Spatial and Temporal Language Understanding: Representation, Reasoning, and Grounding

Part II: IE, QA, and Reasoning of “Time”

Parisa Kordjamshidi, Michigan State University, USA, kordjams@msu.edu

Marie-Francine Moens, KU Leuven, Belgium, sien.moens@cs.kuleuven.be

James Pustejovsky, Brandeis University, USA, jamesp@cs.brandeis.edu

Qiang Ning, Amazon, USA, qiangning.01@gmail.com

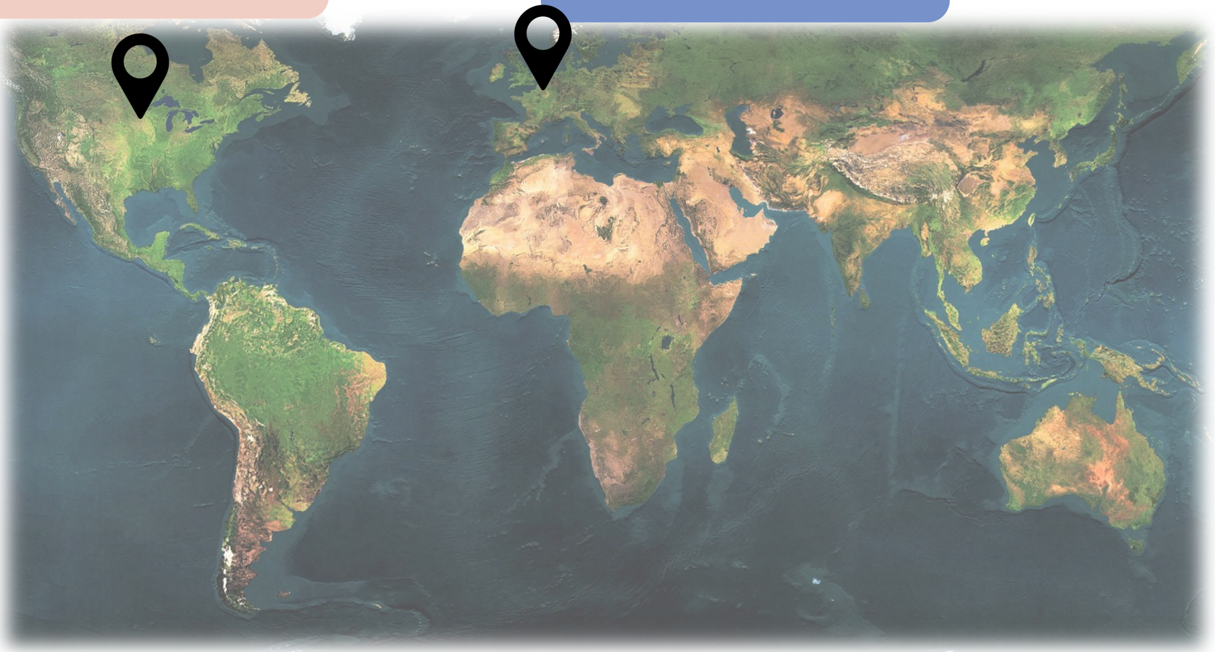
NAACL

Jun 17, 2024

Location is important for understanding events

People were angry

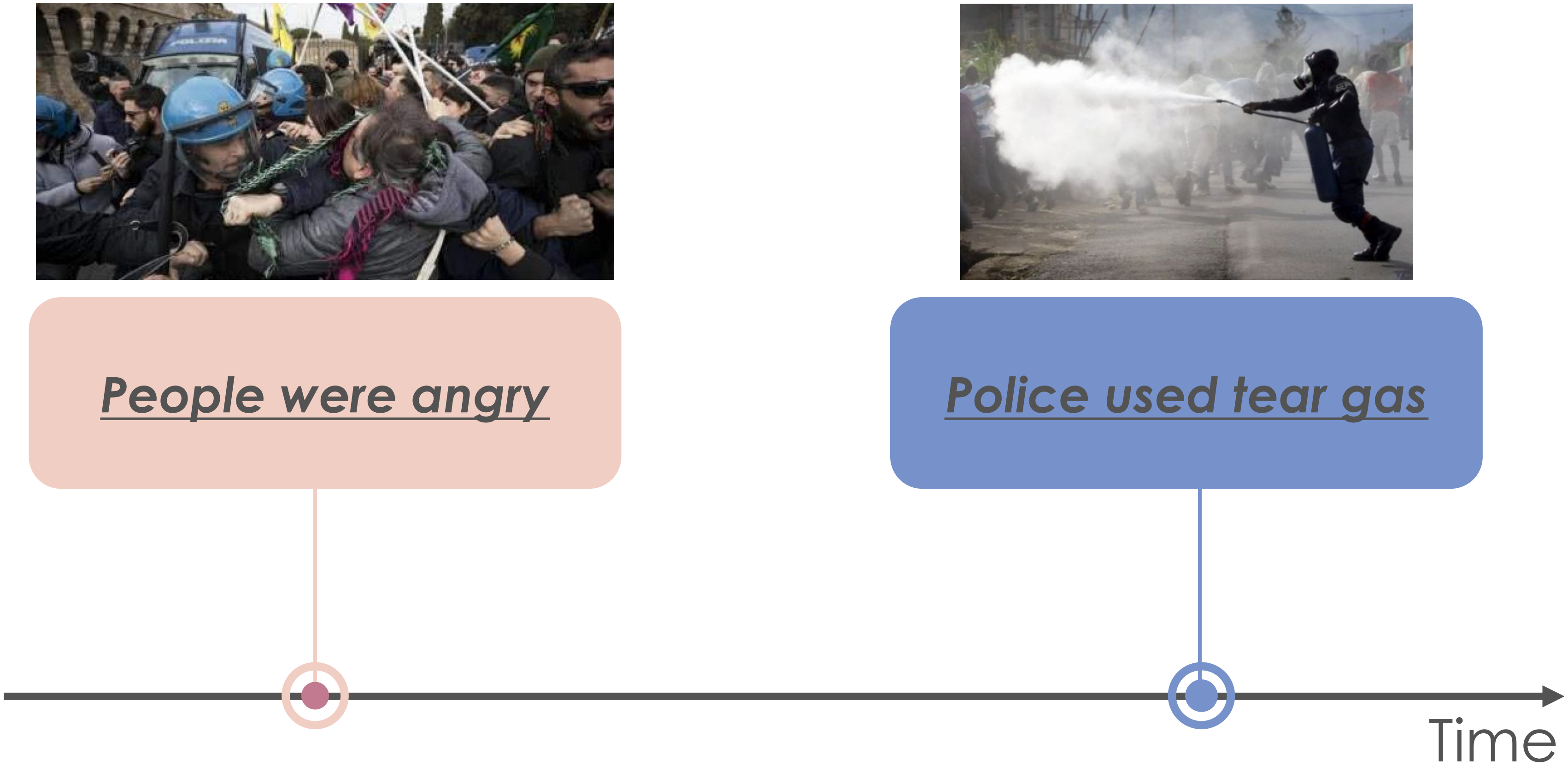
Police used tear gas



There’s often no relationship between these two events if they’re from different places in the world

Pin icon credits to Artdabana@Design from the Noun Project

Time is important for understanding events



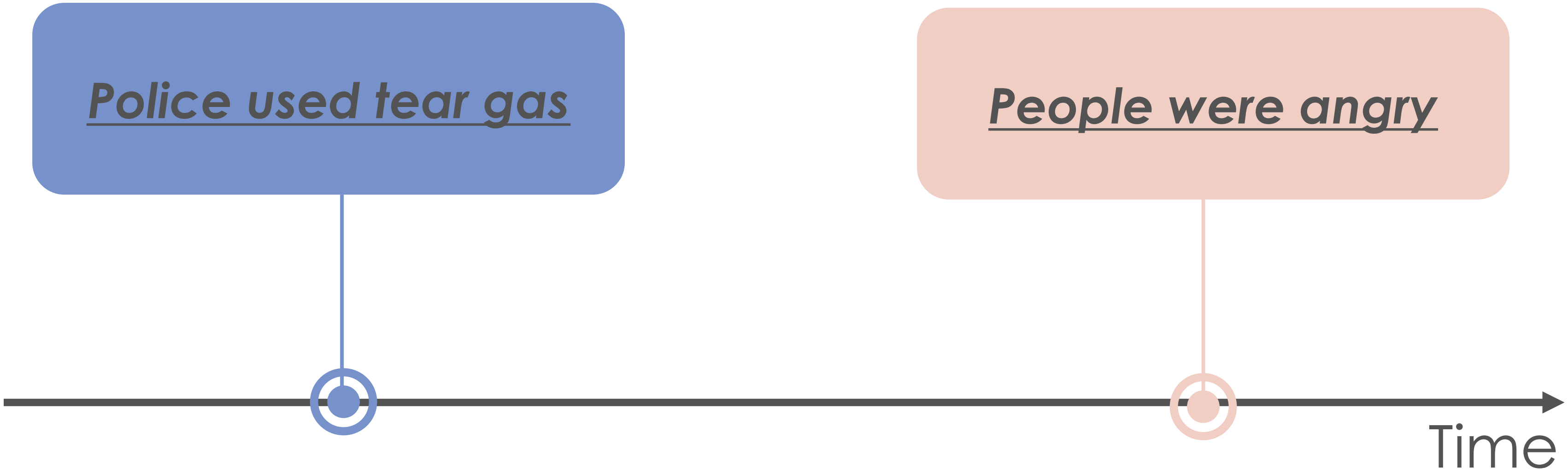
People were angry

Police used tear gas

Time

*People **were angry** first (likely causing chaos), and then the police **used tear gas** (to restore order).*

Same events, different temporal ordering



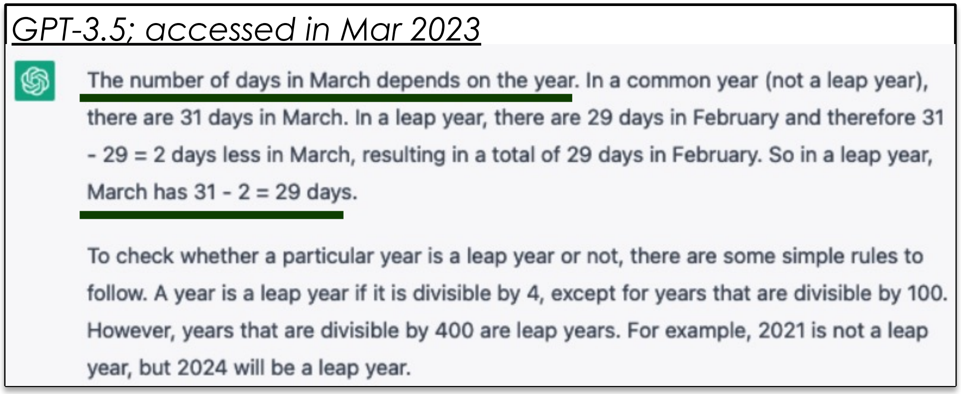
Police used tear gas

People were angry

Police ***used tear gas*** *(during a conflict), and then people* ***were angry*** *(because of the action police took).*

Examples where "time" plays an important role

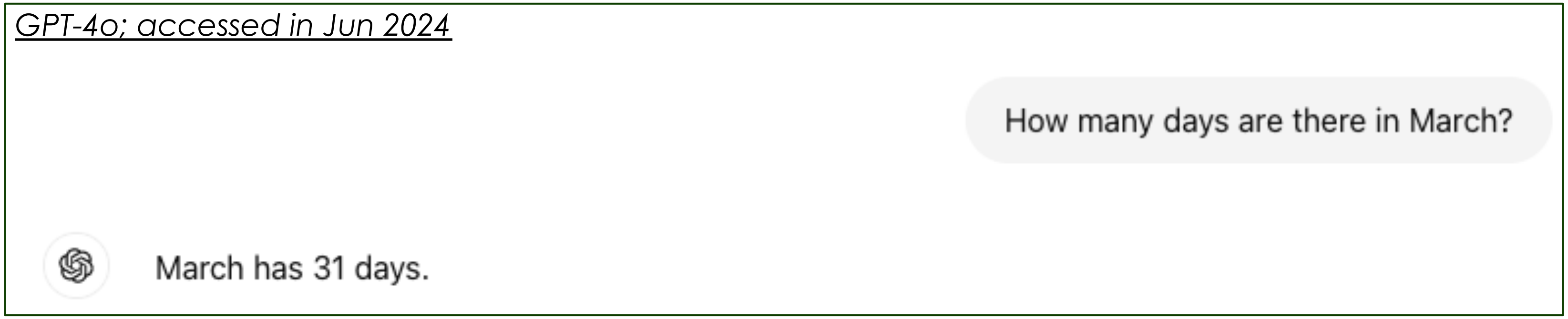
- **How many days are there in March?**



GPT-3.5; accessed in Mar 2023

The number of days in March depends on the year. In a common year (not a leap year), there are 31 days in March. In a leap year, there are 29 days in February and therefore 31 - 29 = 2 days less in March, resulting in a total of 29 days in February. So in a leap year, March has 31 - 2 = 29 days.

To check whether a particular year is a leap year or not, there are some simple rules to follow. A year is a leap year if it is divisible by 4, except for years that are divisible by 100. However, years that are divisible by 400 are leap years. For example, 2021 is not a leap year, but 2024 will be a leap year.



GPT-4o; accessed in Jun 2024

How many days are there in March?

March has 31 days.

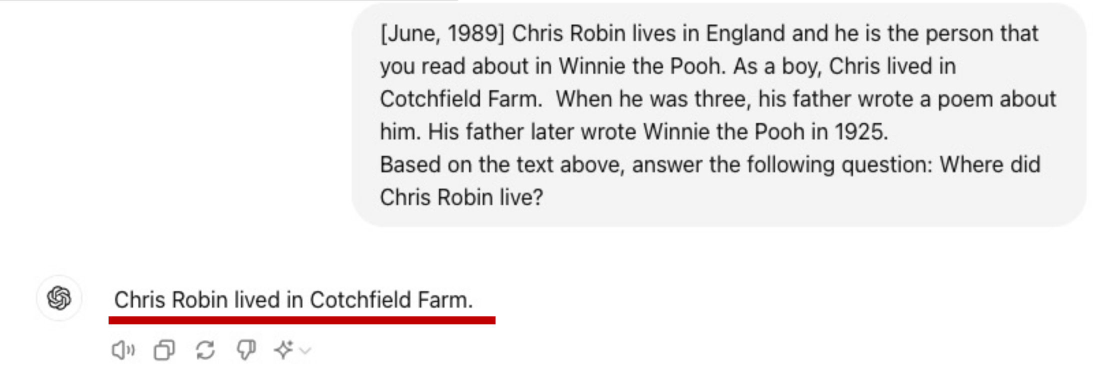
Examples where “time” plays an important role

- **[June, 1989] Chris Robin lives in England and he is the person that you read about in Winnie the Pooh. As a boy, Chris lived in Cotchfield Farm. When he was three, his father wrote a poem about him. His father later wrote Winnie the Pooh in 1925.**
  - ❑ Where did Chris Robin live?

Examples where "time" plays an important role

- **[June, 1989] Chris Robin lives in England and he is the person that you read about in Winnie the Pooh. As a boy, Chris lived in Cotchfield Farm. When he was three, his father wrote a poem about him. His father later wrote Winnie the Pooh in 1925.**
  - ❑ Where did Chris Robin live?
    - This is time sensitive: He lives in England (at least at the time of 1989); as a boy, he lived in Cotchfield Farm.

GPT-4o; accessed in Jun 2024



[June, 1989] Chris Robin lives in England and he is the person that you read about in Winnie the Pooh. As a boy, Chris lived in Cotchfield Farm. When he was three, his father wrote a poem about him. His father later wrote Winnie the Pooh in 1925.
Based on the text above, answer the following question: Where did Chris Robin live?

Chris Robin lived in Cotchfield Farm.

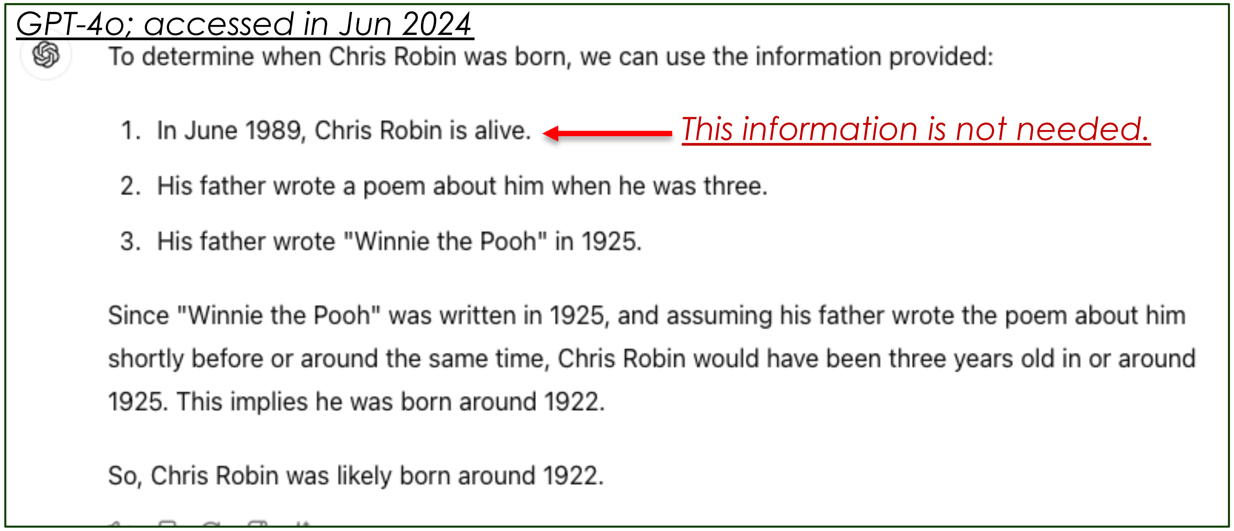
Examples where “time” plays an important role

- **[June, 1989] Chris Robin lives in England and he is the person that you read about in Winnie the Pooh. As a boy, Chris lived in Cotchfield Farm. When he was three, his father wrote a poem about him. His father later wrote Winnie the Pooh in 1925.**
  - ☐ When was Chris Robin born?

Examples where "time" plays an important role

- **[June, 1989] Chris Robin lives in England and he is the person that you read about in Winnie the Pooh. As a boy, Chris lived in Cotchfield Farm. When he was three, his father wrote a poem about him. His father later wrote Winnie the Pooh in 1925.**
  - When was Chris Robin born?
    - This requires reasoning: Chris Robin was at least 3 years old in 1925, so he was likely born **before** 1922.

GPT-4o; accessed in Jun 2024



To determine when Chris Robin was born, we can use the information provided:

1. In June 1989, Chris Robin is alive. ← *This information is not needed.*
2. His father wrote a poem about him when he was three.
3. His father wrote "Winnie the Pooh" in 1925.

Since "Winnie the Pooh" was written in 1925, and assuming his father wrote the poem about him shortly before or around the same time, Chris Robin would have been three years old in or around 1925. This implies he was born around 1922.

So, Chris Robin was likely born around 1922.

“Time” is a broad concept

- Many types of NLP tasks require certain level of understanding of time.
- Examples that the community have focused on:
  - Time expression understanding: “end of the year” = ?
  - Relation extraction: “Event_A” →**before**→“Event_B”
  - QA: What happened **after** Roosevelt died?
  - Knowledge graphs: (Barack Hussein Obama, President of, USA, **[2009, 2017]**)
  - …

A structured view of the NLP literature about "time"

	Information extraction	Question Answering	Reasoning
Intrinsic to "time" (i.e., targeting a better understanding of time)	Time expressions Temporal relations Temporal grounding	Temporal order QA	Temporal relation graph inference Temporal logic
Extrinsic to "time" (i.e., solving tasks that require consideration of time)	Timeline summarization Temporal knowledge graphs Causality	Situated QA Time-sensitive QA Temporal QA over knowledge graphs	Robotic control

- *Purpose is to help understand content in this talk*
- *Not the only way to view the literature*
- *Not exhaustive*

Scope of the talk

	Information extraction	**Question Answering**	**Reasoning**
Intrinsic to "time" (i.e., targeting a better understanding of time)	Time expressions Temporal relations Temporal grounding	Temporal order QA	Temporal relation graph inference Temporal logic
Extrinsic to "time" (i.e., solving tasks that require consideration of time)	Timeline summarization Temporal knowledge graphs Causality	Situated QA Time-sensitive QA Temporal QA over knowledge graphs	Robotic control Trip planning

Main papers to cover in this talk

	Information extraction	**Question Answering**	**Reasoning**
Intrinsic to "time" (i.e., targeting a better understanding of time)	SteQE[1]	TORQUE[2]	NL2TL[3]
Extrinsic to "time" (i.e., solving tasks that require consideration of time)	Timeline summarization Temporal knowledge graphs Causality	Situated QA Time-sensitive QA Temporal QA over knowledge graphs	AutoTEMP[4]

[1] A Meta-framework for Spatiotemporal Quantity Extraction from Text. Ning et al., ACL'22.
[2] TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. Ning et al., EMNLP'20.
[3] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP'23.
[4] AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. Chen et al., ICRA'24.

Start with annotation schemes of “time”

- Two key concepts for annotating “time” are
  - Event mentions (E)
  - Time points (T)

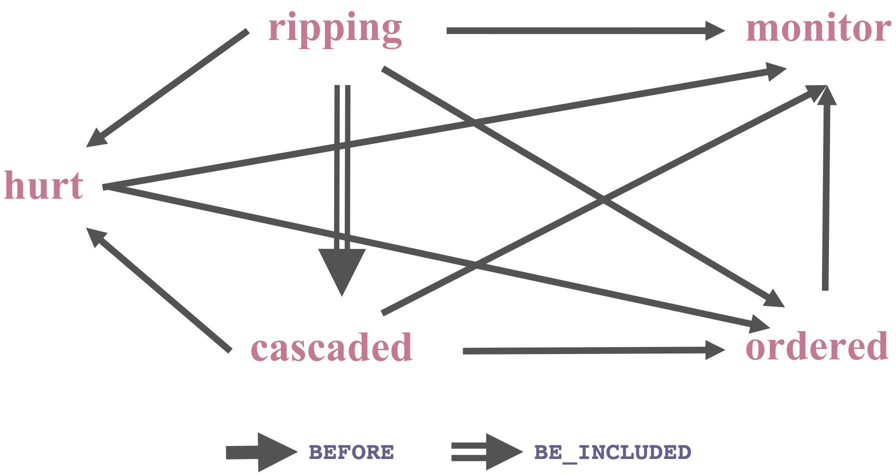
- We can view the existing annotation schemes of “time” based on how they deal with E and T
  - Focus on (E,T) relationship
  - Focus on (E,E) relationship
  - Both (E,T) and (E,E) relationships

TimeML

- ISO-TimeML: An international standard for temporal annotation
  - TLINKs
    - (E,T): “Link events to the document time and to local time expressions”
    - (E,E): “Order event relative to other events”
  - I’ll give a talk on 6/16.

*In Los Angeles that lesson was brought home Friday 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.*

Event-Event (E,E) relations form a graph



TimeML

- ISO-TimeML: An international standard for temporal annotation
  - TLINKs
    - (E,T): “Link events to the document time and to local time expressions”
    - (E,E): “Order event relative to other events”
  - TimeBank[1][2]
    - The classic dataset using TimeML
    - A major driving force for advancement in the last two decades

[1] TimeBank: https://catalog.ldc.upenn.edu/LDC2006T08
[2] TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. UzZaman et al., SemEval 2013.

Focusing on (E,T) relationship -- EventTime 2016[1]

- "Annotation complexity": Quadratic vs linear
  - If there're N events, there're O(N^2) pairs of (E,E) relations, but there're only O(N) time points to consider for (E,T) relations.
  - Advantage: the time of an event doesn't have to come from the text verbatim.

> He was **sent** into space on **May 26, 1980**. He **spent** six days aboard the Salyut 6 spacecraft.

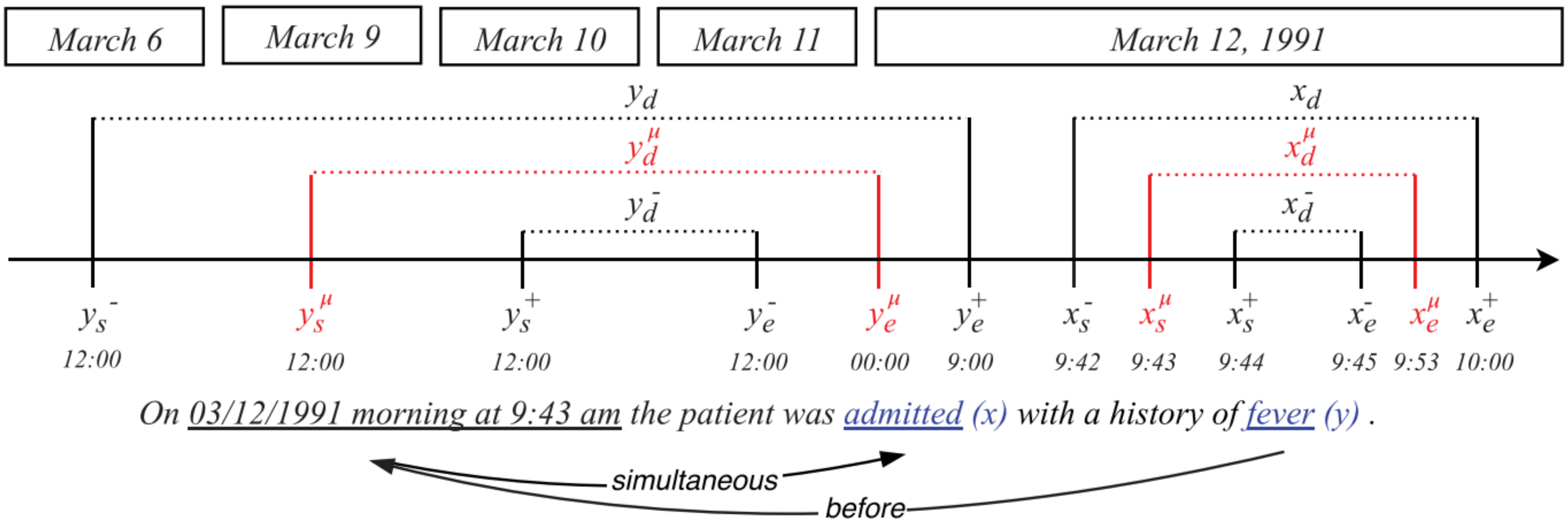
Annotations:

- *Sent: 1980-05-26*
- *Spent: [1980-05-26, 1980-06-01] (this time interval is not from the text verbatim)*

---

[1] Temporal Anchoring of Events for the TimeBank Corpus. Reimers et al., ACL'16.

Focusing on (E,T) relationship -- AbsoluteTimeline 2022[1]



Used time boundaries to represent (reasonable) guesses when text comes with uncertainty. E.g., “a history of fever” is assumed to be in the several days before March 12, 1991, not weeks or months.

---

[1] Towards extracting absolute timelines from English clinical reports. Leeuwenberg & Moens, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 2022.

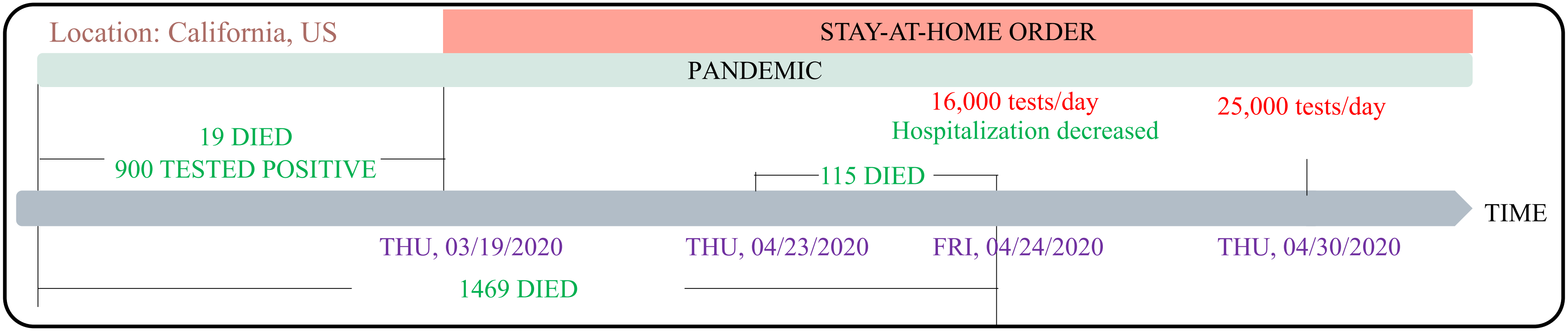
SteQE 2022: when COVID happened

CNN [March 20, 2020]

The nation’s most populous state is ordering its nearly 40 million residents to stay home to prevent the spread of the deadly coronavirus. Nineteen people have died and more than 900 tested positive for coronavirus in California. California Gov. Gavin Newsom's order marks the first statewide mandatory restrictions in the United States to help combat the outbreak. It went into effect at midnight Thursday, meaning Californians should not leave home except for essential things such as food, prescriptions, health care and commuting to jobs considered crucial.

NYT [April 24, 2020]

Over the previous 24 hours, he said, California saw 115 Covid-19 deaths…That brought the total number of those lost to 1,469…Mr. Newsom noted that the number of hospitalizations and patients in intensive care had decreased slightly from the day before…As of this week, Mr. Newsom said that, on average, 16,000 Covid-19 tests were conducted per day across the state. By the end of the month, he said officials are planning for that to increase to 25,000 per day.



Location: California, US

STAY-AT-HOME ORDER

PANDEMIC

19 DIED
900 TESTED POSITIVE

16,000 tests/day
Hospitalization decreased

25,000 tests/day

115 DIED

TIME

THU, 03/19/2020

THU, 04/23/2020

FRI, 04/24/2020

THU, 04/30/2020

1469 DIED

A Meta-framework for Spatiotemporal Quantity Extraction from Text. Ning et al., ACL’22.

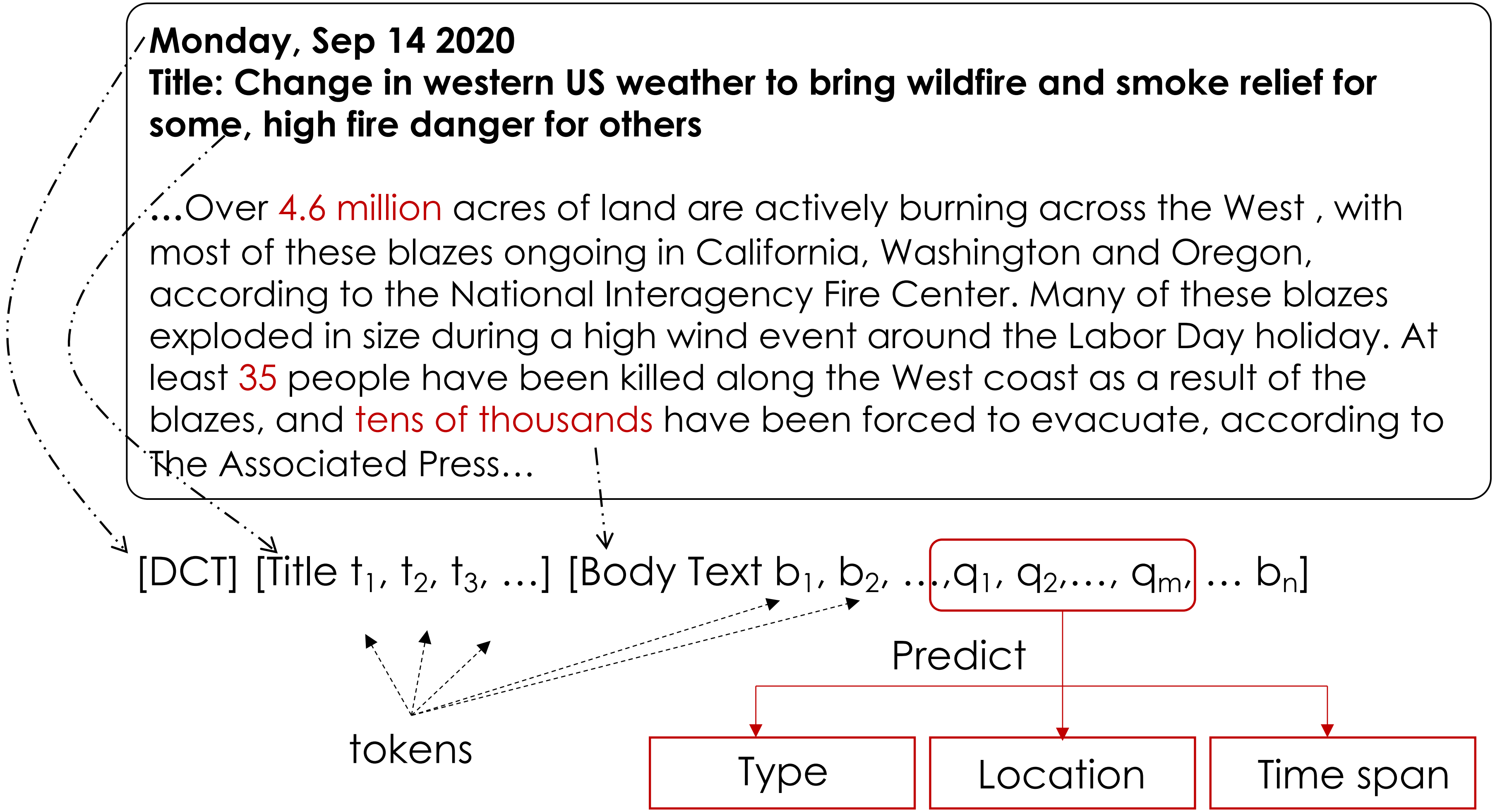
SteQE: Spatiotemporal Quantity Extraction

- This paper proposes a **meta-framework** for studying **spatiotemporal** phenomena of **quantity events**.
  - Quantity events: numbers associated with events
  - Spatiotemporal: when and where did this event happen?
  - Meta-framework: problem formulation, re-usable data annotation pipeline, and baseline models
- SteQE: We release data from three sociopolitical events: the COVID-19 pandemic, Black Lives Matter (BLM) protests, and 2020 California wildfires.
- Original talk in 2022: https://www.youtube.com/watch?v=rNI2Jcu0NgA

Data Preview

Document	Type	Spatial Grounding	Temporal Grounding
Document Creation Time: Sat, 2020-08-15 Tennessee has conducted 1,757,690 tests with **1,631,297** negative results	Test performed for COVID-19: result is negative	US, Tennessee	Overall number ends at DCT
Document Creation Time: Wed, 2020-08-12 Wyandotte County is reporting 4,895 confirmed cases...The county said on Tuesday that **99** people have died from the coronavirus since the start…	Deaths: definitely caused by COVID-19	US, Kansas, Wyandotte County	Overall number ends on 2020-08-11
Document Creation Time: Wed, 2020-08-12 ...large fires across 10 states...At least **35** people have died in California, Oregon and Washington.	People Impacted	US	Overall number ends at DCT
Document Creation Time: Tue, 2020-09-22 the blaze had more than doubled in size over the past week to 170 square miles (**440** square kilometers), ... from Los Angeles.	Physical measurements	US, California, Los Angeles	2020-09-15 to 2020-09-22
Document Creation Time: Tue, 2020-06-16 Black Lives Matter demonstrators in a tiny Ohio town...Sunday. The small demonstration has about **80** people, organized by local Bethel residents.	Number of participants in protests or relevant activities	US, Ohio, Bethel	2020-06-14 to 2020-06-14
Document Creation Time: Sun, 2020-05-31 A CNN analysis found about 80% of the **51** people booked into a Minneapolis jail during two days of protests are actually from Minnesota	Number of arrests due to the protests or following skirmishes	US, Minnesota, Minneapolis	Unknown

SteQE: Task Formulation



Monday, Sep 14 2020
Title: Change in western US weather to bring wildfire and smoke relief for some, high fire danger for others

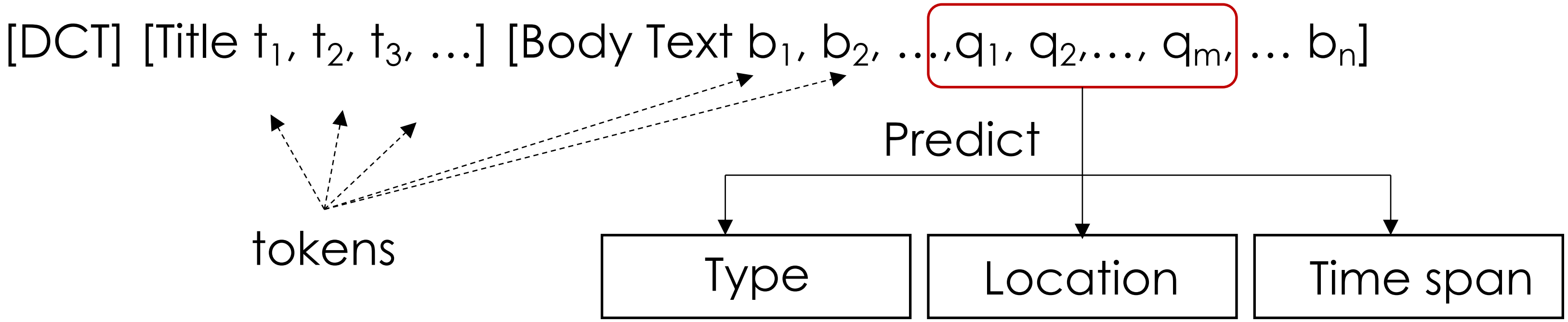
…Over 4.6 million acres of land are actively burning across the West , with most of these blazes ongoing in California, Washington and Oregon, according to the National Interagency Fire Center. Many of these blazes exploded in size during a high wind event around the Labor Day holiday. At least 35 people have been killed along the West coast as a result of the blazes, and tens of thousands have been forced to evacuate, according to The Associated Press…

[DCT] [Title $t_1, t_2, t_3, \ldots$] [Body Text $b_1, b_2, \ldots, q_1, q_2, \ldots, q_m, \ldots b_n$]

SteQE: Task Formulation

1. Quantity Recognition

- Similar to named entity recognition (NER), quantity recognition is defined as a text span detection problem.
  - A BERT model was shown to be very reliable system extracting quantities, so we took BERT predictions as gold.



[DCT] [Title t_1, t_2, t_3, …] [Body Text b_1, b_2, …, q_1, q_2,…, q_m, … b_n]

tokens

Predict

Type

Location

Time span

SteQE: Task Formulation

2. Quantity Typing

- Predefined label set (e.g., positive cases & hospitalization)
- A clear event type is important for subsequent spatiotemporal grounding, but some quantities have multiple types.
  - Single-typing: “*[three] men were hospitalized 5 days after being tested positive*”
  - We enforce single-typing by providing an order of importance. For instance, hospitalization is more important than tested positive, so the spatiotemporal extent of “three” will be that of hospitalizations.

SteQE: Task Formulation

3. Spatial Grounding

- To ground real-world events to a real-world locale
- Format
  - **choose** from a predefined set of questions to determine the country (U.S. vs non-U.S.) and state,
  - use **free text** for the name of the city
  - use **span selection** for more granular locale information (e.g., a pork plant)
- Granularity
  - the **most specific** location mentioned in the text that **contains all** individual cases of a quantity event

SteQE: Task Formulation

4. Temporal Grounding

- To ground each real-world quantity event to a <u>single time span</u>
- Format
  - A time span consists of **two time points**, and the key is the format for time points.
  - Select **calendar dates** from our UI (e.g., Feb 1st, 2021)
  - When a time point is unclear based on the text, we allow annotators to simply select “**Unknown**”

SteQE: Task Formulation

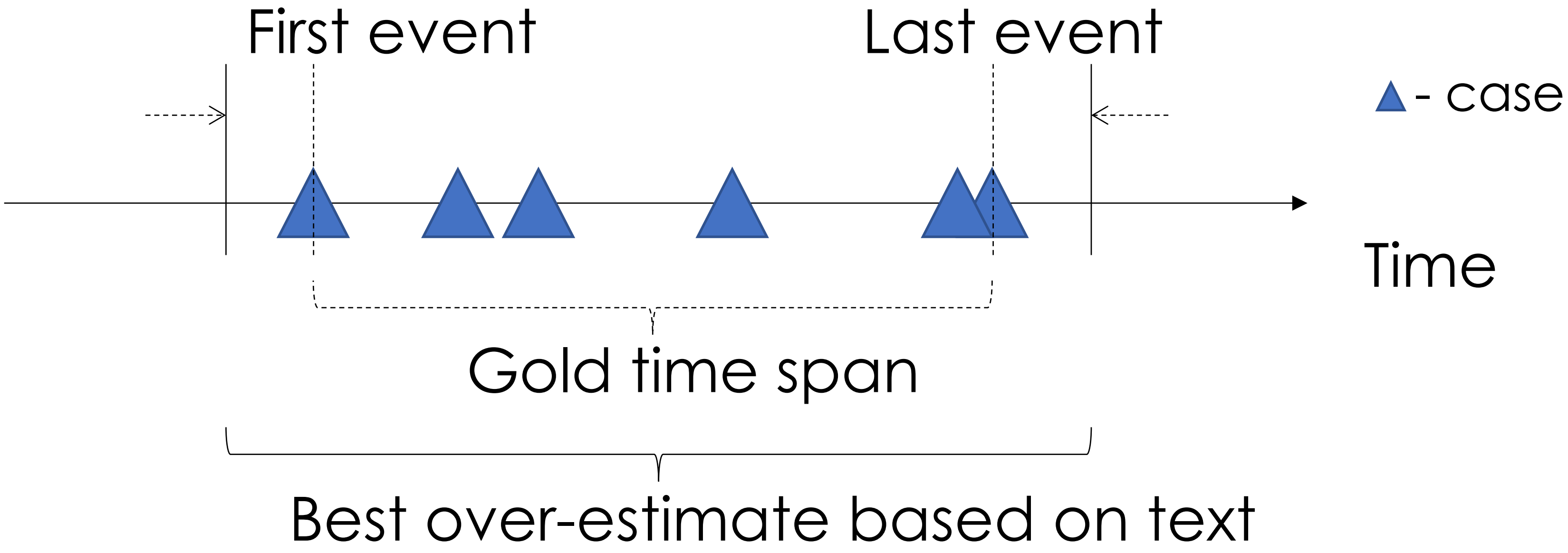
4. Temporal Grounding

- Overall quantity: We identified a special type of quantity – overall quantity – that needs special attention in practice
- “There have been [3 million] COVID-19 positive cases so far”. The start time is the beginning of the pandemic, but people do not always agree on when that was because:
  - the pandemic started at different times in different regions
  - one may argue that the pandemic started either since the first confirmed case, or since the lockdown.
- We allow “overall” as a label for the start time of a quantity.

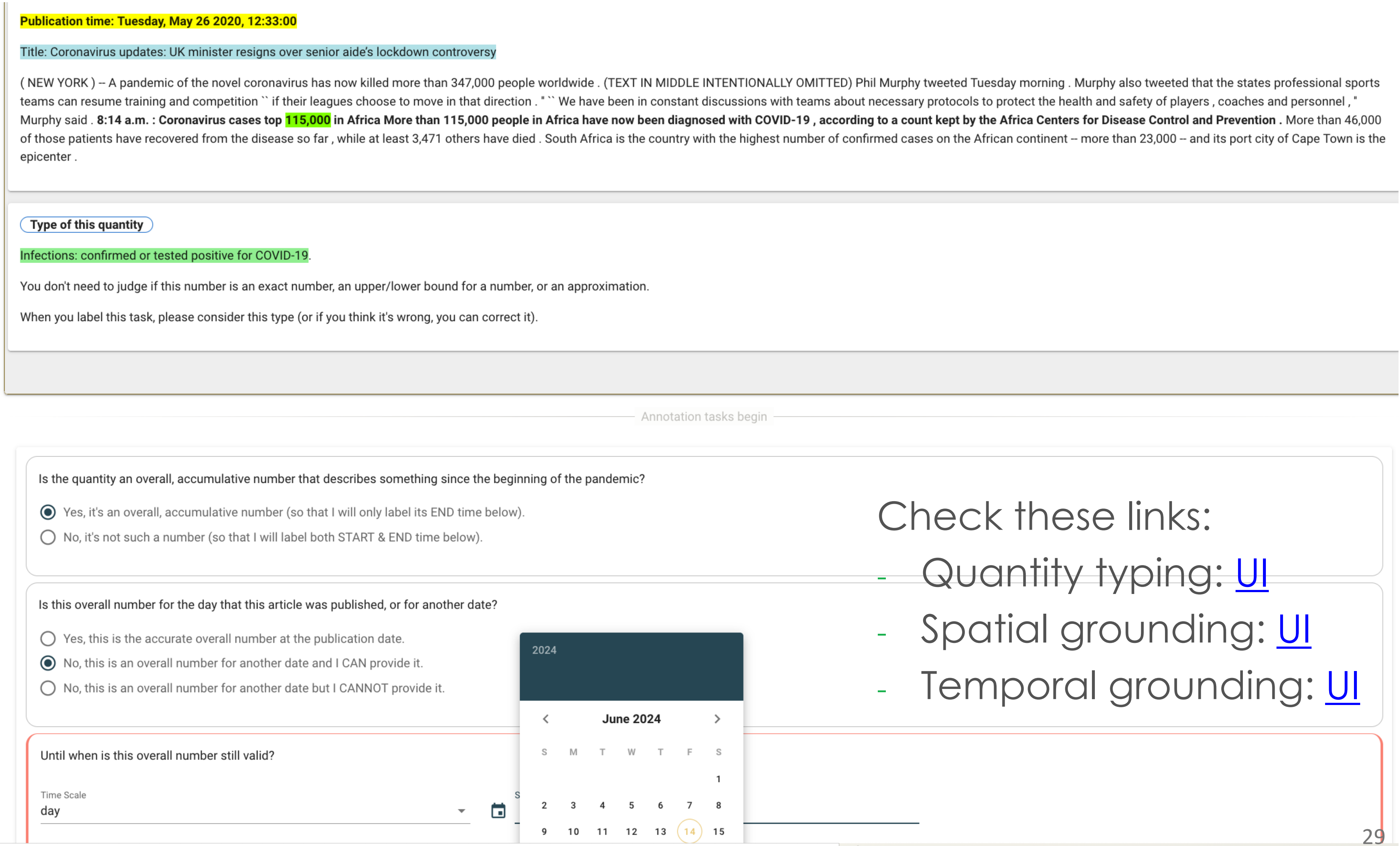
SteQE: Task Formulation

4. Temporal Grounding

- Granularity
  - Given the nature of news events, it is often enough to be specific up to “**day**”
  - The best **over-estimate** of the **gold** time span based on text



Annotation UI on Crowdaq



Publication time: Tuesday, May 26 2020, 12:33:00

Title: Coronavirus updates: UK minister resigns over senior aide's lockdown controversy

(NEW YORK) -- A pandemic of the novel coronavirus has now killed more than 347,000 people worldwide . (TEXT IN MIDDLE INTENTIONALLY OMITTED) Phil Murphy tweeted Tuesday morning . Murphy also tweeted that the states professional sports teams can resume training and competition `` if their leagues choose to move in that direction . " `` We have been in constant discussions with teams about necessary protocols to protect the health and safety of players , coaches and personnel , " Murphy said . **8:14 a.m. : Coronavirus cases top 115,000 in Africa More than 115,000 people in Africa have now been diagnosed with COVID-19 , according to a count kept by the Africa Centers for Disease Control and Prevention .** More than 46,000 of those patients have recovered from the disease so far , while at least 3,471 others have died . South Africa is the country with the highest number of confirmed cases on the African continent -- more than 23,000 -- and its port city of Cape Town is the epicenter .

Type of this quantity

Infections: confirmed or tested positive for COVID-19.

You don't need to judge if this number is an exact number, an upper/lower bound for a number, or an approximation.

When you label this task, please consider this type (or if you think it's wrong, you can correct it).

Annotation tasks begin

Is the quantity an overall, accumulative number that describes something since the beginning of the pandemic?

● Yes, it's an overall, accumulative number (so that I will only label its END time below).

○ No, it's not such a number (so that I will label both START & END time below).

Is this overall number for the day that this article was published, or for another date?

○ Yes, this is the accurate overall number at the publication date.

● No, this is an overall number for another date and I CAN provide it.

○ No, this is an overall number for another date but I CANNOT provide it.

Until when is this overall number still valid?

Time Scale: day

2024
June 2024
S M T W T F S
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
PowerPoint File Edit View Insert Format Arrange Tools Slide Show Window Help

Check these links:

- Quantity typing: UI
- Spatial grounding: UI
- Temporal grounding: UI

Data Statistics

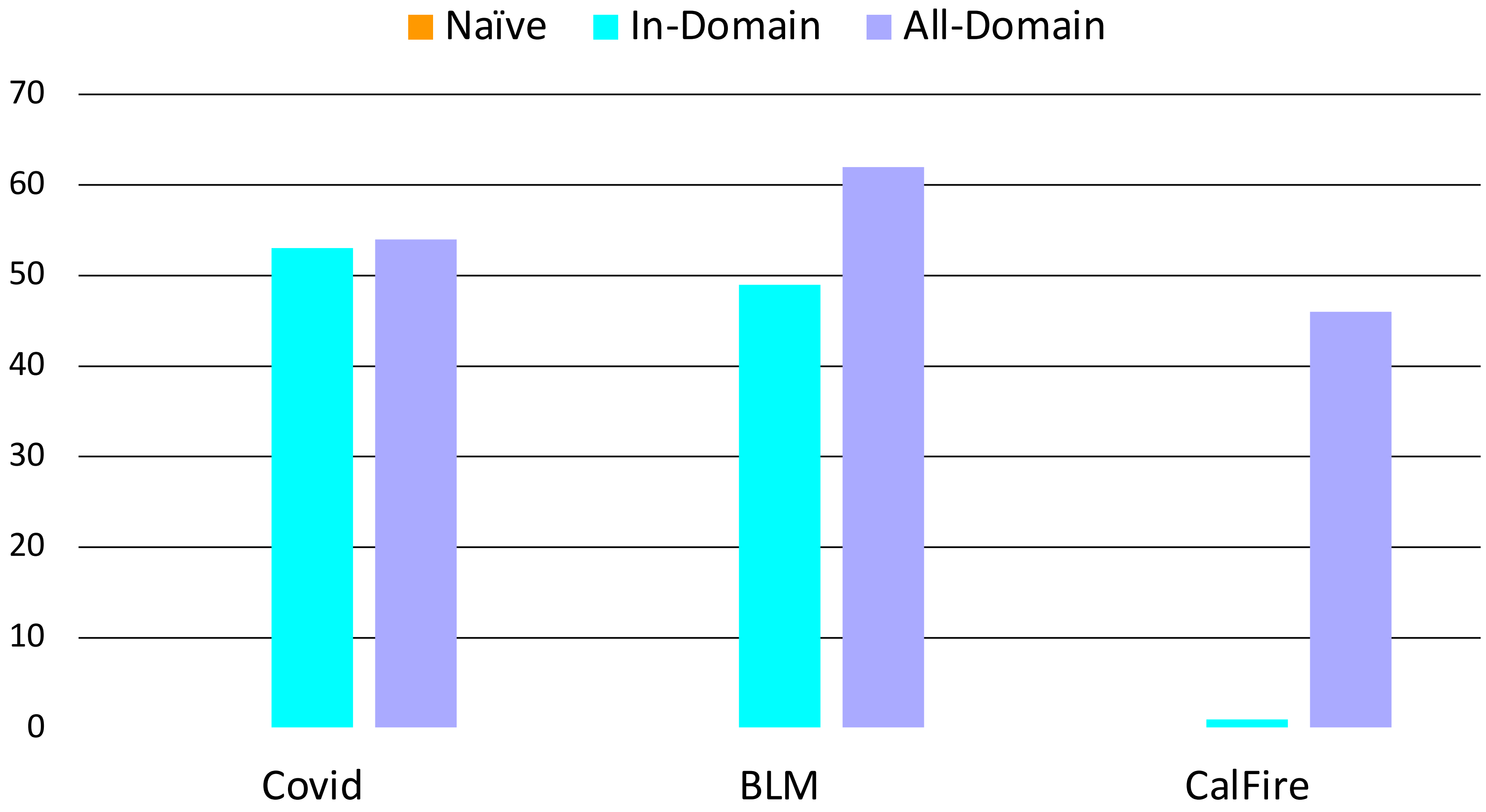
Task	Size			Quality	
	#Quant.	#Sent.	#Doc.	WAWA	Expert
Typ-COVID	1.5k	1.5k	1.3k	95%	100%
Typ-BLM	4k	3.9k	3k	87%	94%
Typ-Fire	2k	2k	1.4k	91%	96%
Spa-COVID	3.4k	3.3k	2.9k	91%	98%
Spa-BLM	1.5k	1.5k	1.2k	80%	96%
Spa-Fire	2k	2k	1.3k	92%	90%
Temp-COVID	4.3k	4.2k	3.5k	86%	100%
Temp-BLM	1.6k	1.6k	1.2k	77%	96%
Temp-Fire	1.6k	1.6k	1.1k	82%	96%

Experiments

- Metrics:
  - Spatial: Exact Match (**city** level)
  - Temporal: **Non-overall quantities (start/end time exact match)**
- Datasets: COVID-19, BLM, CalFire
  - Test set: Each domain has 500 fully annotated test quantities
- Setting:
  - Trained only on in-domain data vs on all 3 domains
- Model:
  - T5-large

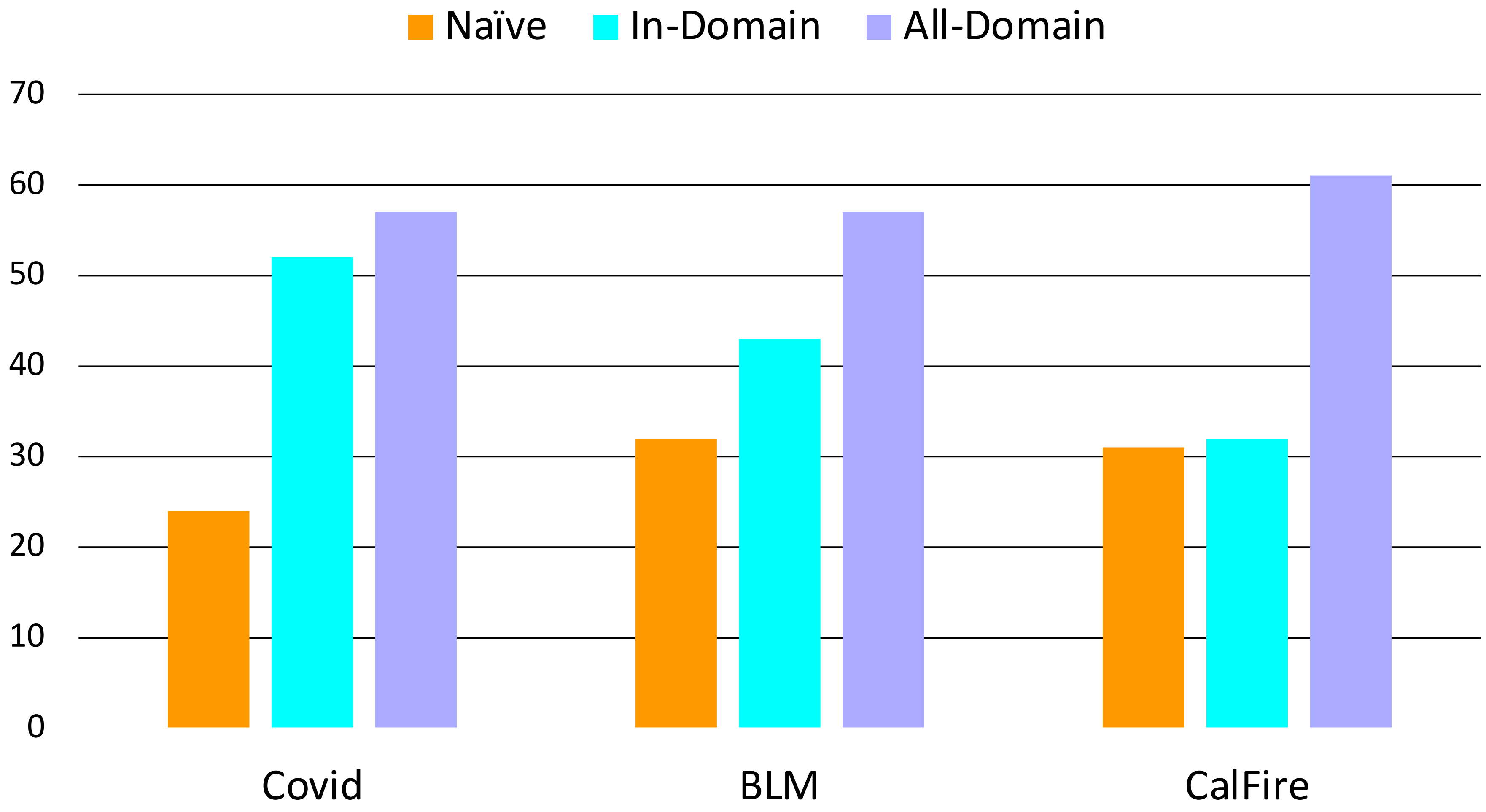
Temporal start time: exact match (EM)

Naïve: always predict the document creation time



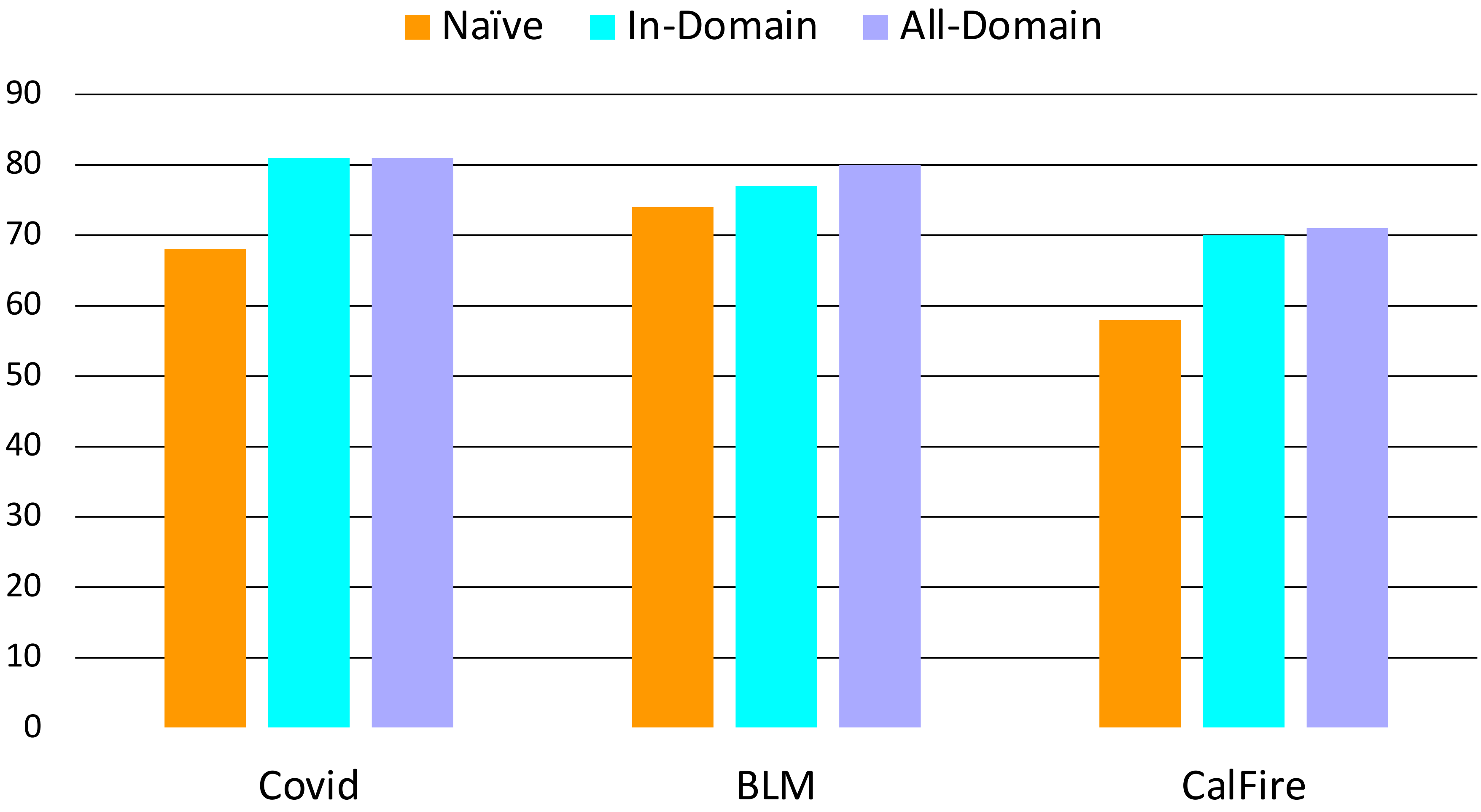
Temporal end time: EM

Naïve: choose the time expression closest to the quantity span



Results – Spatial EM (up to city-level)

Naïve: choose the nearest city mention of the quantity span

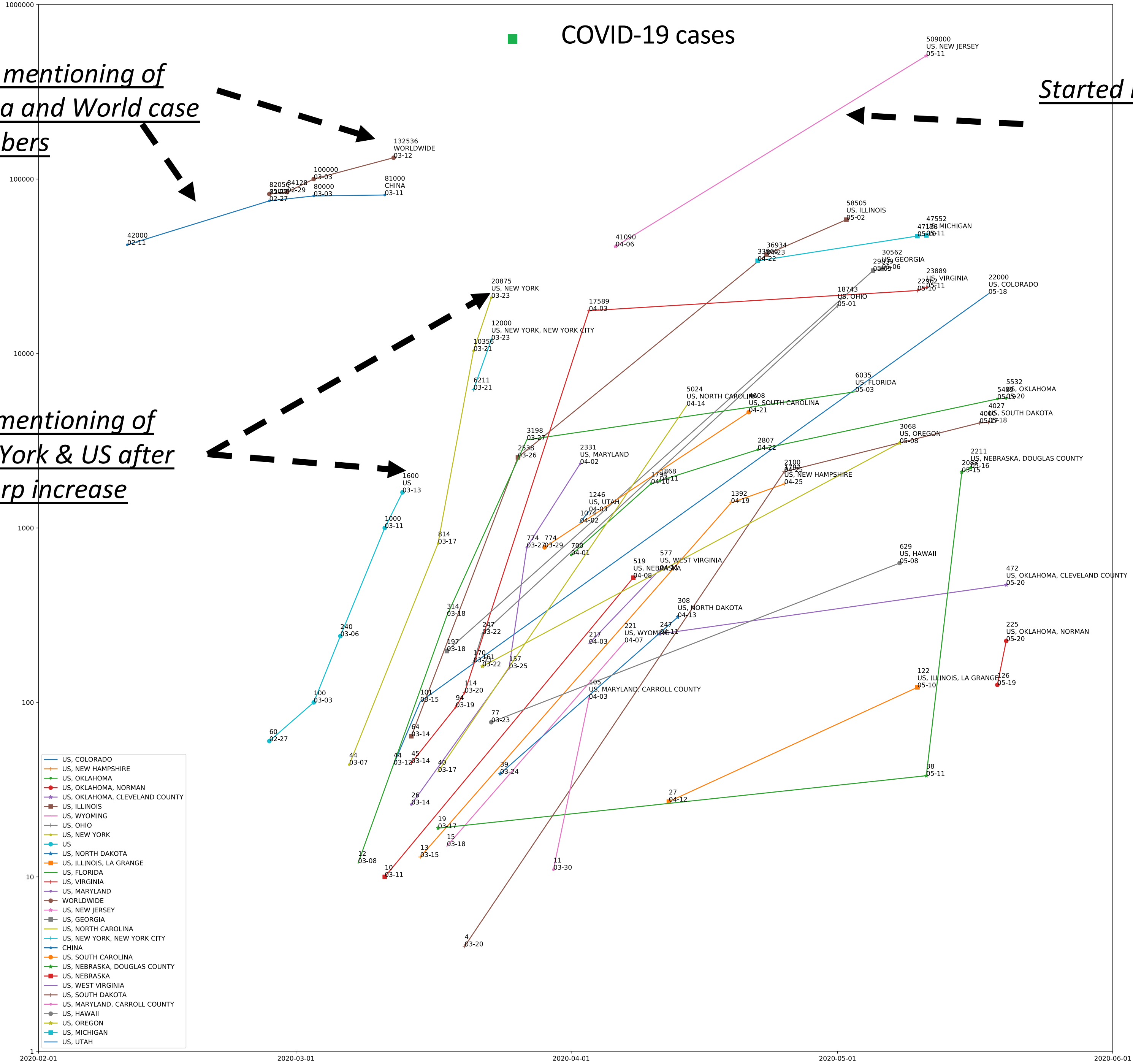


Information Verification

- Discrepancy between two sources
  - Source 1 (therapidian.org at May 11 2020, 19:43:00): …Statewide , there are 47,552 cases , 4,484 deaths , and 22,686 recoveries confirmed …
  - Source 2 (wzzm13.com at May 11 2020, 06:54:00): GRAND RAPIDS , Mich. -- … The total number of cases is 47,552 and 4,584 have died , according to the data from Michigan 's Dept…
- NYT data base: 4,584 total death in Michigan on 5/11/2020. Our hypothesis is that Source 1 made a typo.

Information Verification

- Discrepancy between two time points (same source)
  - **Accessed on 4/3/2020:** Pennsylvania has 7,016 cases of COVID-19 with 18 in Adams County, and 23 in Franklin County. There are 17,589 cases in Virginia with 121 in Loudoun County. West Virginia has 217 cases of coronavirus, 12 in Jefferson County, and 27 in Berkeley County.
  - **Accessed again on 5/8/2020:** Pennsylvania has 7,016 cases of COVID-19 with 18 in Adams County, and 23 in Franklin County. West Virginia has 217 cases of coronavirus, 12 in Jefferson County, and 27 in Berkeley County.
- NYT database: 1,706 cases in Virginia with 121 in Loudoun County. Our hypothesis is that the website found that the red sentence was wrong so they deleted it.



🟩 COVID-19 cases

Less mentioning of China and World case numbers

Started mentioning NJ

Less mentioning of New York & US after a sharp increase

Recap: Annotation schemes of “time”

- Two key concepts for annotating “time” are
  - Event mentions (E)
  - Time points (T)

- We can view the existing annotation schemes of “time” based on how they deal with E and T
  - Focus on (E,T) relationship
  - Focus on (E,E) relationship
  - Both (E,T) and (E,E) relationships

Recap: Annotation schemes of “time”

- Two key concepts for annotating “time” are
  - Event mentions (E)
  - Time points (T)

- We can view the existing annotation schemes of “time” based on how they deal with E and T
  - Focus on (E,T) relationship
  - **Focus on (E,E) relationship**
  - Both (E,T) and (E,E) relationships

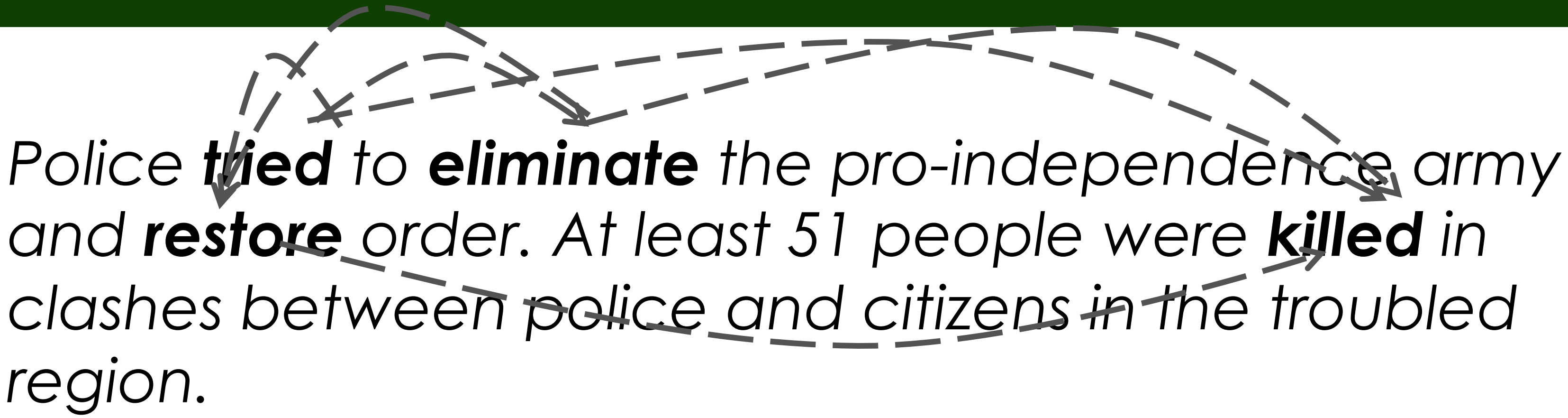
TimeML / TimeBank

*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.*

- Annotation task: to label the TempRels between the **bold** faced events

- TimeBank *[Pustejovsky et al., 2003]*
  - Annotators *freely* add TempRels between those events.
  - It's likely that some TempRels will be missed.
  - E.g., only one relation between "**eliminate**" and "**restore**" is annotated in TimeBank, while other relations such as "**tried**" is before "**eliminate**" and "**tried**" is also before "**killed**" are missed.

TimeBank-Dense[1]



*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.*

- TimeBank-Dense[1]

  "All" – all event pairs are presented, one-by-one, and an annotator *must* provide a response for each of them.

  "1" – In the physical world, time is one dimensional.

  "0" – missing relations anymore.

  "?" – *However*, some pairs of events are very confusing, resulting in disagreement.

[1] An annotation framework for dense event ordering. Cassidy et al., ACL'14.

Problem

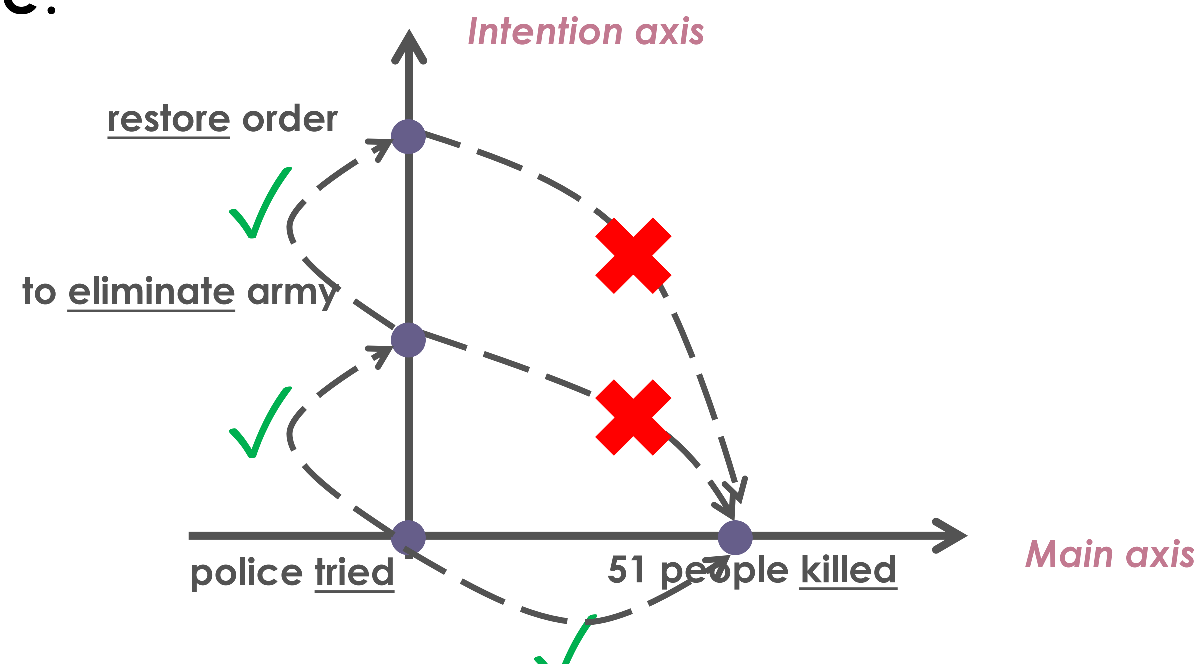
Police ***tried*** *to* ***eliminate*** *the pro-independence army and* ***restore*** *order. At least 51 people were* ***killed*** *in clashes between police and citizens in the troubled region.*

- E.g., what's the relation between ***restore*** and ***killed***?
    - ☐ restore `BEFORE` killed?
    - ☐ restore `INCLUDES` killed?
    - ☐ restore `AFTER` killed?

MATRES: Multi-axis scheme[1]

*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.*

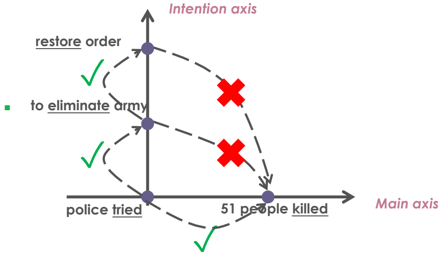
- We suggest that **multiple time axes exist in natural language**.



[1] A Multi-Axis Annotation Scheme for Event Temporal Relations. Ning et al., ACL'18.

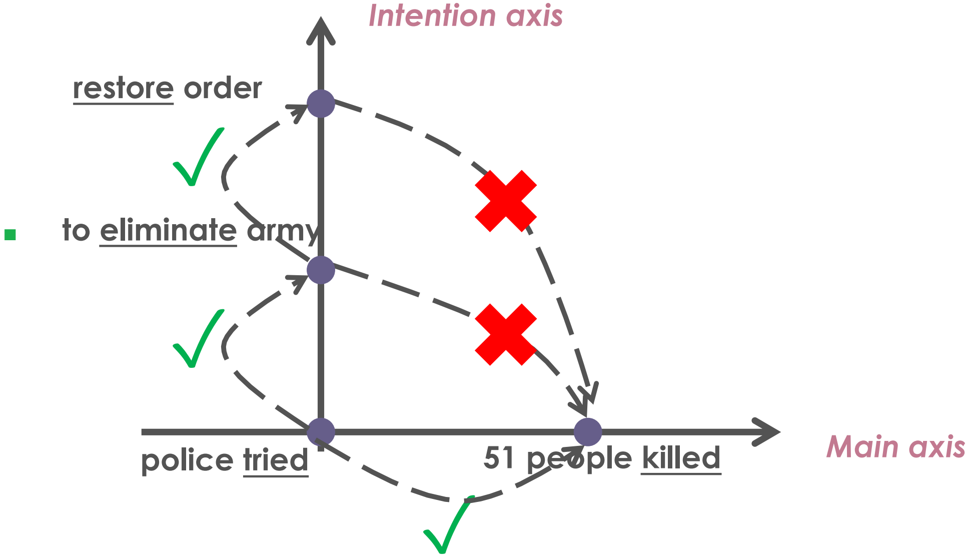
MATRES: Multi-axis scheme

We complete a set of definitions for axes



Event Type	*Axis Type*	%
intention, opinion	orthogonal axis	~20
hypothesis, generic	parallel axis	~10
negation	not on any axis	
static, recurrent	not considered now	
main	main axis	~70

Problem: these formalisms look complicated



Can we use natural language to annotate natural language?

Advantages of QA as a format

1. Handle fuzzy relations

Fuzzy relations

Heavy snow is causing disruption to transport across the UK, with heavy rainfall bringing flooding to the south-west of England.

Q: What happens at about the same time as the disruption?
A: flooding

Advantages of QA as a format

2. Handle events in different modalities

Events in different modes

The lion had a large meal and slept for 24 hours.

[Negated] The lion didn't sleep after having a large meal.

[Uncertain] The lion may have had a large meal before sleeping.

[Hypothetical] If the lion has a large meal, it will sleep for 24 hours.

[Repetitive] The lion used to sleep for 24 hours after having large meals.

[Generic] After having a large meal, lions may sleep longer.

Advantages of QA as a format

2. Handle events in different modalities

Questions that query events in different modes

[Negated] What didn’t the lion do after a large meal?

[Uncertain] What might the lion do before sleeping?

[Hypothetical] What will the lion do if it has a large meal?

[Repetitive] What did the lion use to do after large meals?

[Generic] What do lions do after a large meal?

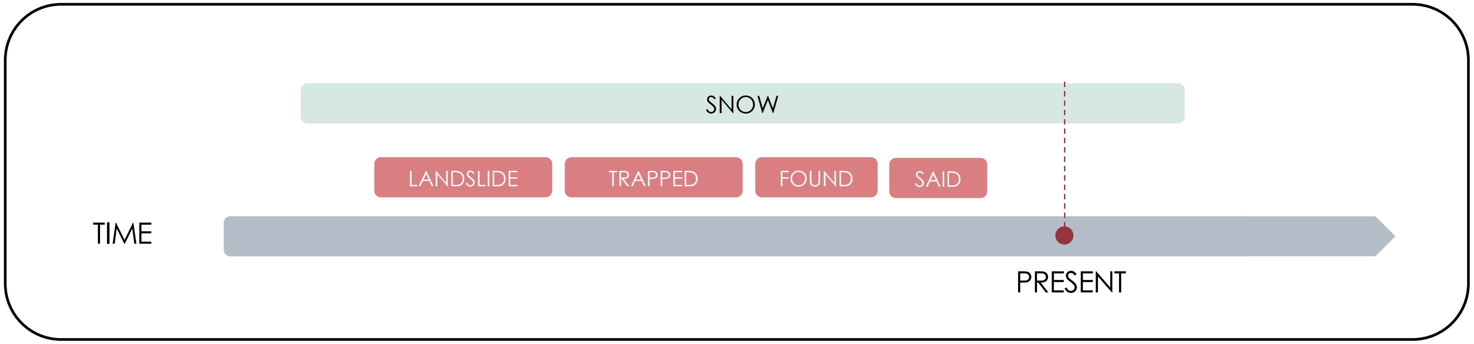
Advantages of QA as a format

3. Prior formalisms had a major issue with defining when two events should have a relation.

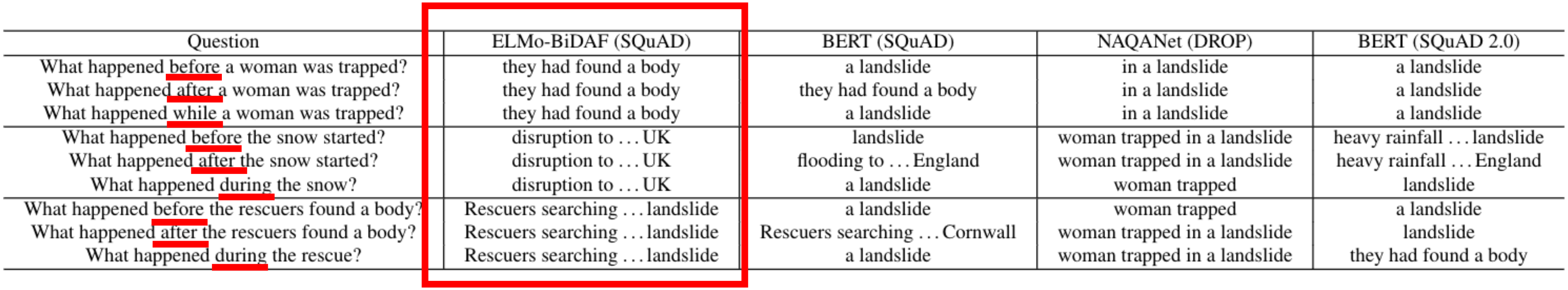
- *e.g., TimeBank, TimeBank-dense, VerbClause, RED, and MATRES*

In TORQUE, the format of QA naturally bypasses this issue – we don't need to teach annotators these linguistic formalisms.

Text: 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.

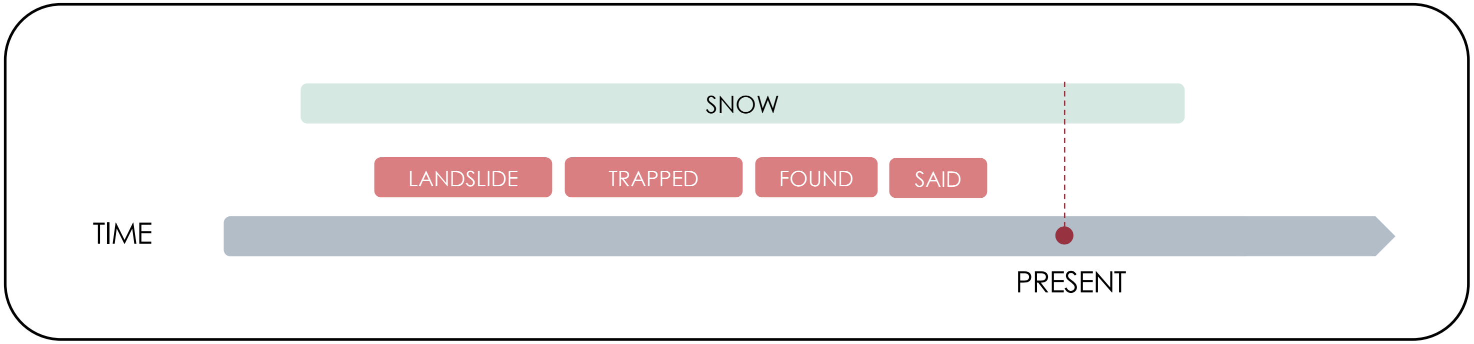


Existing systems don't really use "time" to answer these questions.

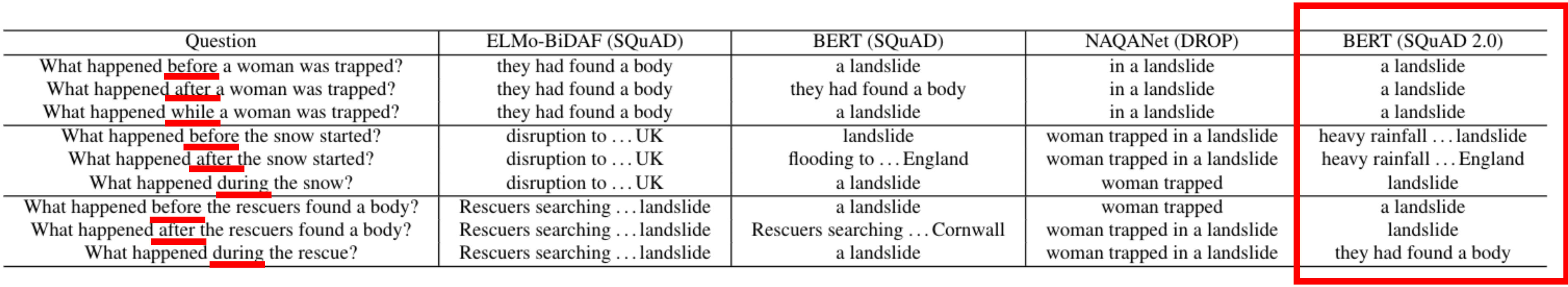


Question	ELMo-BiDAF (SQuAD)	BERT (SQuAD)	NAQANet (DROP)	BERT (SQuAD 2.0)
What happened before a woman was trapped?	they had found a body	a landslide	in a landslide	a landslide
What happened after a woman was trapped?	they had found a body	they had found a body	in a landslide	a landslide
What happened while a woman was trapped?	they had found a body	a landslide	in a landslide	a landslide
What happened before the snow started?	disruption to . . . UK	landslide	woman trapped in a landslide	heavy rainfall . . . landslide
What happened after the snow started?	disruption to . . . UK	flooding to . . . England	woman trapped in a landslide	heavy rainfall . . . England
What happened during the snow?	disruption to . . . UK	a landslide	woman trapped	landslide
What happened before the rescuers found a body?	Rescuers searching . . . landslide	a landslide	woman trapped	a landslide
What happened after the rescuers found a body?	Rescuers searching . . . landslide	Rescuers searching . . . Cornwall	woman trapped in a landslide	landslide
What happened during the rescue?	Rescuers searching . . . landslide	a landslide	woman trapped in a landslide	they had found a body

Text: 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.

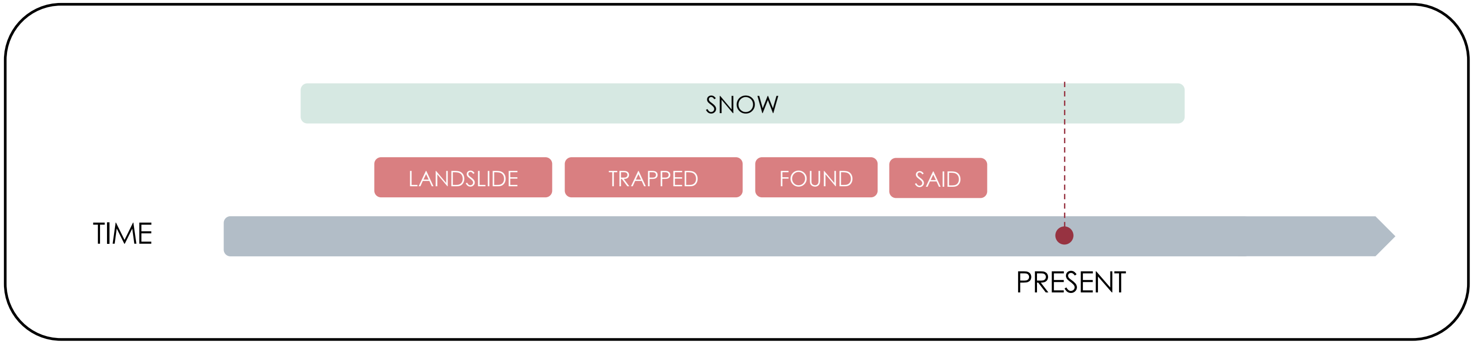


Existing systems don't really use "time" to answer these questions.



Question	ELMo-BiDAF (SQuAD)	BERT (SQuAD)	NAQANet (DROP)	BERT (SQuAD 2.0)
What happened before a woman was trapped?	they had found a body	a landslide	in a landslide	a landslide
What happened after a woman was trapped?	they had found a body	they had found a body	in a landslide	a landslide
What happened while a woman was trapped?	they had found a body	a landslide	in a landslide	a landslide
What happened before the snow started?	disruption to . . . UK	landslide	woman trapped in a landslide	heavy rainfall . . . landslide
What happened after the snow started?	disruption to . . . UK	flooding to . . . England	woman trapped in a landslide	heavy rainfall . . . England
What happened during the snow?	disruption to . . . UK	a landslide	woman trapped	landslide
What happened before the rescuers found a body?	Rescuers searching . . . landslide	a landslide	woman trapped	a landslide
What happened after the rescuers found a body?	Rescuers searching . . . landslide	Rescuers searching . . . Cornwall	woman trapped in a landslide	landslide
What happened during the rescue?	Rescuers searching . . . landslide	a landslide	woman trapped in a landslide	they had found a body

Text: 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.



Existing systems don't really use "time" to answer these questions.

<u>GPT-4o; accessed in Jun 2024</u>

Please choose from the following options: (a) snow, (b) a landslide, (c) rescuers said something, (d) rescuers found a body, (e) I don't know. You can choose multiple options if multiple answers apply.

Q: What happened **before** the snow started?

A: landslide, said ✗

Q: What happened **after** the snow started?

A: found ✗

Q: What happened **during** the snow?

A: snow ✗

TORQUE[1]: Data collection process

Given a passage, an annotator needs to

1. Label all the events
2. Repeat the following steps
    - ☐ Ask a question querying temporal relations
    - ☐ Select all the correct answers from the event list
    - ☐ Slightly perturb the question and answer it

[1] TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. Ning et al., EMNLP'20.

Perturbations of existing questions

- A temporal relation can be represented as a triplet: (Event A, Relation r, Event B)
- A valid question is: (?, r, B)
    - What happened after B?
- Perturbation: (?, r’, B)
    - What happened before/while …?
    - What had begun before B?
    - What happened before B finished?
    - …

TORQUE

3.2k passages from the 2.8k articles in TempEval3

- Two sentences per passage
- 50 tokens per passage

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.

The loan may be extended by the McAlpine group for an additional year with an increase in the conversion price to $2.50 a share. The sale of shares to the McAlpine family along with the recent sale of 750,000 shares of Meridian stock to Haden MacLellan Holding PLC of Surrey, England and a recent public offering have increased Meridian's net worth to $8.5 million, said William Feniger, chief executive officer of Toledo, Ohio-based Meridian.

This time, however, some analysts think he could face a real battle. "Without some unexpected coup de theatre, I don't see what will block the Paribas bid," said Philippe de Cholet, analyst at the brokerage Cholet-Dupont amp Cie. Mr. de Cholet said Mr. Fournier's biggest hope was to somehow persuade regulatory authorities to block the bid.

3.2k passages from the 2.8k articles in TempEval3

- Two sentences per passage
- 50 tokens per passage

24.9k events

For each passage, we have

- 3 hard-coded questions

- on average, 7 user-generated questions

- Grouped as contrast sets

TORQUE

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

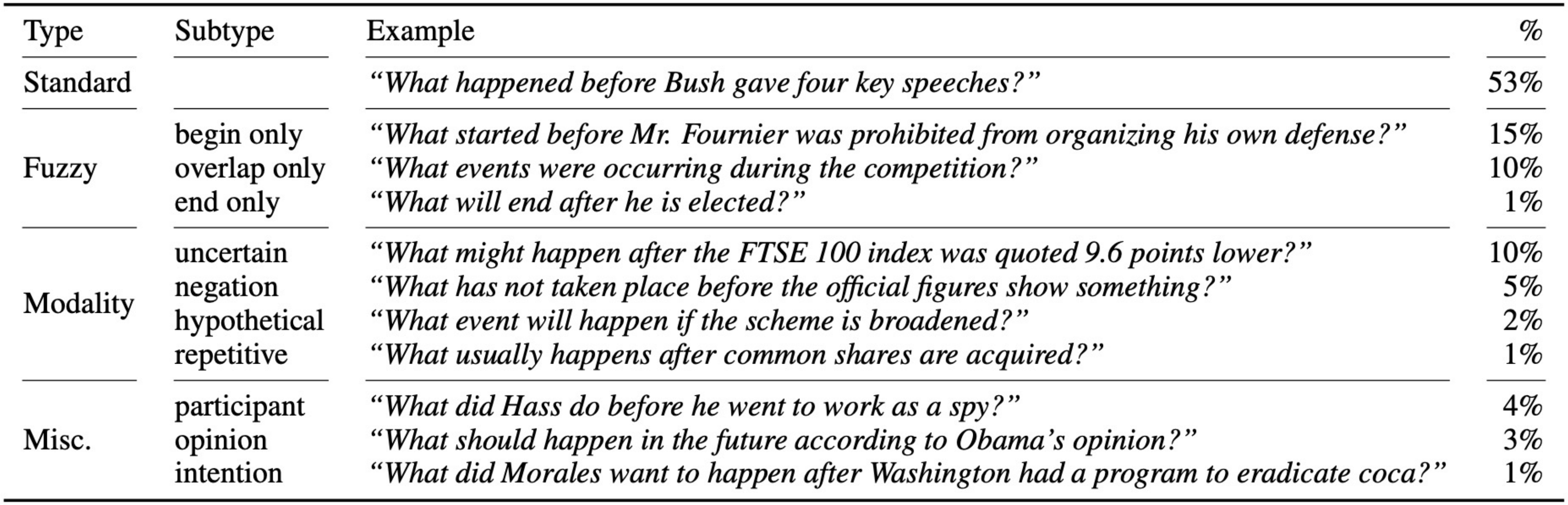
Group of contrast questions

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

Temporal phenomena in TORQUE

- By manually checking a sample of 200 questions, we have the following categorization on their temporal phenomena (some questions can have multiple types):

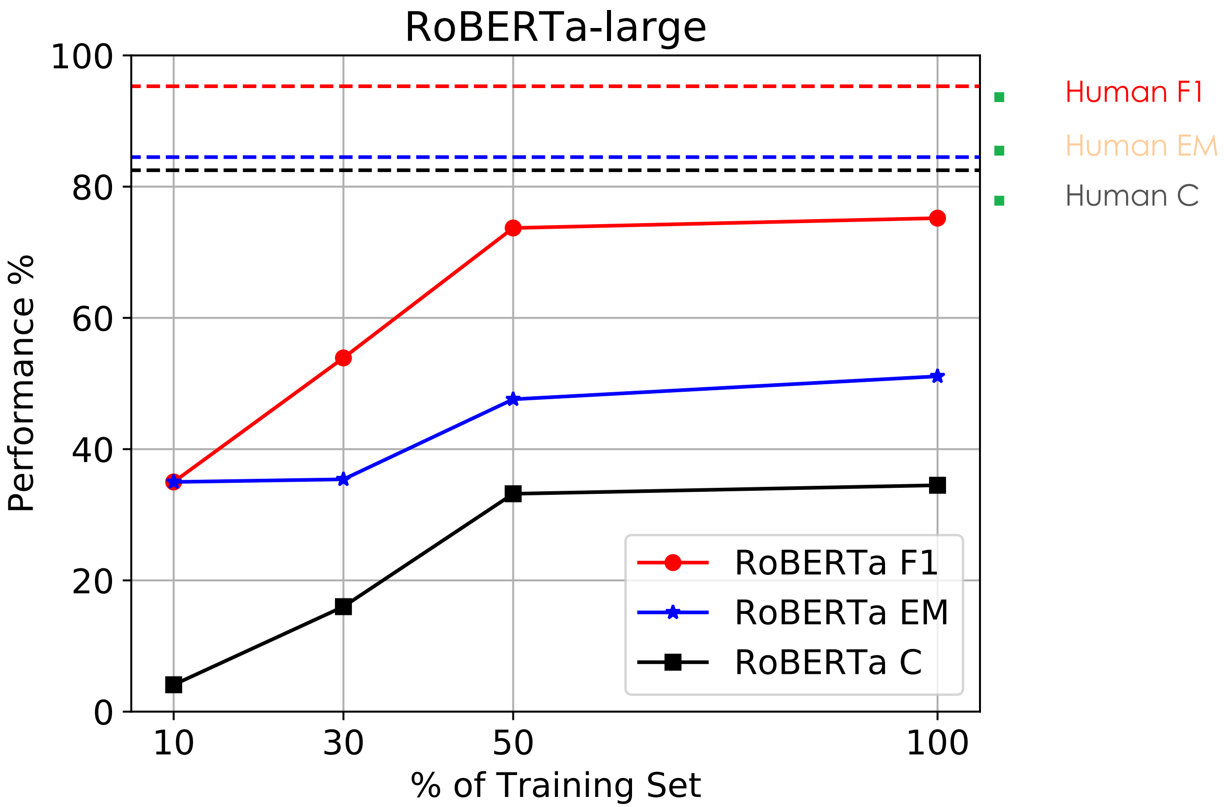


Type	Subtype	Example	%
Standard		*"What happened before Bush gave four key speeches?"*	53%
Fuzzy	begin only	*"What started before Mr. Fournier was prohibited from organizing his own defense?"*	15%
	overlap only	*"What events were occurring during the competition?"*	10%
	end only	*"What will end after he is elected?"*	1%
Modality	uncertain	*"What might happen after the FTSE 100 index was quoted 9.6 points lower?"*	10%
	negation	*"What has not taken place before the official figures show something?"*	5%
	hypothetical	*"What event will happen if the scheme is broadened?"*	2%
	repetitive	*"What usually happens after common shares are acquired?"*	1%
Misc.	participant	*"What did Hass do before he went to work as a spy?"*	4%
	opinion	*"What should happen in the future according to Obama's opinion?"*	3%
	intention	*"What did Morales want to happen after Washington had a program to eradicate coca?"*	1%

Experiment

- Setup:
  - Given a passage and a question,
  - model looks at every token in the passage
  - makes a binary classification of whether this token is an answer to the question or not
- Model:
  - One-layered perceptron on top of RoBERTa
- Split:
  - Train 80%, dev 5%, test 15%
  - Dev & test are provided by 5 different annotators on their events; by 3 different annotators on their answers

Experiment



F1 & EM: standard metrics when there're multiple reference answers

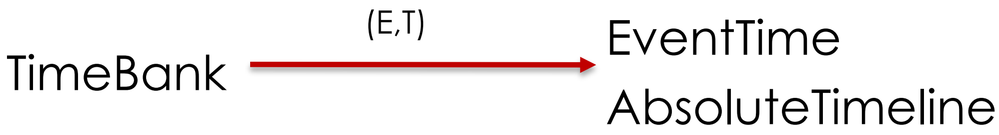
C (consistency) is the percentage of contrast groups for which a model's predictions have F1>=80% for all questions in a group

Original talk in 2020: https://slideslive.com/38938807

A (very biased) view of advancement in ”time”

TimeBank

A (very biased) view of advancement in ”time”



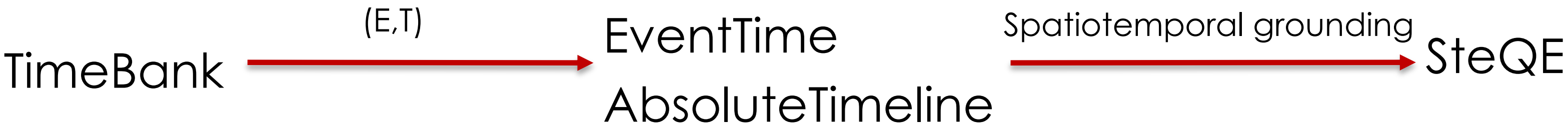
TimeBank

(E,T)

EventTime

AbsoluteTimeline

A (very biased) view of advancement in "time"



TimeBank

(E,T)

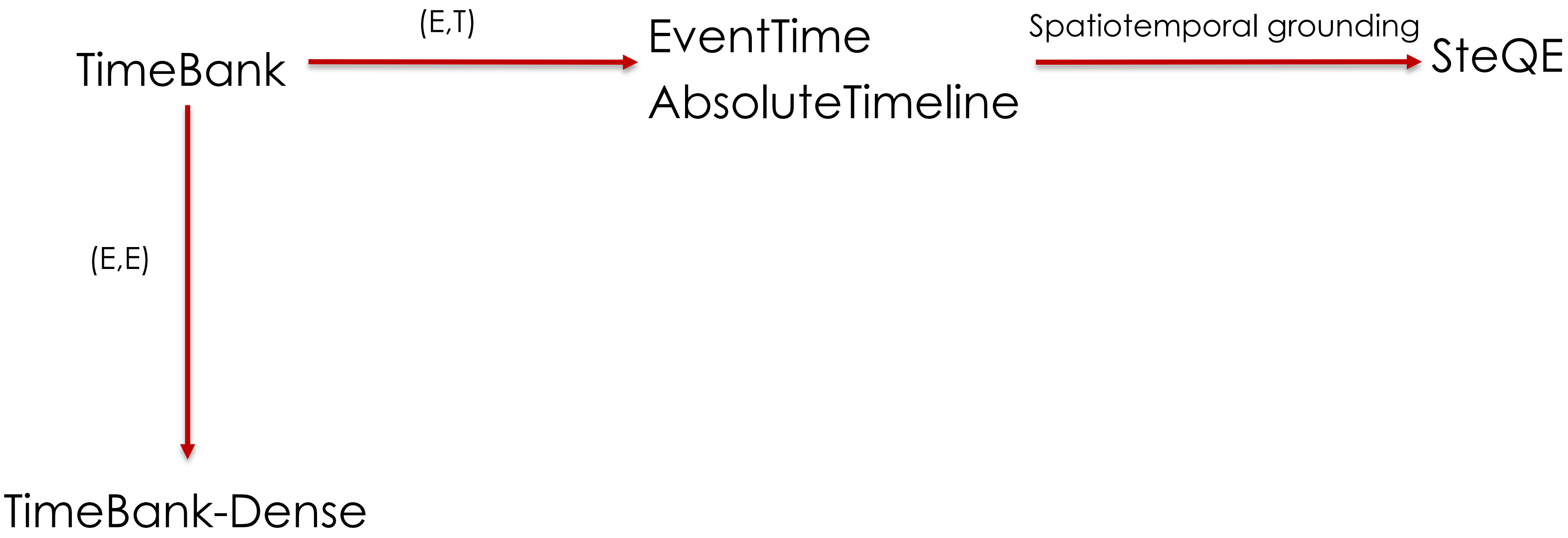
EventTime

AbsoluteTimeline

Spatiotemporal grounding

SteQE

A (very biased) view of advancement in "time"



TimeBank

(E,T)

EventTime

AbsoluteTimeline

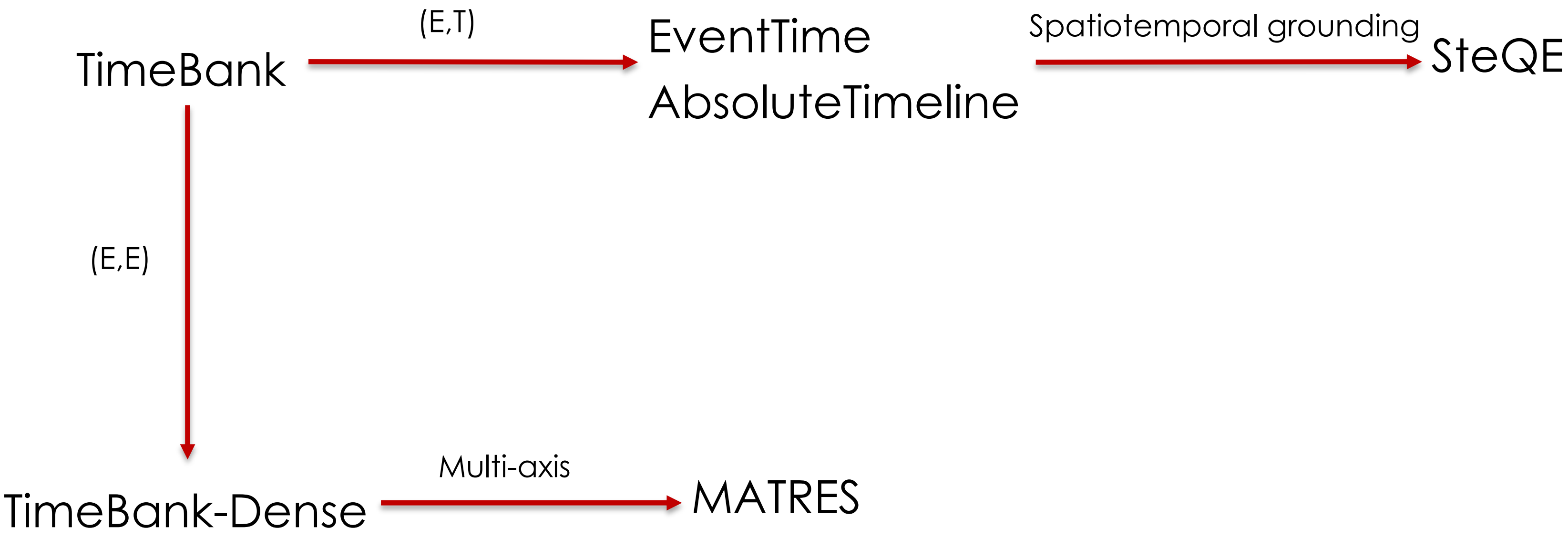
Spatiotemporal grounding

SteQE

(E,E)

TimeBank-Dense

A (very biased) view of advancement in ”time”



TimeBank

(E,T)

EventTime

AbsoluteTimeline

Spatiotemporal grounding

SteQE

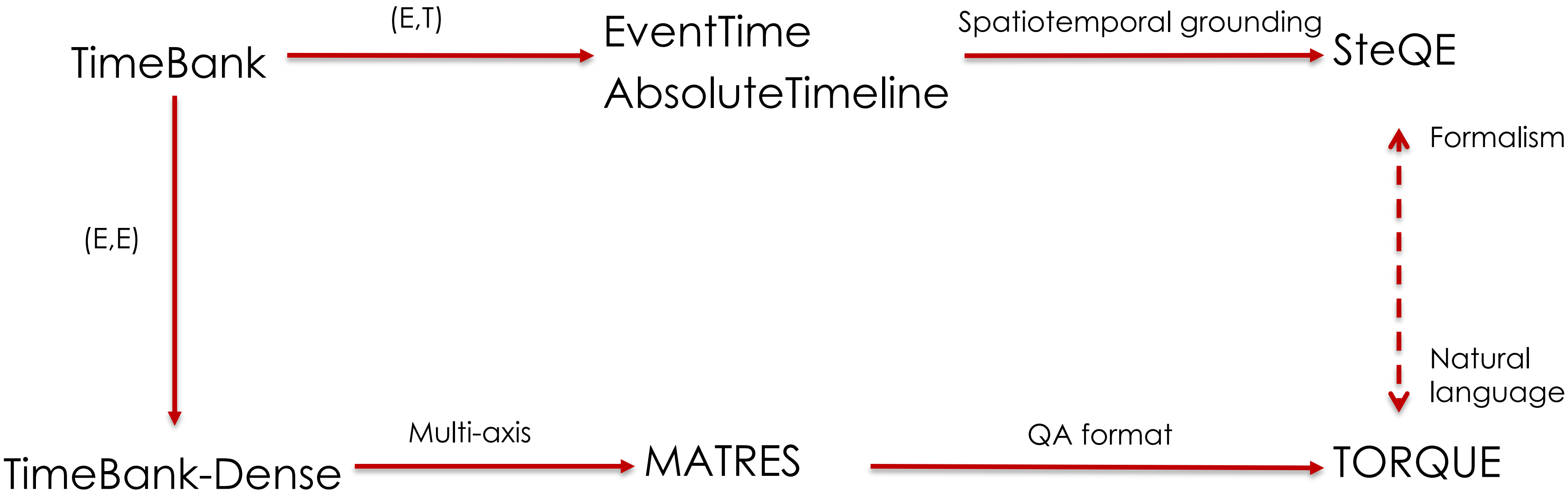
(E,E)

TimeBank-Dense

Multi-axis

MATRES

A (very biased) view of advancement in ”time”



TimeBank

(E,T)

EventTime

AbsoluteTimeline

Spatiotemporal grounding

SteQE

(E,E)

Formalism

Natural language

TimeBank-Dense

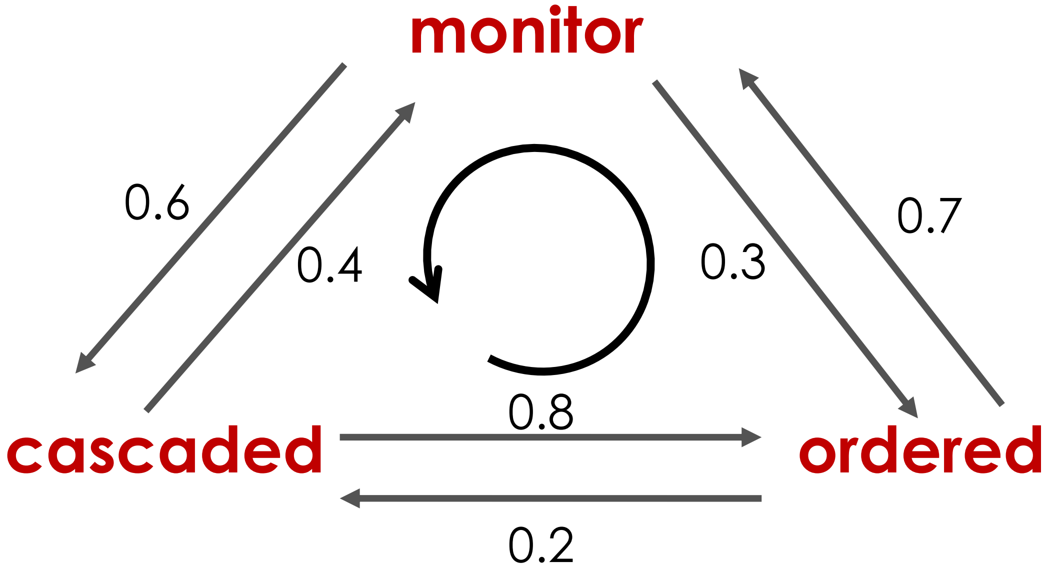
Multi-axis

MATRES

QA format

TORQUE

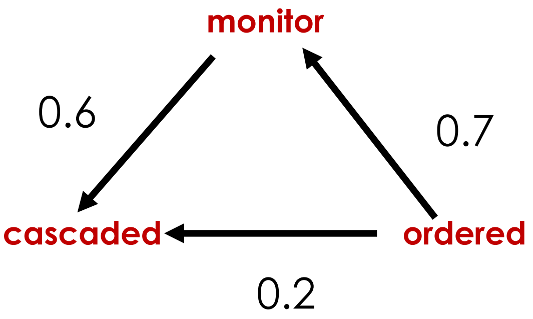
Time is a good vehicle to study reasoning



Reasoning constraint: time cannot be a loop

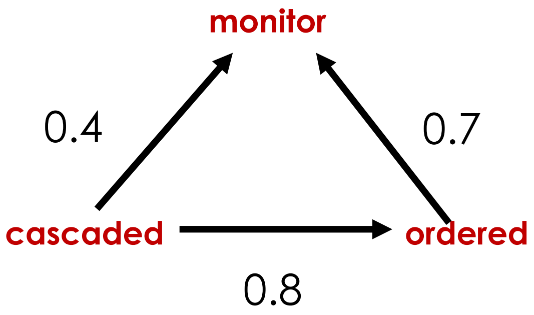
Global inference (a toy example)

Option 1



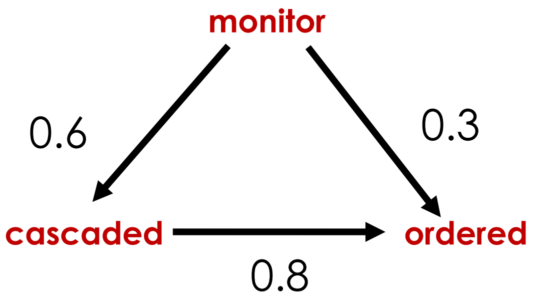
0.6+0.2+0.7=1.5

Option 2



0.4+0.8+0.7=1.9

Option 3

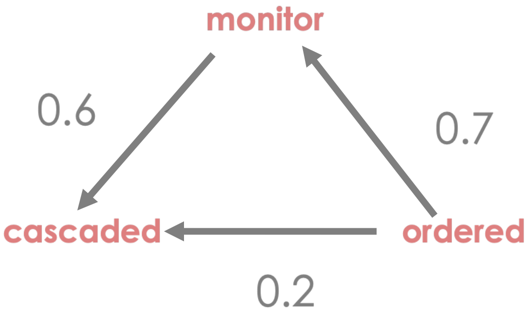


0.6+0.3+0.8=1.7

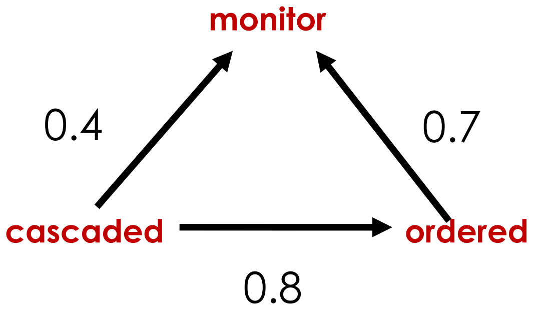
We should not only select the assignment with the best score, but also avoid loops

Global inference (a toy example)

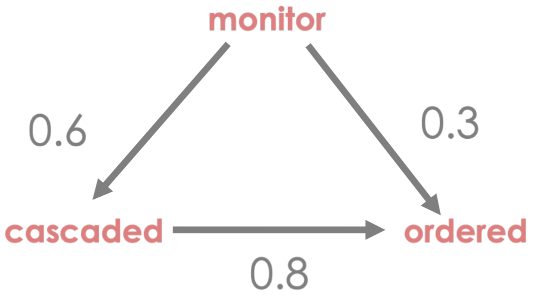
Option 2


monitor
0.6
0.7
cascaded
ordered
0.2

0.6+0.2+0.7=1.5


monitor
0.4
0.7
cascaded
ordered
0.8

0.4+0.8+0.7=1.9

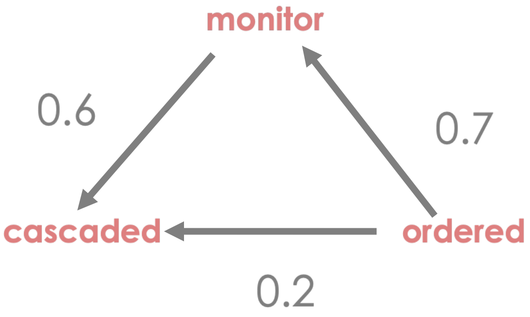

monitor
0.6
0.3
cascaded
ordered
0.8

0.6+0.3+0.8=1.7

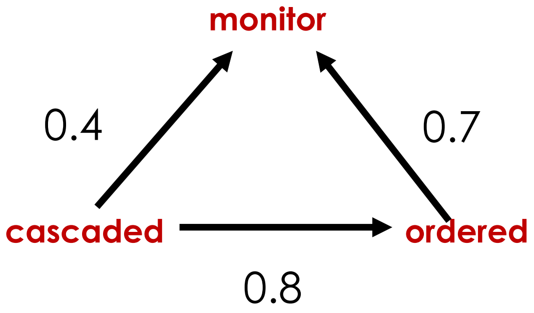
We should not only select the assignment with the best score, but also avoid loops

Global inference (a toy example)

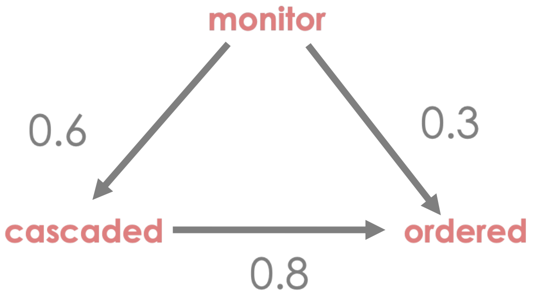
Option 2



0.6+0.2+0.7=1.5



0.4+0.8+0.7=1.9

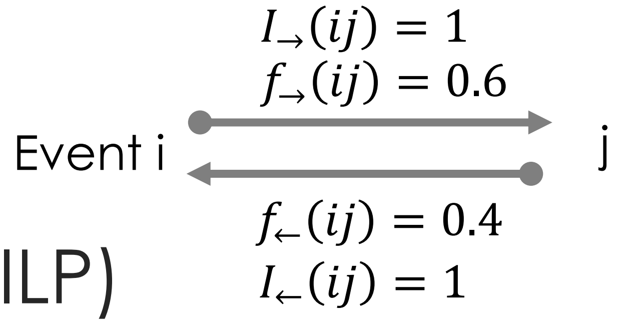


0.6+0.3+0.8=1.7

How should we formulate this procedure?

Global inference via ILP *[Roth & Yih, 2004]*

Global inference via ILP


$I_{\rightarrow}(ij) = 1$
$f_{\rightarrow}(ij) = 0.6$
Event i
j
$f_{\leftarrow}(ij) = 0.4$
$I_{\leftarrow}(ij) = 1$

Integer Linear Programming (ILP)

real variable

$$\hat{I} = arg\max_{I} \sum_{i<j} \sum_{r} f_r(ij) \boxed{I_r(ij)}$$

boolean variable

s.t. $\forall i, j, k$

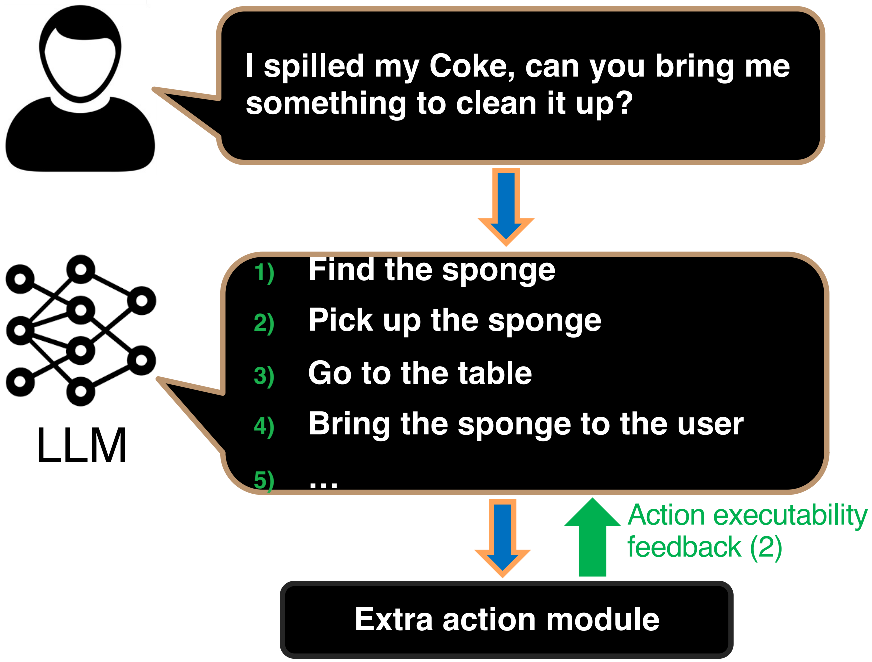
$$\sum_{r} I_r(ij) = 1, \qquad I_{r1}(ij) + I_{r2}(jk) - I_{r3}(ik) \leq 1$$

Uniqueness

Reasoning constraints

We're maximizing the score of an entire graph **while enforcing reasoning constraints**.

(Naïve) Robot motion planning



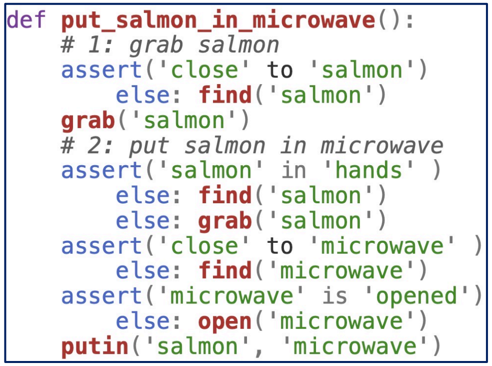
I spilled my Coke, can you bring me something to clean it up?

LLM

1) **Find the sponge**
2) **Pick up the sponge**
3) **Go to the table**
4) **Bring the sponge to the user**
5) **…**

Action executability feedback (2)

Extra action module



```
def put_salmon_in_microwave():
    # 1: grab salmon
    assert('close' to 'salmon')
        else: find('salmon')
    grab('salmon')
    # 2: put salmon in microwave
    assert('salmon' in 'hands' )
        else: find('salmon')
        else: grab('salmon')
    assert('close' to 'microwave' )
        else: find('microwave')
    assert('microwave' is 'opened')
        else: open('microwave')
    putin('salmon', 'microwave')
```

[3] Singh et al. ProgPrompt: Generating situated robot task plans using large language models, ICRA 2023.

LLMs provide common knowledge

LLMs as task planners

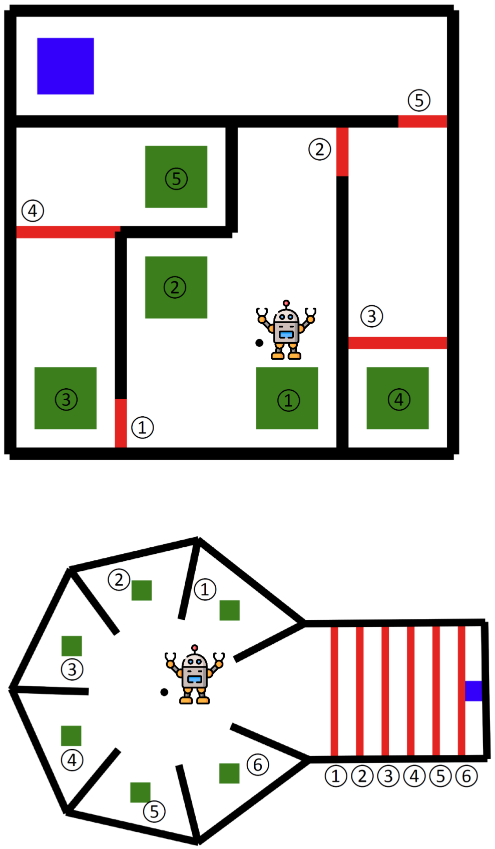
[1] Huang et al. Language Models as Zero-Shot Planners, ICML 2022.
[2] Ahn et al. Do as I can, not as I say: Grounding language in robotic affordances. CoRL 2022.

Robot motion planning via “Temporal Logic”

- Temporal Logic is a rigorous system of rules for requirement engineering that can capture many complex spatial, temporal, and logical requirements.

- If we can translate natural language task descriptions into temporal logic, then robots will know what to do and when to do it, not how to do it.

A simple example of temporal logic



Robot puzzle-solving

Robots are asked to visit the blue goal region

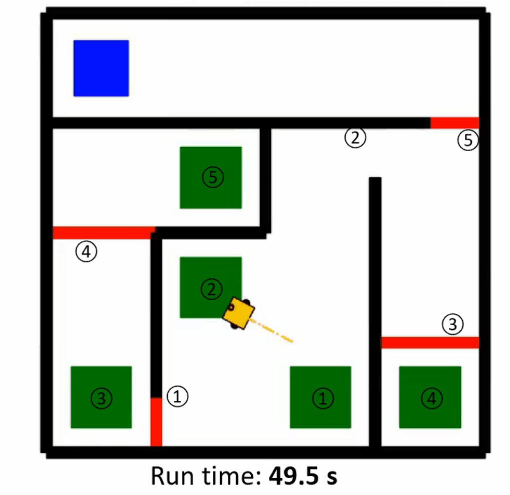
1. Robots should always reach their destinations with minimum traveling time, regardless of the puzzle type
2. Robots cannot travel through walls (black) or closed doors (red)
3. Robots must visit the correspondingly numbered green region to pick the keys to open the doors

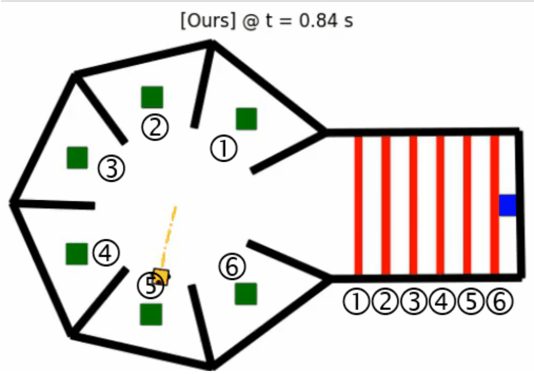
Temporal logic representation of these rules is:

$$\left(\mathbf{F}_{[0,10]}\text{Destination}\right) \wedge \left(\mathbf{G}_{[0,10]}\neg\text{Wall}\right) \wedge \left(\bigwedge_{i=1}^{N} \neg\text{Door}_i \ \mathbf{U}\ \text{Key}_i\right)$$

Robot task and motion planning from temporal logic

- Planning algorithms can plan motions for agents with provably optimal computational complexity based on temporal logic representation[1]





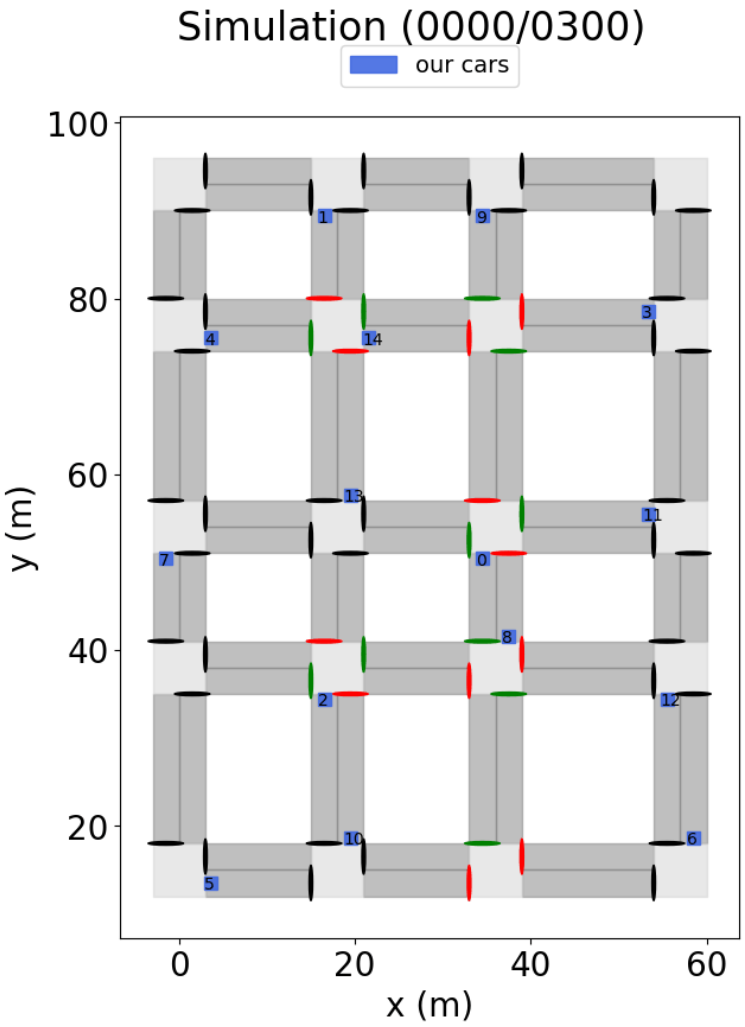
$$\left(\mathbf{F}_{[0,10]}\text{Destination}\right) \wedge \left(\mathbf{G}_{[0,10]}\neg\text{Wall}\right) \wedge \left(\bigwedge_{i=1}^{N} \neg\text{Door}_i \;\; \mathbf{U}\; \text{Key}_i\right)$$

[1] Sun, Dawei, Jingkai Chen, Sayan Mitra, and Chuchu Fan. "Multi-agent motion planning from signal temporal logic specifications." *IEEE Robotics and Automation Letters (2022)*.

Robot task and motion planning from temporal logic

Embed the following rules in one vehicle's action, trained in a single intersection[1]

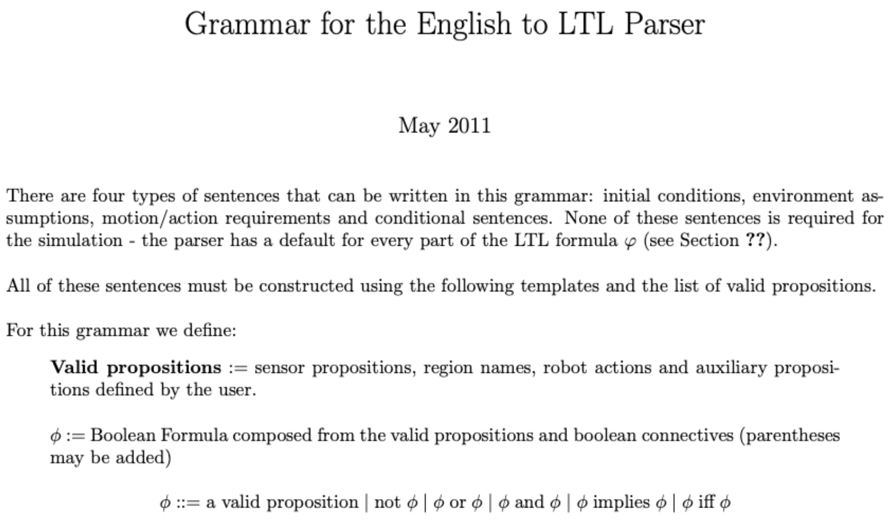
- Avoid rear-end collision
- Stop sign rules
- Traffic light rules



[1] Meng, Yue, and Chuchu Fan. "Signal Temporal Logic Neural Predictive Control." *IEEE Robotics and Automation Letters (2023).*

Natural language to temporal logic (NL2TL)

- Previous works usually used hard-code parsing rules and made strong assumptions on input/output format.



Grammar for the English to LTL Parser

May 2011

There are four types of sentences that can be written in this grammar: initial conditions, environment assumptions, motion/action requirements and conditional sentences. None of these sentences is required for the simulation - the parser has a default for every part of the LTL formula φ (see Section **??**).

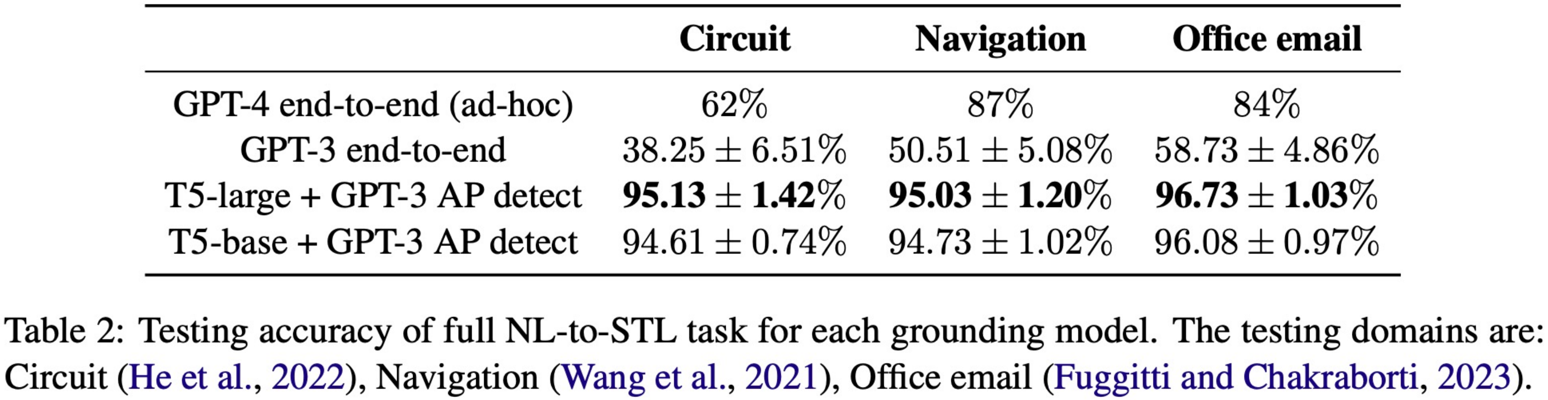
All of these sentences must be constructed using the following templates and the list of valid propositions.

For this grammar we define:

> **Valid propositions** := sensor propositions, region names, robot actions and auxiliary propositions defined by the user.
>
> ϕ := Boolean Formula composed from the valid propositions and boolean connectives (parentheses may be added)

$$\phi ::= \text{a valid proposition} \mid \text{not } \phi \mid \phi \text{ or } \phi \mid \phi \text{ and } \phi \mid \phi \text{ implies } \phi \mid \phi \text{ iff } \phi$$

Natural language to temporal logic (NL2TL)[1]



	Circuit	Navigation	Office email
GPT-4 end-to-end (ad-hoc)	62%	87%	84%
GPT-3 end-to-end	$38.25 \pm 6.51\%$	$50.51 \pm 5.08\%$	$58.73 \pm 4.86\%$
T5-large + GPT-3 AP detect	$\mathbf{95.13 \pm 1.42}\%$	$\mathbf{95.03 \pm 1.20}\%$	$\mathbf{96.73 \pm 1.03}\%$
T5-base + GPT-3 AP detect	$94.61 \pm 0.74\%$	$94.73 \pm 1.02\%$	$96.08 \pm 0.97\%$

Table 2: Testing accuracy of full NL-to-STL task for each grounding model. The testing domains are: Circuit (He et al., 2022), Navigation (Wang et al., 2021), Office email (Fuggitti and Chakraborti, 2023).

[1] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP'23.

Natural language to temporal logic (NL2TL)[1]

Atomic proposition recognition:

If a response is created in Slack , or the Acoustic Campaign contact is being updated then in response the scenario that a response is created in Asana shall be instantly observed .

↓

Lifted natural sentence:

If (prop_1) , or (prop_2) then in response the scenario that (prop_3) shall be instantly observed .

↓

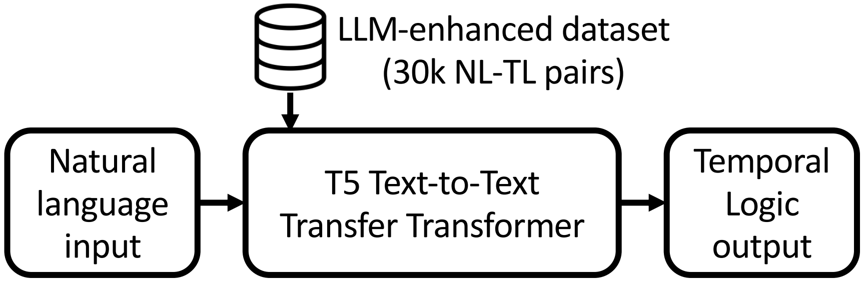
Lifted STL prediction:

globally (((prop_1) or (prop_2)) imply (prop_3))

↓

Full STL prediction:

globally (((create_Slack) or (update_Acoustic Campaign)) imply (created_Asana))


LLM-enhanced dataset (30k NL-TL pairs)

Natural language input → T5 Text-to-Text Transfer Transformer → Temporal Logic output

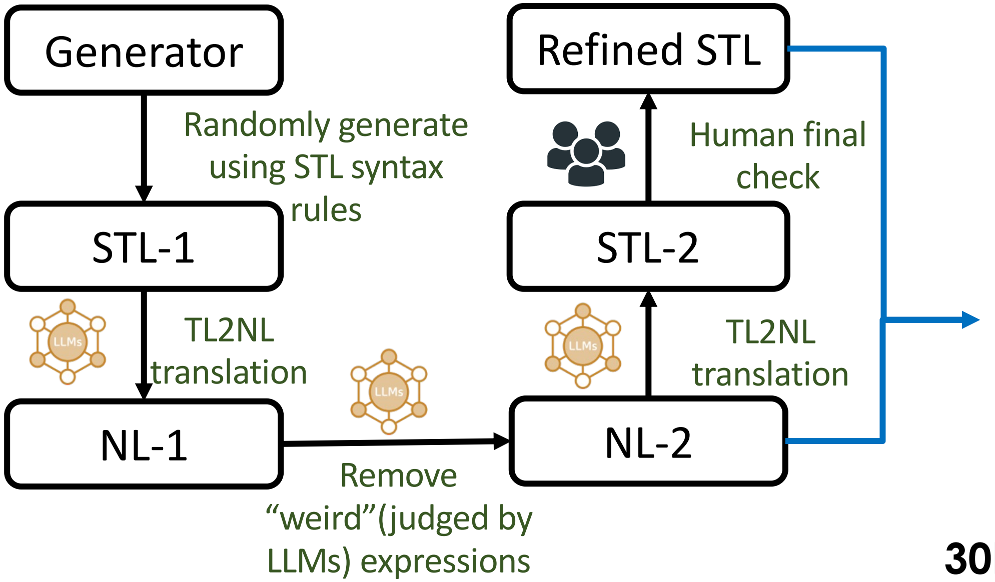
The translation problem is split into two steps[1]

- Identify specific basic operating units (atomic propositions or APs) and hide them from translation.
- Translate the Temporal Logic structure with hidden APs (lifted version) independent of application domains.

---

[1] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP'23.

LLM-assisted data generation



Generator

Randomly generate using STL syntax rules

STL-1

TL2NL translation

NL-1

Remove “weird”(judged by LLMs) expressions

NL-2

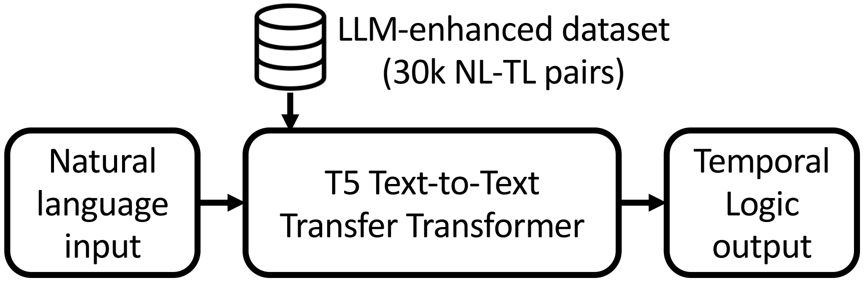
TL2NL translation

STL-2

Human final check

Refined STL

Output NL-TL pairs



LLM-enhanced dataset (30k NL-TL pairs)

Natural language input

T5 Text-to-Text Transfer Transformer

Temporal Logic output

# Operators per STL		# Words per sentence (no APs)	
Median	Max	Median	Max
3	7	16	72

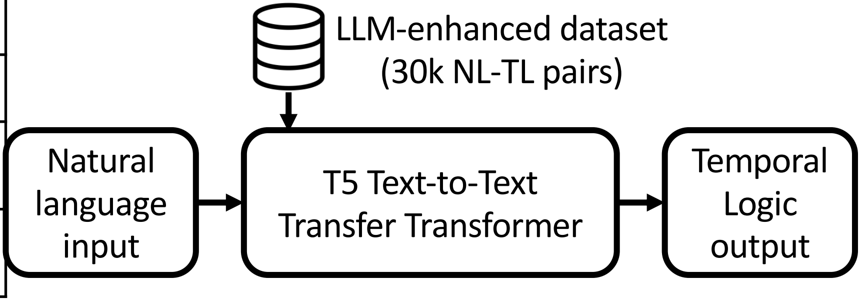
30k NL-TL pairs, across 5 different application domains (Circuit design, navigation, manipulation, household, email)

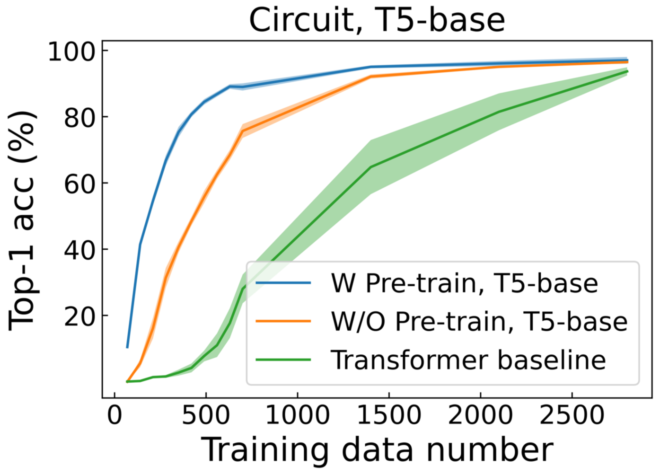
- 14k pairs collected from existing datasets.
- 16k pairs synthesized via GPT-3 and re-annotated.

[1] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP’23.

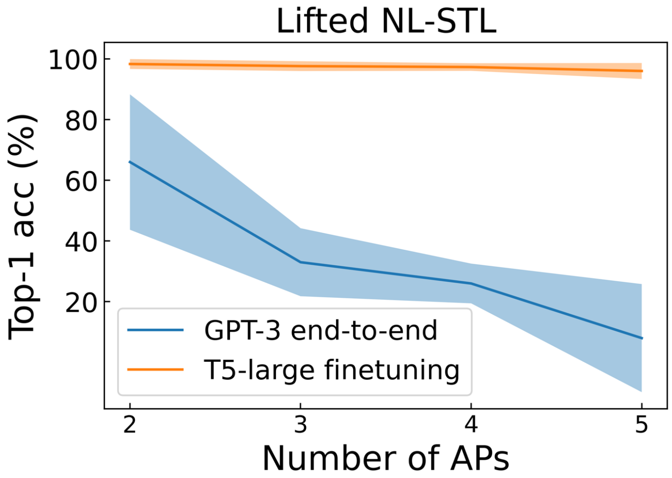
Translate English instructions to Temporal Logic specifications

Model	#parameters	Accuracy (%)
GPT-3	175B	38.25-58.73
GPT-4	n/a	73
NL2TL[1]	770M	95.03-96.73
NL2TL[1]	220M	94.73-96.08





Pretrained model needs half the data to be fine-tuned within a previously unseen domain

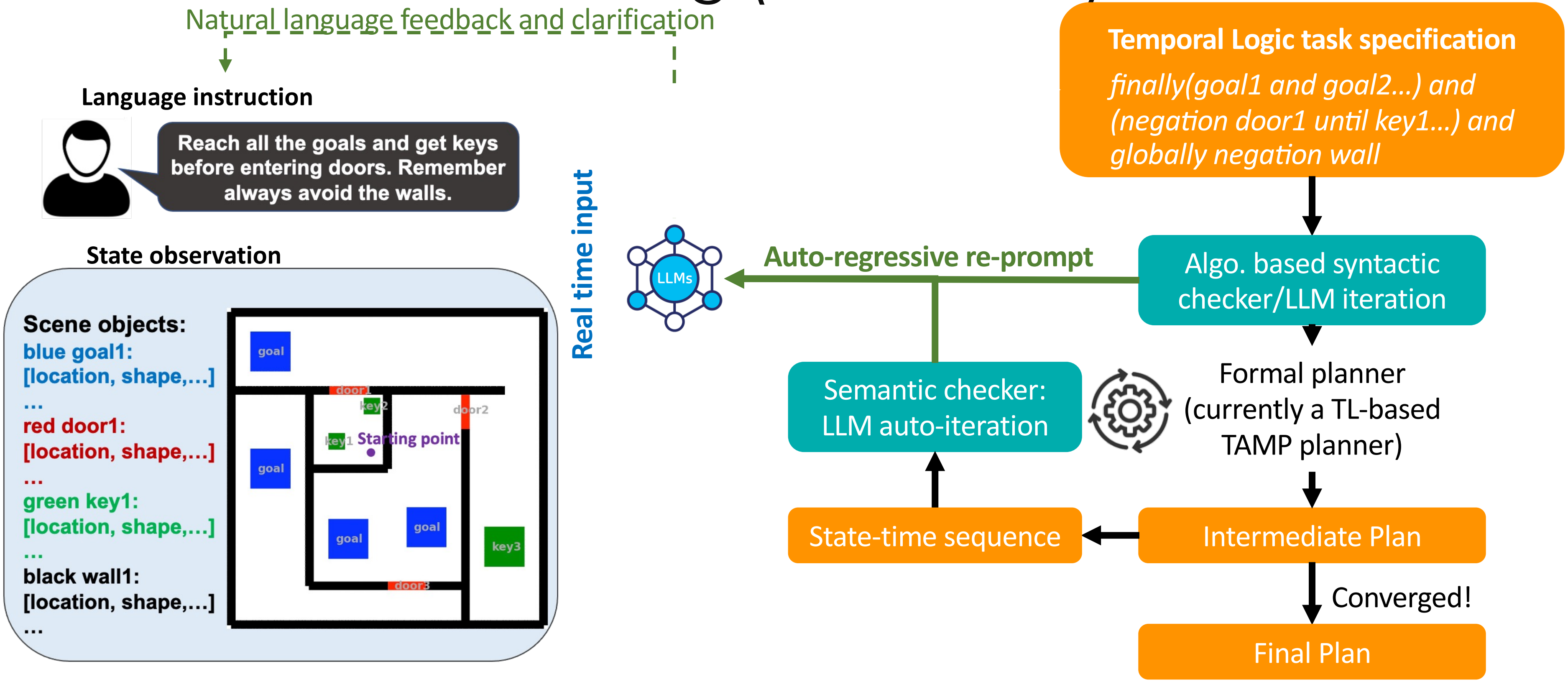


End-to-end GPT-3's performance gets worse as the #AP increases

---

[1] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP'23.

Auto-regressive LLM-based Task and Motion Planning (AutoTAMP)



Natural language feedback and clarification

Language instruction

Reach all the goals and get keys before entering doors. Remember always avoid the walls.

State observation

Scene objects:
blue goal1:
[location, shape,...]
...
red door1:
[location, shape,...]
...
green key1:
[location, shape,...]
...
black wall1:
[location, shape,...]
...

Real time input

Auto-regressive re-prompt

Temporal Logic task specification

finally(goal1 and goal2...) and (negation door1 until key1...) and globally negation wall

Algo. based syntactic checker/LLM iteration

Formal planner (currently a TL-based TAMP planner)

Semantic checker: LLM auto-iteration

State-time sequence

Intermediate Plan

Converged!

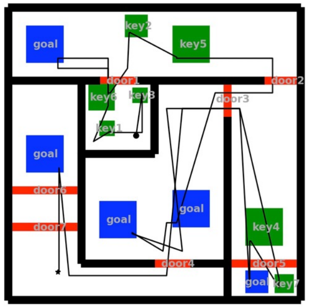
Final Plan

LLM-As-Translator & Checker

[1] AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. Chen et al., ICRA'24.

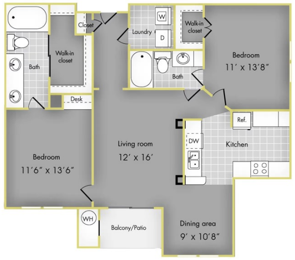
AutoTAMP results: Single-agent, time & geometric constraints

Solving puzzles

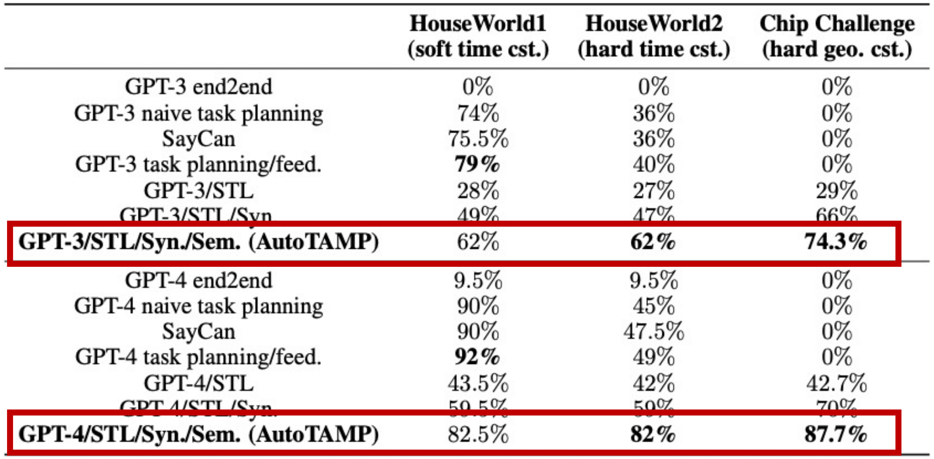


“Try to reach all the goals but you have to reach the corresponding key first to open the specific door. For example, you have to reach key1 ahead to open door1. Also, remember always do not touch the walls.”

Household work



“Navigate into the pink bedroom and then go to the yellow restroom, but remember do not touch the livingroom at anytime.”



	HouseWorld1 (soft time cst.)	HouseWorld2 (hard time cst.)	Chip Challenge (hard geo. cst.)
GPT-3 end2end	0%	0%	0%
GPT-3 naive task planning	74%	36%	0%
SayCan	75.5%	36%	0%
GPT-3 task planning/feed.	**79%**	40%	0%
GPT-3/STL	28%	27%	29%
GPT-3/STL/Syn.	49%	47%	66%
GPT-3/STL/Syn./Sem. (AutoTAMP)	62%	**62%**	**74.3%**
GPT-4 end2end	9.5%	9.5%	0%
GPT-4 naive task planning	90%	45%	0%
SayCan	90%	47.5%	0%
GPT-4 task planning/feed.	**92%**	49%	0%
GPT-4/STL	43.5%	42%	42.7%
GPT-4/STL/Syn.	59.5%	59%	70%
GPT-4/STL/Syn./Sem. (AutoTAMP)	82.5%	**82%**	**87.7%**

Task success rates of different approaches in single-agent TAMP (results obtained in May 2023).

AutoTAMP uses TL as the intermediate language, which enables auto-regressive reasoning and significantly improves the success rate of using LLM for TAMP when there are hard timing and geometric constraints.

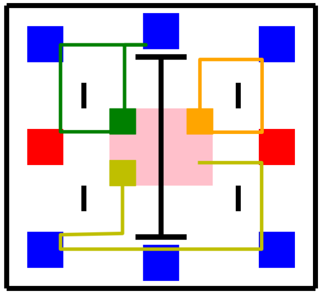
[1] AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. Chen et al., ICRA’24.

AutoTAMP results: Multi-agent, time & geometric constraints

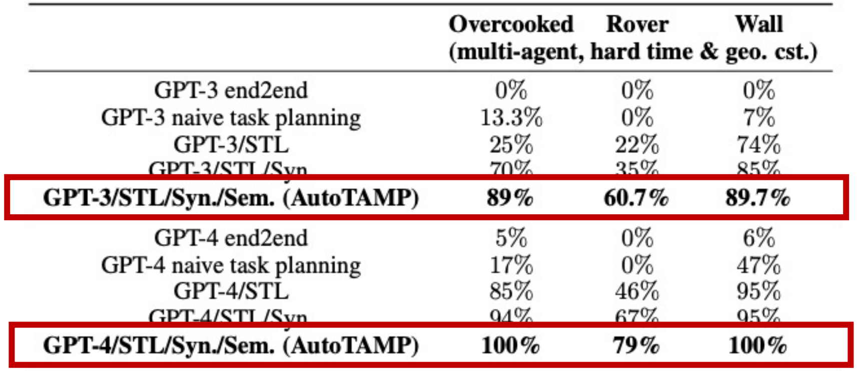
Overcooked game

"Enter all the ingredient room to pick up food. Once entered ingredient rooms, go to cooking room within 3 seconds. And all the agents should not collide each other and black obstacles."

Multi-agent surveillance



"All rovers must reach the pink charging station within 5 units of time each time they exit it. Once they reach their (blue) destination, they need to get to a red transmitter within 2-time units to send the collected information to the remote control. Rovers must keep clear of black walls and other rovers. All target areas need to be visited."



	Overcooked	Rover	Wall
	(multi-agent, hard time & geo. cst.)		
GPT-3 end2end	0%	0%	0%
GPT-3 naive task planning	13.3%	0%	7%
GPT-3/STL	25%	22%	74%
GPT-3/STL/Syn	70%	35%	85%
GPT-3/STL/Syn./Sem. (AutoTAMP)	**89%**	**60.7%**	**89.7%**
GPT-4 end2end	5%	0%	6%
GPT-4 naive task planning	17%	0%	47%
GPT-4/STL	85%	46%	95%
GPT-4/STL/Syn	94%	67%	95%
GPT-4/STL/Syn./Sem. (AutoTAMP)	**100%**	**79%**	**100%**

Task success rates of different approaches in multi-agent TAMP (results obtained in May 2023).

Replacing the TAMP planner with an any planner, this framework will interactively provide rigorous and explainable assistants for human operators, with improved success rates even with unstructured human user inputs.

[1] AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. Chen et al., ICRA'24.

Recap: structured view of “temporal” work

	Information extraction	**Question Answering**	**Reasoning**
Intrinsic to “time” (i.e., targeting a better understanding of time)	Time expressions Temporal relations Temporal grounding	Temporal order QA	Temporal relation graph inference Temporal logic
Extrinsic to “time” (i.e., solving tasks that require consideration of time)	Timeline summarization Temporal knowledge graphs Causality	Situated QA Time-sensitive QA Temporal QA over knowledge graphs	Robotic control

Recap: main papers covered in this talk

	Information extraction	**Question Answering**	**Reasoning**
Intrinsic to "time" (i.e., targeting a better understanding of time)	SteQE[1]	TORQUE[2]	NL2TL[3]
Extrinsic to "time" (i.e., solving tasks that require consideration of time)	Timeline summarization Temporal knowledge graphs Causality	Situated QA Time-sensitive QA Temporal QA over knowledge graphs	AutoTEMP[4]

[1] A Meta-framework for Spatiotemporal Quantity Extraction from Text. Ning et al., ACL'22.
[2] TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. Ning et al., EMNLP'20.
[3] NL2TL: Transforming Natural Languages to Temporal Logics using Large Language Models. Chen et al., EMNLP'23.
[4] AutoTAMP: Autoregressive task and motion planning with LLMs as translators and checkers. Chen et al., ICRA'24.

Summary

- We only focused on tasks that are intrinsic to time. Many existing works where "time" is supporting the main tasks are skipped.
  - SteQE exemplifies "what could happen if we move closer to the extreme of formalism"
  - TORQUE shows "what could happen if we use natural language to annotate natural language"
- "Time" is a good vehicle to study reasoning. LLM still struggles but can improve
  - Once we teach formalism to it (ILP, temporal logic, etc.)
  - Then we adopt solvers, tools, APIs to perform reasoning