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

Time

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

Understanding Time is Important

Police used tear gas

People were angry

Time

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

Understanding Time is Important

Police used tear gas

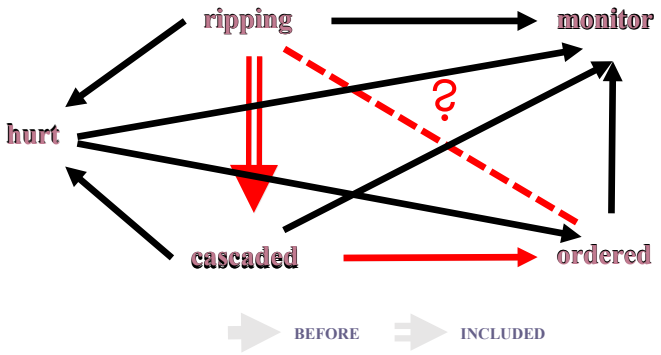
People were angry

Time

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

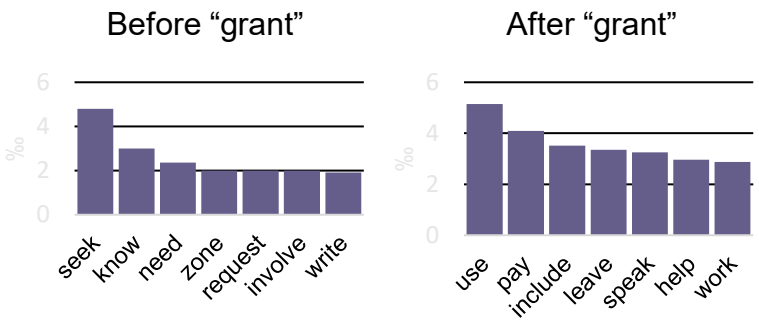
Temporal Relation Extraction

- Structured Learning (Ning et al. EMNLP 17)



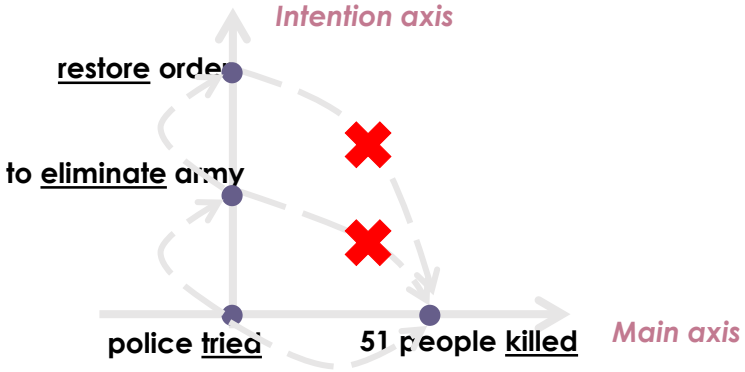
- Prior (Commonsense) Knowledge (Ning et al. NAACL 18)
  - More than 10 people have (**event1: died**), police said. A car (**event2**: **exploded**) on Friday in a group of men.

Example pairs		Before (%)	After (%)
Event 1	Event 2		
Ask	Help	86	9
Attend	Schedule	1	82
Accept	Propose	10	77
Die	Explode	14	83

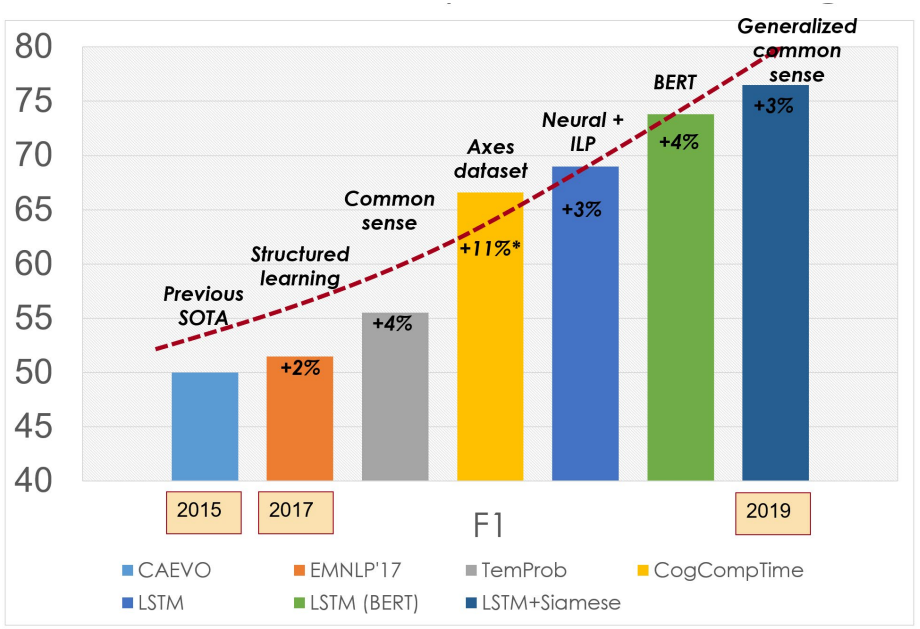


Temporal Relation Extraction

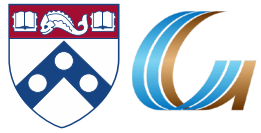
- Data Annotation (Ning et al. ACL 18)
  - *Police **tried** to **eliminate** the pro-independence army and **restore** order. At least 51 people were **killed** in clashes between police and citizens in the troubled region.*



- Joint Learning
  - Ning et al. EMNLP 18, EMNLP 19
  - *Using all these to develop powerful models*



Defining the Temporal Commonsense Challenge



- Moving Foward
  - *[Ning et al. 18]* considered relationtional common sense between event triggers
  - Other aspects: Duration, Frequency, Typical Time, Stationarity
- Why should we care?

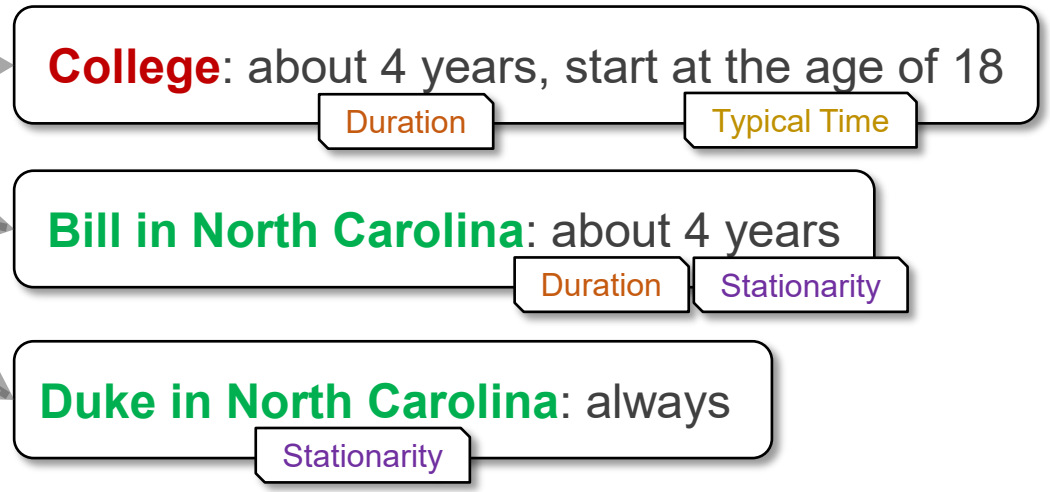
The Temporal Commonsense Challenge

- Most recent line of research : how to deal with temporal information that is implicit in text

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.

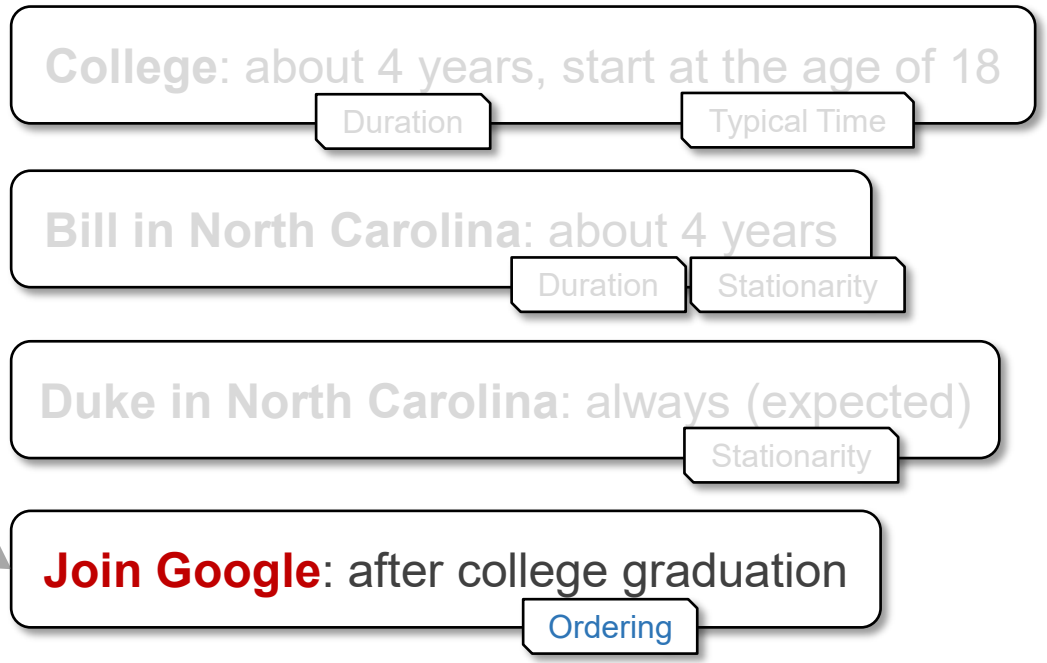
The Temporal Commonsense Challenge

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.



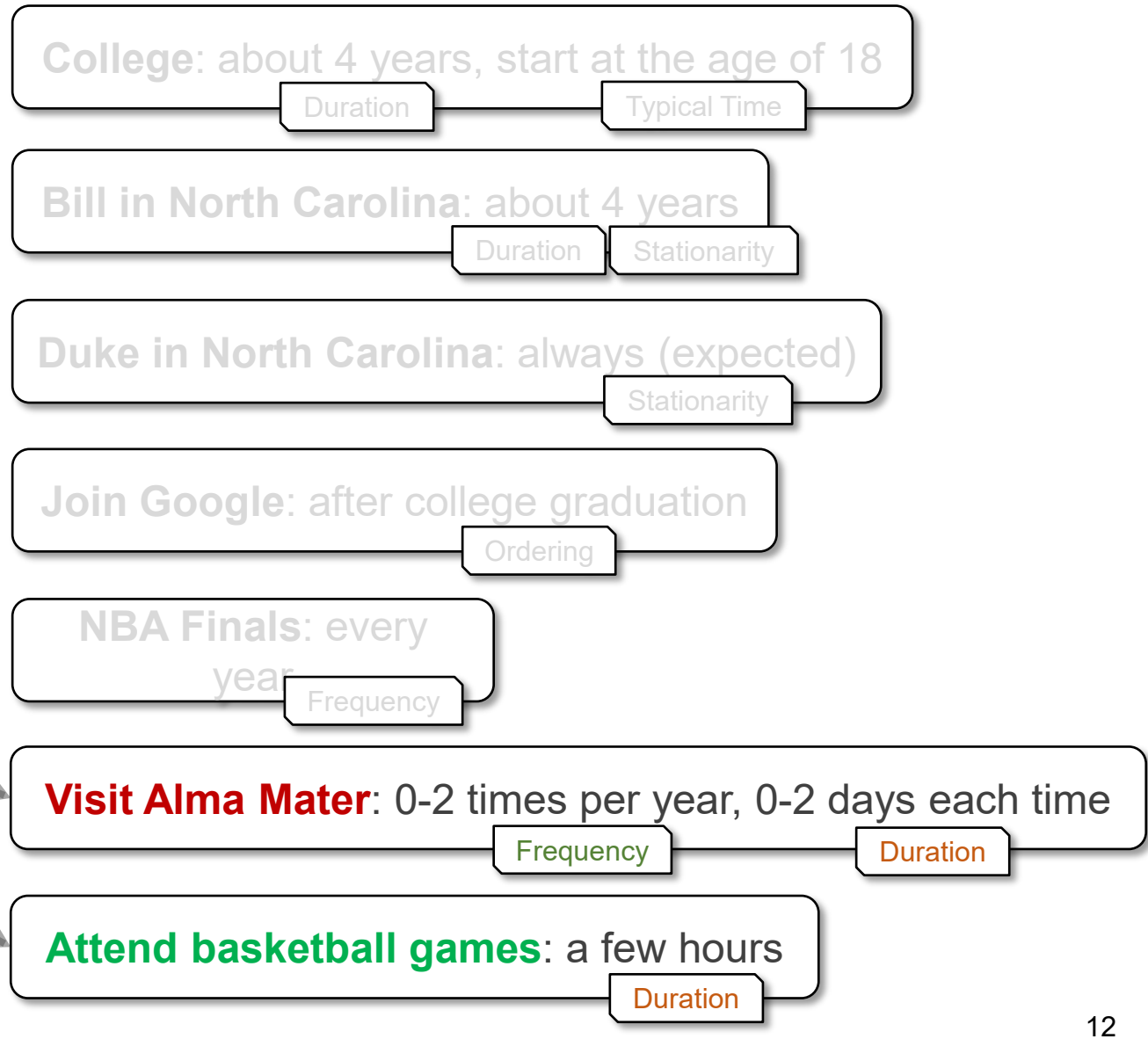
The Temporal Commonsense Challenge

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.



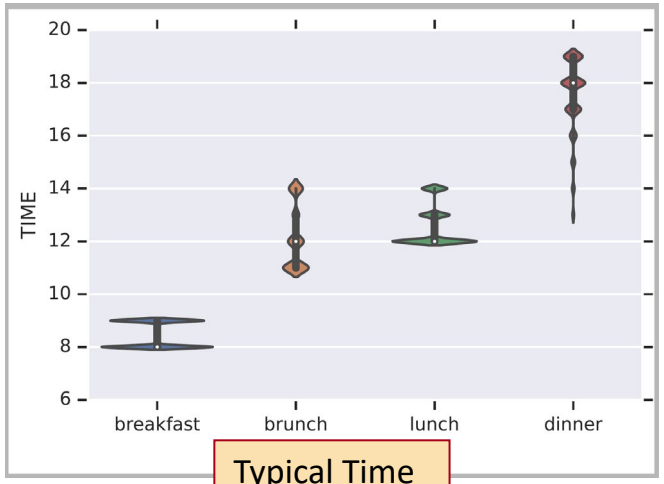
The Temporal Commonsense Challenge

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.

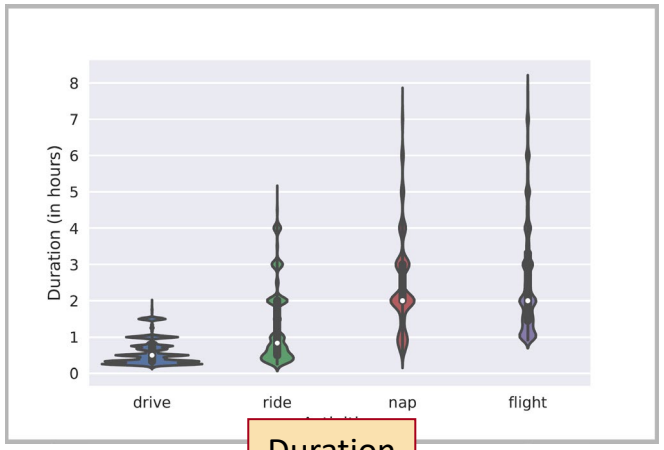


Temporal Common Sense

- Most of the effort is in Knowledge Acquisition: Duration, typical time, frequency.



Typical Time



Duration

- Missing: Stationarity, what things can happen together? (cleaning KBs)
- How to use?

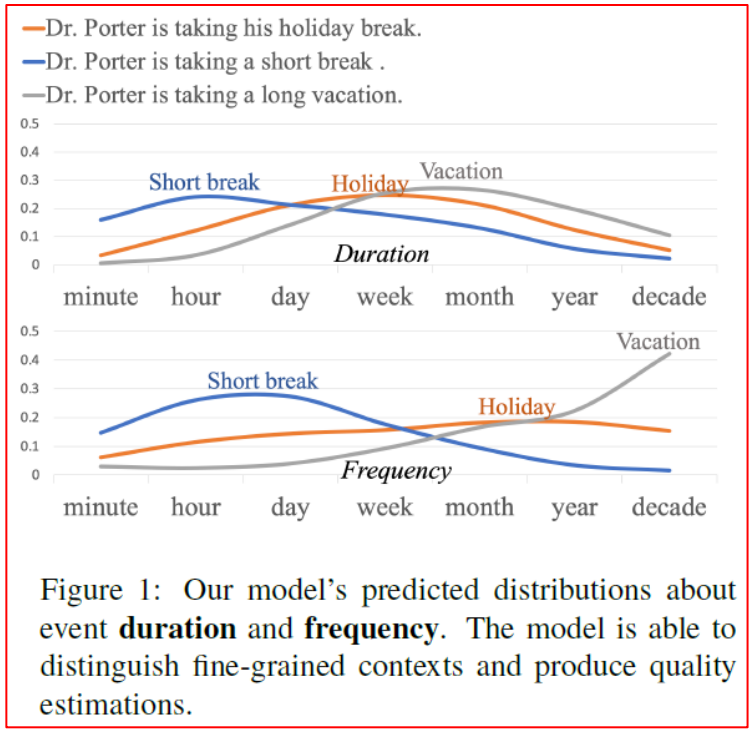
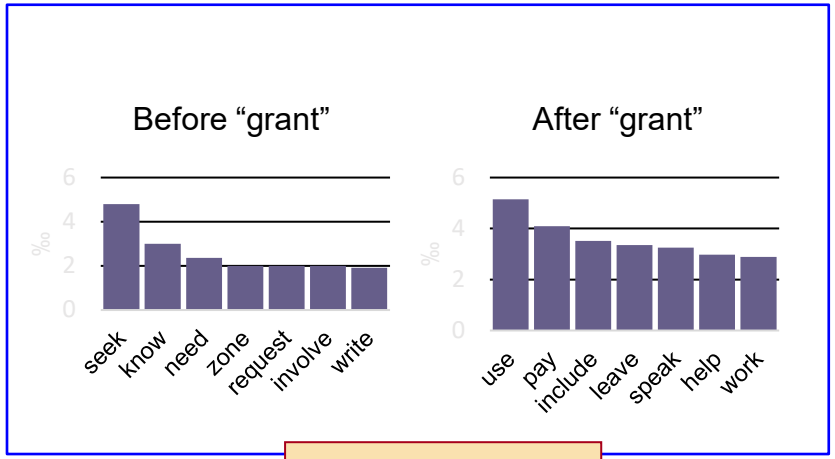


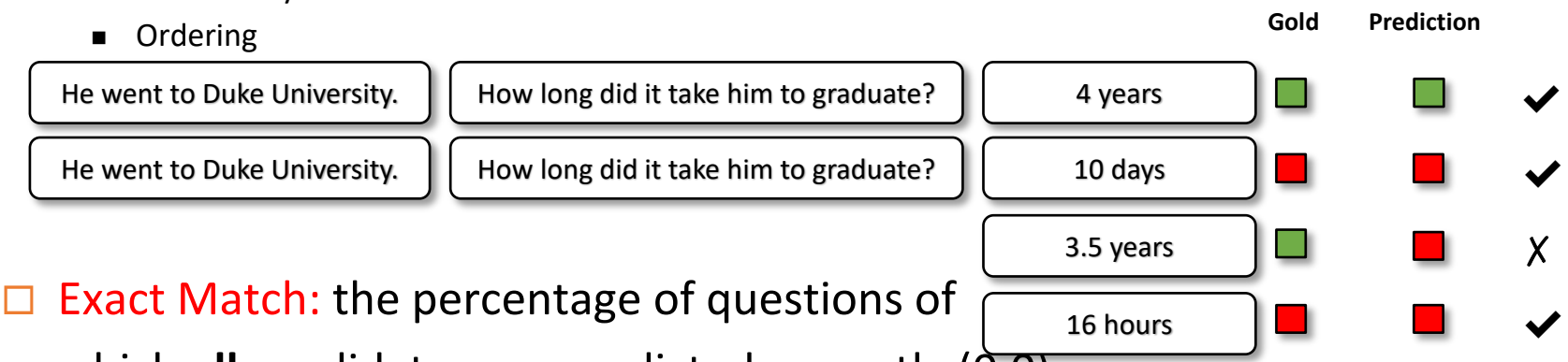
Figure 1: Our model's predicted distributions about event **duration** and **frequency**. The model is able to distinguish fine-grained contexts and produce quality estimations.



Typical Temporal Relations

Defining the Temporal Commonsense Challenge

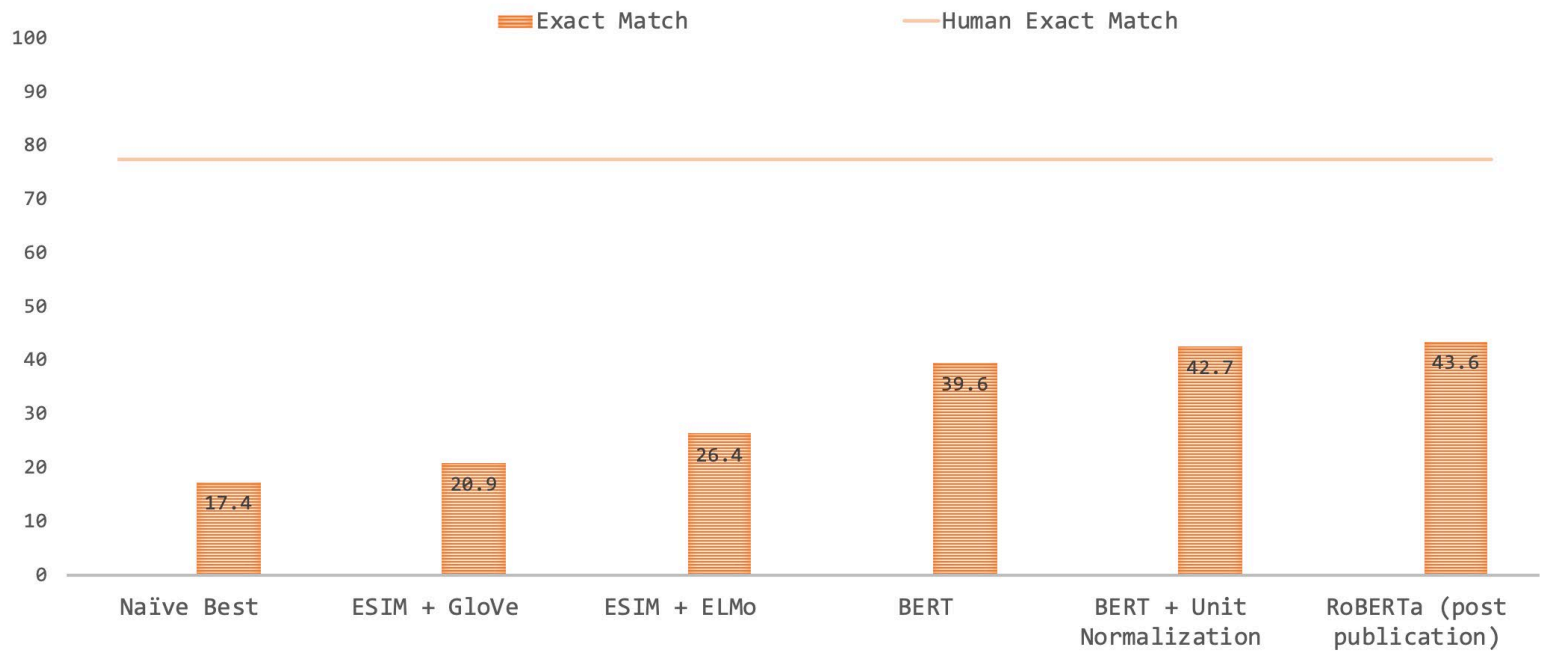
- MC-TACO (Zhou et al. EMNLP 2019)
  - **m**ultiple **c**hoice **t**empor**a**l **co**mmon-sense
  - 1,893 questions and 13,225 question-answer pairs
  - Querying at least one of the five dimensions:
    - Duration
    - Frequency
    - Typical Occurring Time
    - Stationarity
    - Ordering



  - Exact Match: the percentage of questions of which **all** candidates are predicted correctly (0.0)
  - F1: The F1 score of "plausible" (66.7)

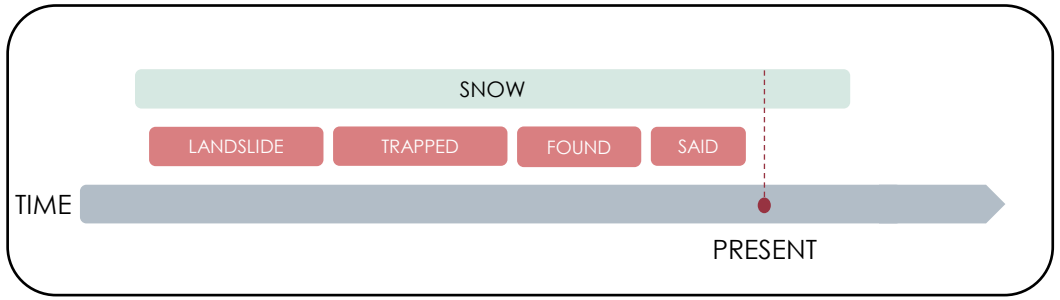
Defining the Temporal Commonsense Challenge

- MC-TACO (Zhou et al. EMNLP 2019)
  - RoBERTa is 30% behind human performance on Exact Match



Temporal Relation QA Task

- Existing QA models are not capable of handling time
- *Heavy **snow** is causing disruption to transport across the UK, with heavy rainfall bringing flooding to the south-west of England. Rescuers searching for a woman **trapped** in a **landslide** at her home **said** they had **found** a body.*



- *What happened **before** a woman was trapped?*
  - ELMo-BiDAF (SQuAD): they found a body
  - BERT (SQuAD 2.0): a landslide
- *What happened **after** a woman was trapped?*
  - ELMo-BiDAF (SQuAD): they found a body
  - BERT (SQuAD 2.0): a landslide

Temporal Relation QA Task

- There is a need for a more comprehensive annotation
- Current “standard” annotation leaves out a lot of phenomena
  - an inherent limitation of a fixed annotation scheme:
  - The lion had a large meal and slept for 24 hours
    - [Negated] What *didn’t* the lion do after the large meal?
    - [Uncertain] What *might* the lion do after the large meal?
  - Before v. Often Before
    - Bob ran last evening
    - What does Bob *often* do in the evening? -> No answer
    - What does Bob *sometimes* do in the evening? -> Running
  - Event describing occurring actions v. Single action
    - I *played* basketball v. I *play* basketball

Temporal Relation QA Task

- TORQUE
  - Temporal ORder QUEstion-answering
- Question Answering
  - Easy to annotate
  - Covers some phenomena that are hard to annotate with previous schemes
- 3k new passages, 21k questions, 25k events
- RoBERTa-Large: 53% Exact Match
- Human (preliminary): 71-88%

Heavy snow is causing disruption to transport across the UK, with heavy rainfall bringing flooding to the south-west of England. Rescuers searching for a woman trapped in a landslide at her home said they had found a body.

Q1: What event has already finished?
A: searching trapped landslide said found
Q2: What event has begun but has not finished?
A: snow causing disruption rainfall bringing flooding
Q3: What will happen in the future?
A: No answers.

Hard-coded questions

Q4: What happened before a woman was trapped?
A: landslide
Q5: What had started before a woman was trapped?
A: snow rainfall landslide
Q6: What happened while a woman was trapped?
A: searching
Q7: What happened after a woman was trapped?
A: searching said found

Group of contrast questions

Q8: What happened at about the same time as the snow?
A: rainfall
Q9: What happened after the snow started?
A: causing disruption bringing flooding searching trapped landslide said found
Q10: What happened before the snow started?
A: No answers.

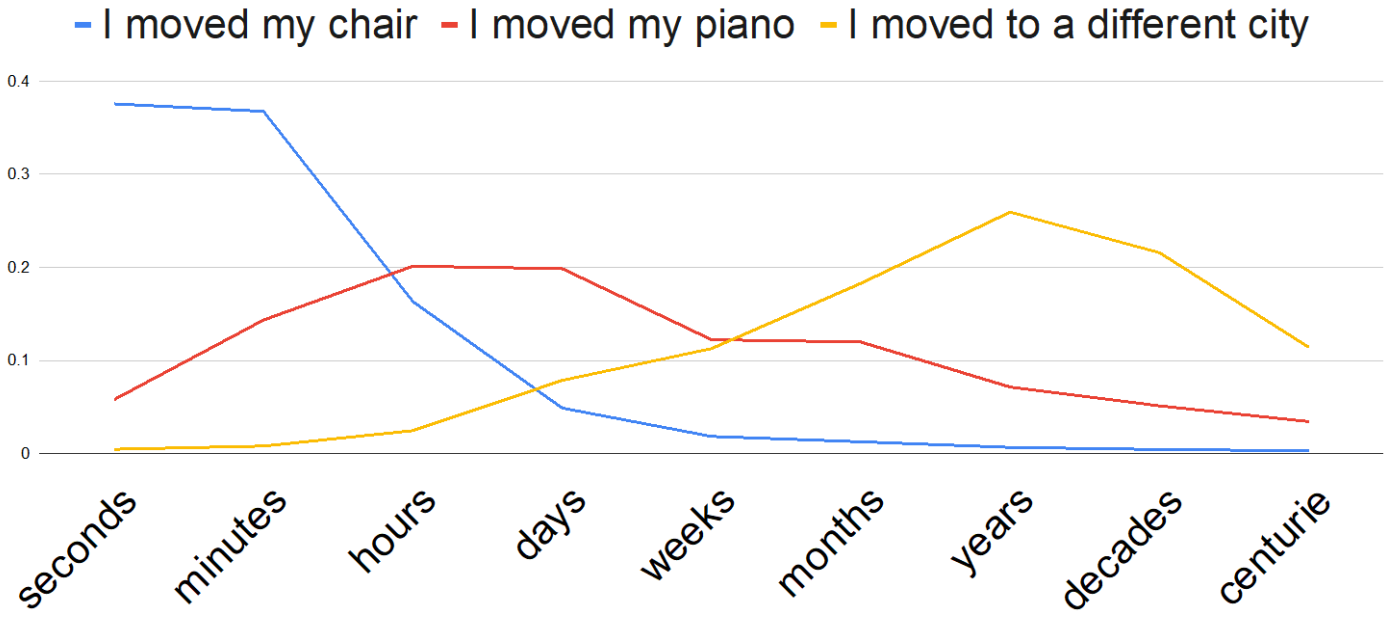
Group of contrast questions

Modeling Temporal Common Sense

- Goal: acquire temporal commonsense knowledge
  - Duration, Frequency, Typical time
  - Minimal supervision
  - Maximum generalization
- It is challenging:
  - How long does “move” take?
    - Highly contextual
  - How long does “I moved to a different city” take?
    - Needs more than direct event arguments
  - Which one is longer? Moving my chair or moving my piano?
    - Hard to categorize/type the argument in the temporal dimension
  - Do people often write how long they brushed their teeth in text?
    - Reporting biases
- Our view: we need to model distributions of temporal properties of events in fine grained contexts

Modeling Temporal Common Sense

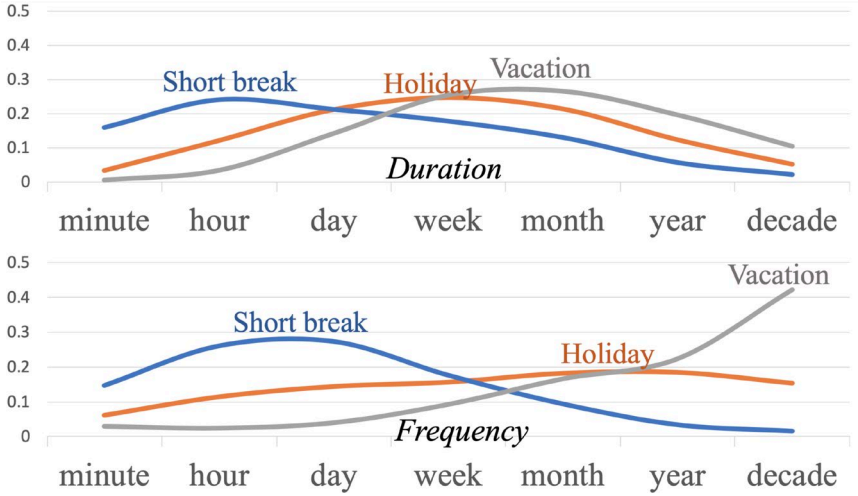
- How long does it take to move?
  - A real example where we model duration as distribution over units



- Our view: we need to model distributions of temporal properties of events in fine grained contexts

Technical highlights

- 1) Unsupervised direct/auxiliary signals collection
  - Using patterns from free text
- 2) Joint model across interrelated dimensions from 1)
  - 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”
  - Exist natural constraints: duration <= 1/frequency



Technical highlights

- 1) Unsupervised direct/auxiliary signals collection
  - Using patterns from free text
- 2) Joint model across interrelated dimensions from 1)
  - 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”
  - Exist natural constraints: duration <= 1/frequency
- 3) We employ several techniques in our model
  - Force continuous relationships between labels (seconds < week < year)
  - Reduce reporting biases on labels (”seconds” is mentioned 7 times less than “years”)
- (Over) simply put:
  - We train a BERT that is aware of time in a more unbiased way
  - With the pre-trained top layers, the model can directly predict temporal attributes of events

1) Unsupervised Signal Collection

- Representation of events: verb in a sentence
- High-precision patterns (with SRL)
  - Duration: [Event] for [Timex]
  - Frequency: [Event] every [Timex]
  - Typical Time: [Event] on [DayOfWeek]
  - Duration Upperbound: [Event] in [Timex]
  - Hierarchy [Event] while [Event]
- Labels:
  - Units (seconds, … centuries)
  - Temporal keywords (Monday, January, …)
- Run on entire Gigaword
  - 76M sentences, 25M pattern-selected, 4.3M filtered

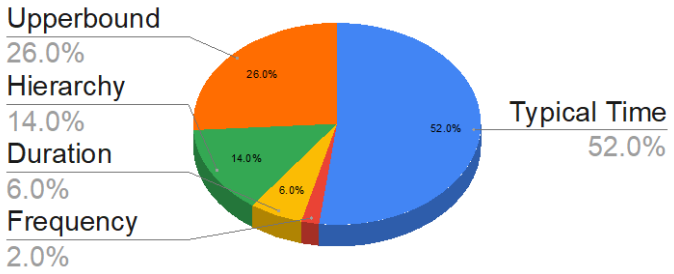
Jack rested for 2 hours before the speech.
(*Jack rested before the speech*, hours, **duration**)

We went to a bar last Friday.
(*We went to a bar*, Friday, **typical**)

Jane makes breakfast for herself everyday.
(*Jane makes breakfast for herself*, days, **frequency**)

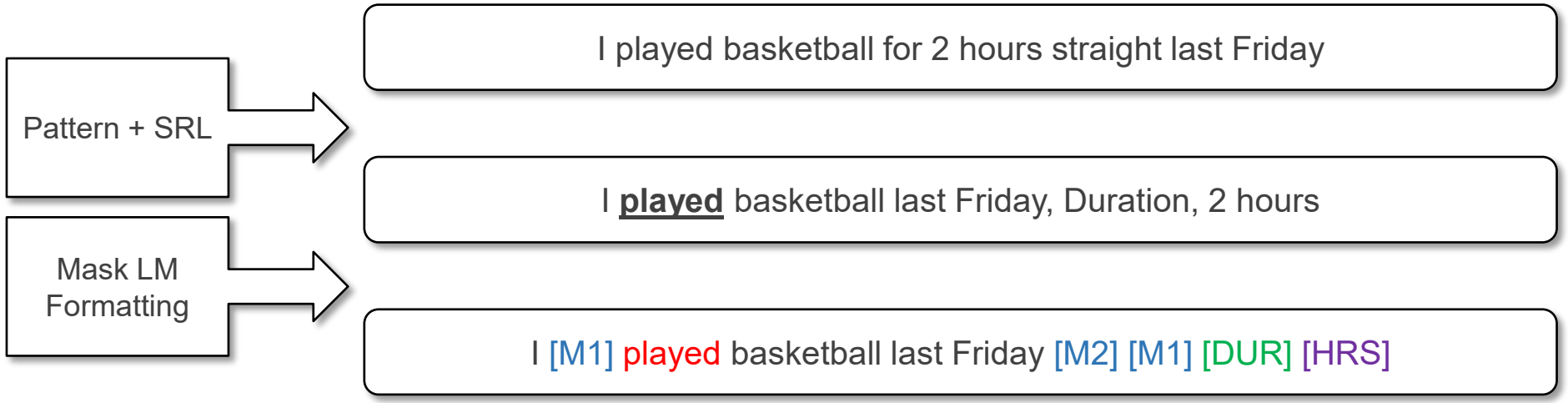
The city was surrounded by police yesterday.
(*The city was surrounded by police*, days, **upper-bound**)

The phone rang while I was in the bathroom.
(*The phone rang*, while I was in the bathroom, **hierarchy**)



2) Joint Model with Masked LM

- Consider [Event] [Value] [Dimension] tuples in each instance
- [E1, E2, … M1, ET … En, M2, M1, D, V]
  - Where E1, E2,… En are tokens in the event, ET is the trigger, M1, M2 are special markers, D is a marker for each dimension, V is a marker for the value of the dimension
- With an example:



2) Joint Model with Masked LM

I [M1] played basketball last Friday [M2] [M1] [DUR] [HRS]

- Main objective: mask some tokens and recover them
- How we mask:
  - With some probability, mask temporal value while keeping others

    I [M1] played basketball last Friday [M2] [M1] [DUR] **[MASK]**
  - Otherwise, mask a certain portion of E1...En while keeping temporal value unchanged

    [MASK] [M1] **[MASK] [MASK]** last Friday [M2] [M1] [DUR] [HRS]
  - Max (P(E|D,V) + P(V|E,D))
- Benefits:
  - Jointly learn **one** transformer towards **all** dimensions
  - Labels play a role in the transformer
  - One event may contain more than one (D + V), thus the model learns dimension relationships

    I [M1] played basketball **last Friday** [M2] [M1] [DUR] **[MASK]**
    - Last Friday -> Less than 24 hours -> [HRS]

3) Additional Techniques

I [M1] played basketball last Friday [M2] [M1] [DUR] [HRS]

- 1: Soft cross entropy for recovering V
  - If gold label is “hours”, the label vector **y** for “minutes, hours, days” will be [0.16, 0.47, 0.25] instead of [0.0, 1.0, 0.0] when computing cross entropy

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

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

- 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

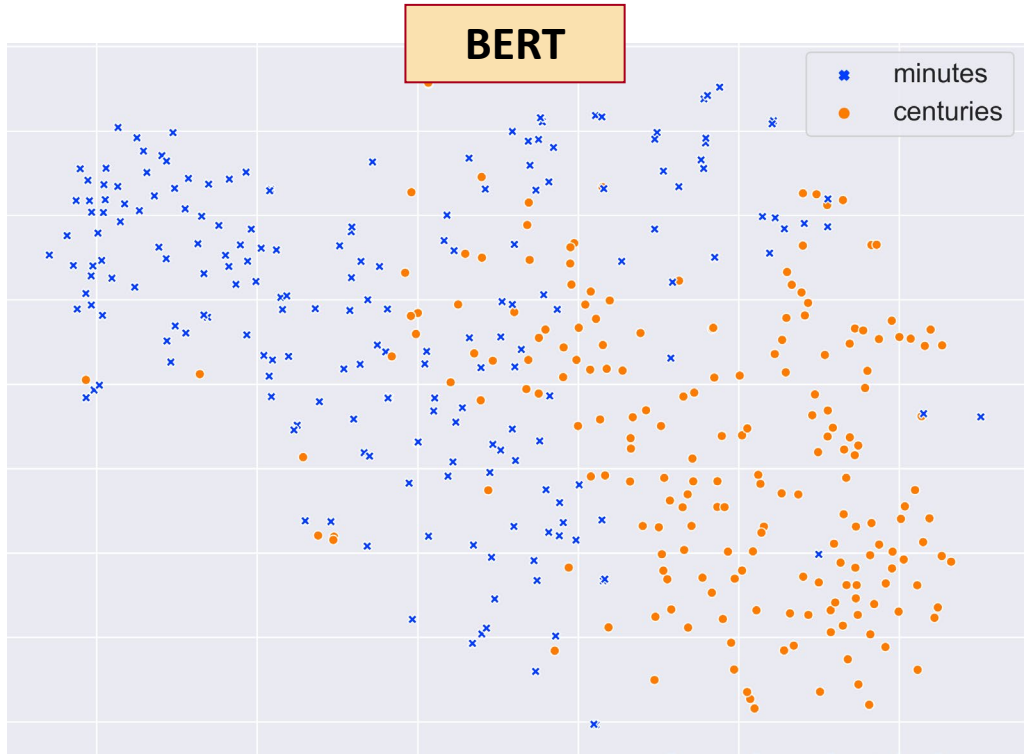
-> MASK = coffee, because “cup”

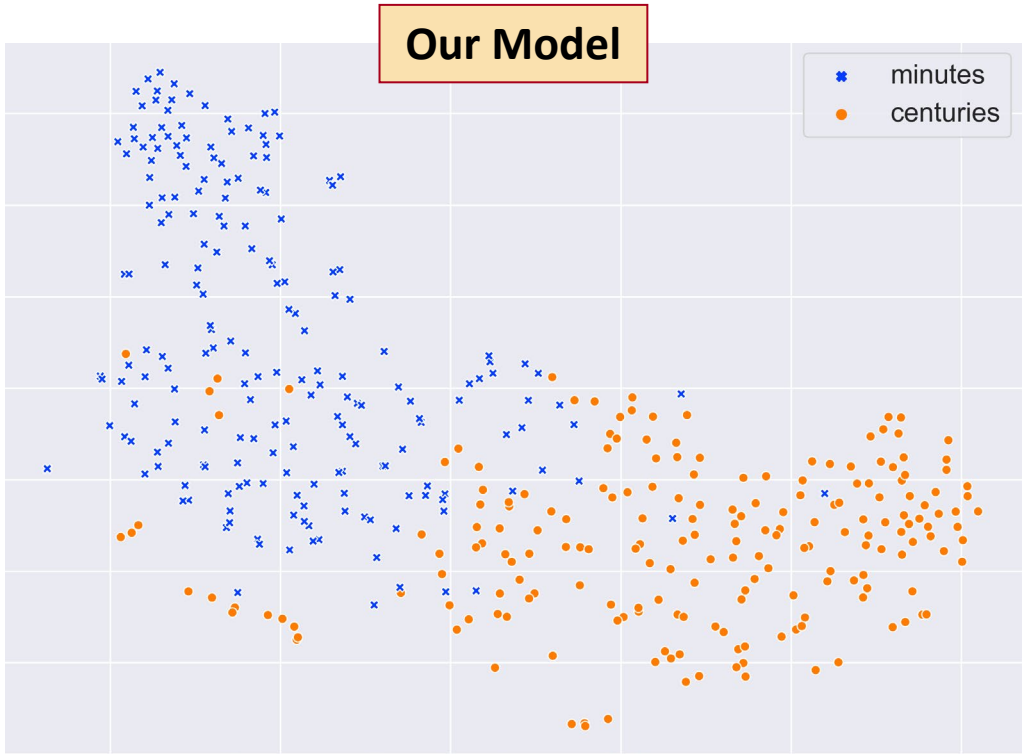
I [M1] had a cup of **[MASK]** last Friday [M2] [M1] [TYP] [Evening]

I [M1] had **[MASK] [MASK]** of **[MASK]** last Friday [M2] [M1] [TYP] [Evening]

Evaluation: Embedding Space

- A collection of events with duration of “minutes” or “centuries” (two 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 (Quantitively)

- Metric: Distance to gold label.
  - Dist (seconds, hours)=2, Dist (minutes, hours)=1
- RealNews: no document overlap, label generated through our patterns
- UDS-T: Vashishtha 2019, duration only
- TmpBERT: A naïve baseline that only uses soft cross entropy on top of BERT

Systems	**RealNews**						**UDS-T**
	Duration	Freq	Typical Time				Duration
			Day	Week	Month	Season	
BERT	1.09	1.68	1.75	1.53	3.78	0.87	1.84
TmpBERT	1.21	1.45	1.47	1.28	3.28	1.08	2.06
Ours	**0.60**	**1.17**	**1.68**	**1.36**	**2.75**	**0.67**	**1.77**
Ours (normalized)	7%	13%	21%	19%	23%	17%	20%

Evaluation: Extrinsic (HiEVE)

- Task: Identify event-event hierarchical relations
- Model: sentence pair classification, with triggers marked by special markers.
- C-P: Child-Parent relation
  - A bomb exploded and killed 6 civilians. This is the sixth accident since the war started.

Systems F1	NoRel	Coref	C-P	P-C
BERT	90.5	47.9	40.7	40.6
Ours	**91.2**	**51.7**	**49.6**	**48.3**

- Temporal common sense contributes to this task

*U.S. President Barack Obama **sparred** with Russia's Vladimir Putin over how to end the **war** in Syria on Monday during an icy **encounter** at a G8 **summit**. **Speaking** after **talks** with Obama, Putin **said** Moscow and Washington agreed the **bloodshed** must stop...*