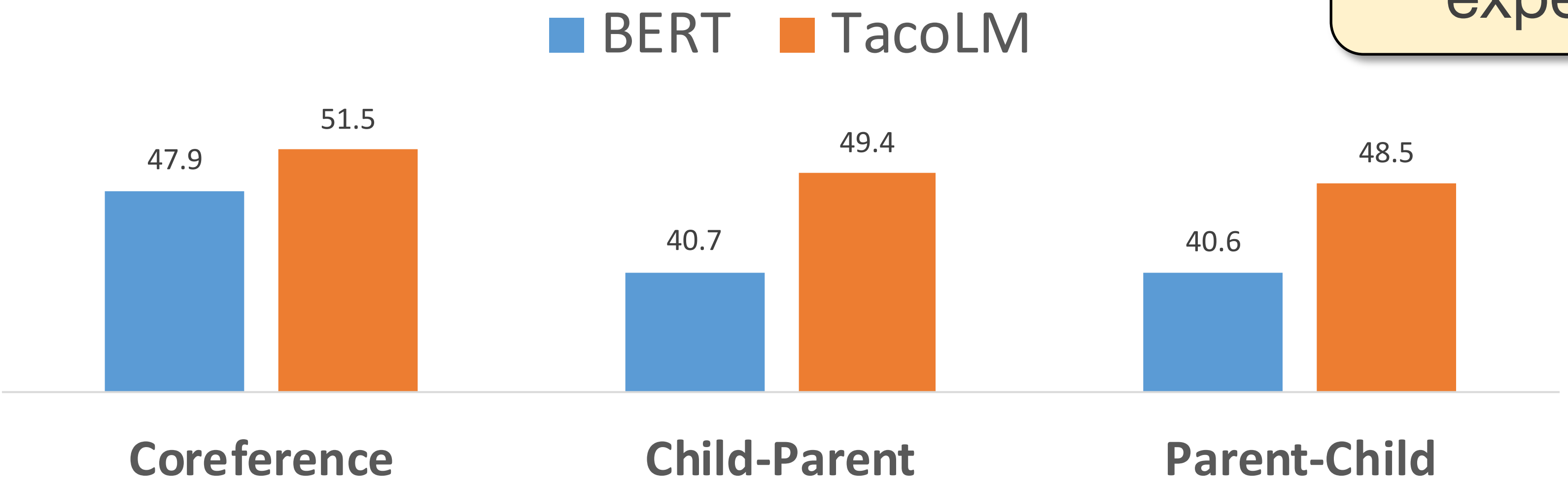


Evaluation: Extrinsic

- 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.



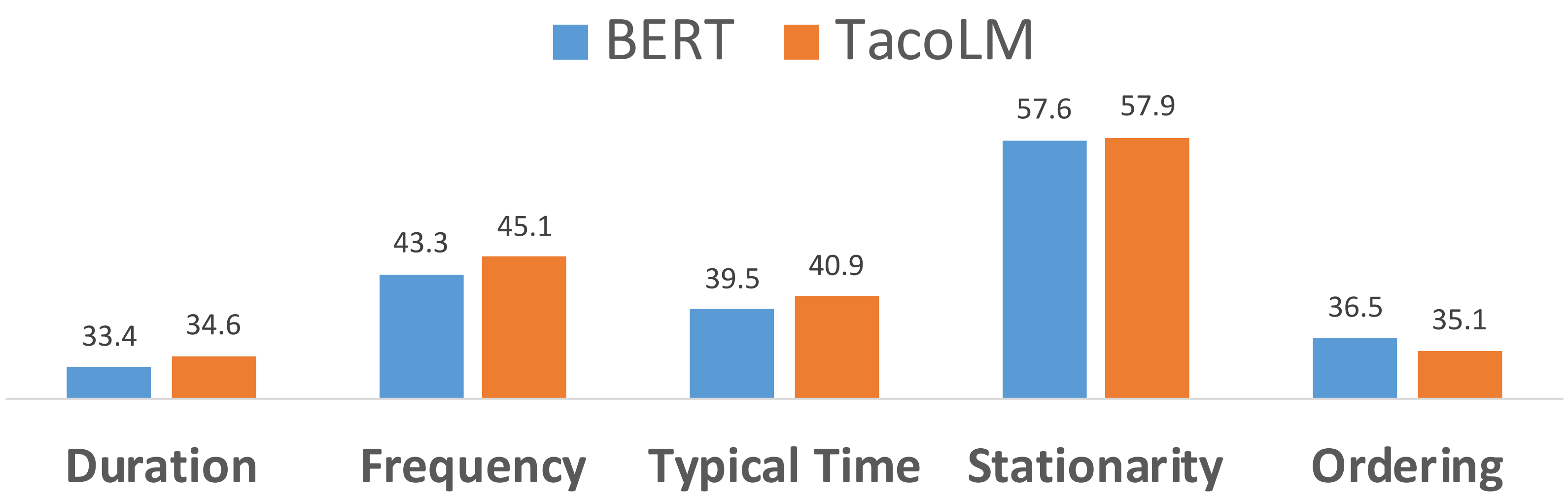
- Model (finetuned):
  - Sentence pair classification
- Results (F1, **higher the better**)



More Intrinsic/Extrinsic experiments in the paper!

Evaluation: Extrinsic (MC-TACO)

- Task: QA on temporal related questions. (how long, how often, etc.)
- Model (finetuned)
  - Standard BERT QA model
- Results



Conclusion - TacoLM

Conclusion - TacoLM

- Time-aware with minimal supervision
- Joint pre-training over multiple temporal dimensions
- Able to directly predict events' duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts
- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

- Able to directly predict events' duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball <u>for 2 hours</u>

- Joint pre-training over multiple temporal dimensions

- Able to directly predict events' duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts

- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

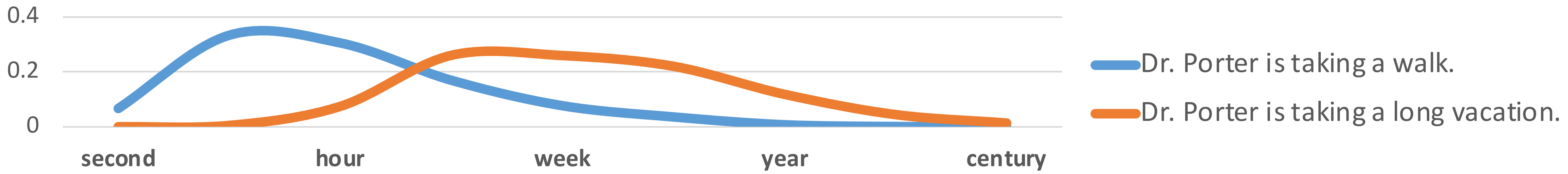
- Time-aware with minimal supervision

I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

Frequency of "brushing teeth" = every morning" → Duration of "brushing teeth" < morning

- Able to directly predict events' duration, frequency or typical time
  - ☐ 19% better on direct prediction tasks
  - ☐ Bell-shaped predictive distributions
  - ☐ Differentiates fine grained event contexts



- Works as a general language model
  - ☐ 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

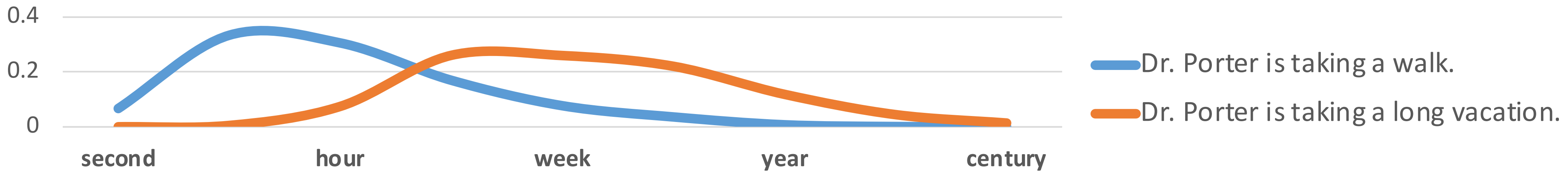
- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts



- Works as a general language model
  - 8% improvement on child-parent event relation extraction

Conclusion - TacoLM

Thank you!
Code & Data:
https://github.com/CogComp/TacoLM

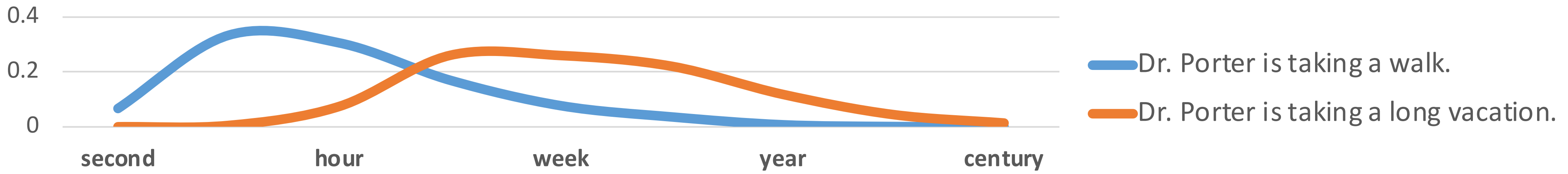
- Time-aware with minimal supervision

  I played basketball for 2 hours

- Joint pre-training over multiple temporal dimensions

  Frequency of “brushing teeth” = every morning” → Duration of “brushing teeth” < morning

- Able to directly predict events’ duration, frequency or typical time
  - 19% better on direct prediction tasks
  - Bell-shaped predictive distributions
  - Differentiates fine grained event contexts



- Works as a general language model
  - 8% improvement on child-parent event relation extraction