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

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

*Police used tear gas*  
*People were angry*
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.

- 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 23\textsuperscript{rd}.
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(N²) 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)
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.

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

<table>
<thead>
<tr>
<th>Example pairs</th>
<th>Temporal Before (%)</th>
<th>Temporal After (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Before</td>
<td>Text After</td>
<td></td>
</tr>
<tr>
<td>Ask</td>
<td>Help</td>
<td>86</td>
</tr>
<tr>
<td>Attend</td>
<td>Schedule</td>
<td>1</td>
</tr>
<tr>
<td>Accept</td>
<td>Propose</td>
<td>10</td>
</tr>
<tr>
<td>Die</td>
<td>Explode</td>
<td>14</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Example pairs</th>
<th>Temporal Before (%)</th>
<th>Temporal After (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Before</td>
<td>Text After</td>
<td></td>
</tr>
<tr>
<td>Ask</td>
<td>Help</td>
<td>86</td>
</tr>
<tr>
<td>Attend</td>
<td>Schedule</td>
<td>1</td>
</tr>
<tr>
<td>Accept</td>
<td>Propose</td>
<td>10</td>
</tr>
<tr>
<td>Die</td>
<td>Explode</td>
<td>14</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Example pairs</th>
<th>Temporal Before (%)</th>
<th>Temporal After (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Before</td>
<td>Text After</td>
<td></td>
</tr>
<tr>
<td>Ask</td>
<td>Help</td>
<td>86</td>
</tr>
<tr>
<td>Attend</td>
<td>Schedule</td>
<td>1</td>
</tr>
<tr>
<td>Accept</td>
<td>Propose</td>
<td>10</td>
</tr>
<tr>
<td>Die</td>
<td>Explode</td>
<td>14</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Example pairs</th>
<th>Temporal Before (%)</th>
<th>Temporal After (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text Before</td>
<td>Text After</td>
<td></td>
</tr>
<tr>
<td>Ask</td>
<td>Help</td>
<td>86</td>
</tr>
<tr>
<td>Attend</td>
<td>Schedule</td>
<td>1</td>
</tr>
<tr>
<td>Accept</td>
<td>Propose</td>
<td>10</td>
</tr>
<tr>
<td>Die</td>
<td>Explode</td>
<td>14</td>
</tr>
</tbody>
</table>

Priors on order are often different than order of occurrence in text.
Event Order Distributions

Before “grant”

After “grant”

<table>
<thead>
<tr>
<th>Event</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>seek</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>know</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>need</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>zone</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>request</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>involve</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>write</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>use</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>pay</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>include</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>leave</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>speak</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>help</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>work</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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**)

![Diagram of the neural architecture](image)
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**)

![Diagram of the neural architecture](image)
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.
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.
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.
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
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
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.
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.
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.

<table>
<thead>
<tr>
<th>Stationarity</th>
<th>Duration</th>
<th>Typical Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>College: about 4 years, starts at the age of 18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bill in North Carolina: about 4 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duke in North Carolina: always</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Event</th>
<th>Duration</th>
<th>Typical Time</th>
<th>Stationarity</th>
<th>Order</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>about 4 years</td>
<td>start at the age of 18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bill in North Carolina</td>
<td>about 4 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duke in North Carolina</td>
<td>always (expected)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Join Google</td>
<td>after college graduation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBA Finals</td>
<td>every year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
My friend Bill went to Duke University in North Carolina. With a degree in CS, he joined Google MTV as a software engineer. As a huge basketball fan, he has attended all 3 NBA finals since then. He also plans to visit Duke regularly as an alumnus to attend their home games.
Temporal Common Sense

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

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.

[Elazar et al. ACL’19]

Ning et al. NAACL’18
Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - Multiple Choice Temporal Commonsense
  - 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]

- Multiple Choice Temporal Commonsense
- 1,893 questions; 13,225 question-answer pairs
- Querying at least one of the five dimensions:
  - Duration
  - Frequency
  - Typical Occurring Time
  - Stationarity
  - Ordering

Example: He went to Duke University. How long did it take him to graduate?

- 4 years (Gold)
- 10 days
- 3.5 years
- 16 hours
Defining the Temporal Commonsense Challenge

- **MC-TACO** [Zhou et al. EMNLP 2019]
  - Multiple Choice Temporal Commonsense
  - 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?

<table>
<thead>
<tr>
<th>Gold</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 years</td>
<td></td>
</tr>
<tr>
<td>10 days</td>
<td></td>
</tr>
<tr>
<td>3.5 years</td>
<td></td>
</tr>
<tr>
<td>16 hours</td>
<td></td>
</tr>
</tbody>
</table>
### Defining the Temporal Commonsense Challenge

- **MC-TACO [Zhou et al. EMNLP 2019]**
  - Multiple Choice TemporAl COmmon-sense
  - 1,893 questions; 13,225 question-answer pairs
  - Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

---

<table>
<thead>
<tr>
<th>He went to Duke University.</th>
<th>How long did it take him to graduate?</th>
<th>Gold</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 years</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>10 days</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td></td>
<td>3.5 years</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td></td>
<td>16 hours</td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>
Defining the Temporal Commonsense Challenge

- MC-TACO [Zhou et al. EMNLP 2019]
  - Multiple Choice Temporal Common-sense
  - 1,893 questions; 13,225 question-answer pairs
  - Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering

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

**Reasoning type:** 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

**Reasoning type:** stationarity

- [ ] yes

**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?

- [ ] 19 hours
- [x] 45 minutes
- [ ] 15 days
- [ ] 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
- [x] 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.

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

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
**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

**Predicted Duration from TacoLM**

- Dr. Porter is taking a walk.
- Dr. Porter is taking a long vacation.
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 \leq 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

**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

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

**Step 2: Joint Masked Language Model**
TacoLM – the Big Picture

**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 $\sim 1 / \text{Frequency}$
    - “I brush my teeth every morning”
    - Duration of “brushing teeth” < morning
  - Further generalization to combat reporting biases

**Goal:** build a general time-aware LM with minimal supervision
TacoLM – the Big Picture

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

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

Original sentence: I played basketball for 2 hours.
Information Extraction

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

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

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

I played basketball for 2 hours.

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

Original sentence
I played basketball for 2 hours.

SRL Parse

Pattern Matching
for 2 hours: matches Duration pattern

Formatted Output
I played basketball, Duration, Hours
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

Original sentence: I played basketball for 2 hours.

Verb: played
Arg-0: basketball
Arg-1: for
Arg-Tmp: 2

Pattern Matching: for 2 hours: matches Duration pattern

Event: played basketball
Dimension: Duration
Value: Hours

Formatted Output Instance: I played basketball, Duration, Hours
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.
  I played basketball, Duration, Hours
  I [M] played basketball [SEP] [M] [DUR] [HRS]
Joint Model with Masked LM

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

I [M] played basketball [SEP] [M] [DUR] [HRS]
Joint Model with Masked LM

- 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

- 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

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

   $$\hat{x} = \log(\text{softmax}(x))$$
   $$\text{loss} = -\hat{x}^T y$$

2: Label weight adjustment
   - Instances with “seconds” have higher loss than those with “years”
Joint Model with Masked LM

1: Soft cross entropy for recovering Val
   - If gold label is “hours”, the label vector $y$ for “minutes, hours, days” will be $[0.16, 0.47, 0.25]$
   $$\hat{x} = \log(\text{softmax}(x))$$
   $$\text{loss} = -\hat{x}^T y$$

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

- [M] played basketball [SEP] [M] [DUR] [HRS]
- [M] had a cup of [MASK] [SEP] [M] [TYP] [Evening]  $\Rightarrow$ MASK = coffee, because “cup of”
- [M] had [MASK] [MASK] of [MASK] [SEP] [M] [TYP] [Evening]
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)
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

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

<table>
<thead>
<tr>
<th>BERT</th>
<th>TacoLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.33</td>
<td>0.75</td>
</tr>
<tr>
<td>1.68</td>
<td>1.17</td>
</tr>
<tr>
<td>1.98</td>
<td>1.74</td>
</tr>
</tbody>
</table>

19% average improvement

<table>
<thead>
<tr>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.77</td>
</tr>
<tr>
<td>1.49</td>
</tr>
</tbody>
</table>
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.

Diagram:
- war
- accident = exploded
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.
- Model (finetuned):
  - Sentence pair classification
- Results (F1, higher is better)

<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>TacoLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coreference</td>
<td>47.9</td>
<td>51.5</td>
</tr>
<tr>
<td>Child-Parent</td>
<td>40.7</td>
<td>49.4</td>
</tr>
<tr>
<td>Parent-Child</td>
<td>40.6</td>
<td>48.5</td>
</tr>
</tbody>
</table>
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!
Ways to acquire, represent and distill commonsense knowledge
- Along multiple dimensions: Physical, Social, Temporal
- Some require crowdsourcing, some can be extracted directly from text

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 text

- 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 text

- 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 text

- 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

- Ways to acquire, represent and distill commonsense knowledge
  - Along multiple dimensions: Physical, Social, Temporal
  - Some require crowdsourcing, some can be extracted directly from text

- 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”
  - ....

Thank you!

So, did Aristotle have a laptop?