UNIVERSITY of PENNSYLVANIA

C3SR center for cognitive computing systems research

IBM | ILLINOIS

A MULTI-AXIS ANNOTATION SCHEME FOR EVENT TEMPORAL RELATIONS



Qiang Ning, Hao Wu, and Dan Roth

07/17/2018

University of Illinois, Urbana-Champaign & University of Pennsylvania

COGNITIVE COMPUTATION GROUP

Towards Natural Language Understanding

1. 

2. 

3. 

4.

.....

.....

11. Reasoning about Time

TIME IS IMPORTANT

- **[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?

TIME IS IMPORTANT

- **[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.
  - When was Chris Robin born?

- **[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.
  - When was Chris Robin born?
    - Based on text: <=1922 (Wikipedia: 1920)
  - Requires identifying **relations** between events, and temporal reasoning.

poem [Chris at age 3] —before→ Winnie the Pooh [1925]

- Temporal relation extraction
  - "A" happens BEFORE/AFTER "B";
  - Events are associated with time intervals: $[t^1_{start}, t^1_{end}], [t^2_{start}, t^2_{end}]$
  - 12 temporal relations in every 100 tokens (in TempEval3 datasets)

"Time" could be expressed **implicitly**

Temporal Relations: A Key Component

- **Temporal Relation (TempRel):** *I* ***turned*** *off the lights and* ***left****.*
- **Challenges** faced by existing datasets/annotation schemes:
  - Low inter-annotator agreement (IAA)
    - TB-Dense: Cohen's κ 56%~64%
    - RED: F1<60%
    - EventTimeCorpus: Krippendorff's $\alpha \approx 60\%$
  - Time consuming: Typically, 2-3 hours for a single document.

- Our goal is to address these challenges,
  - And, understand the task of temporal relations better.

HIGHLIGHTS AND OUTLINE

What we did:

- **276 docs:** Annotated the 276 documents from TempEval3
- **1 week:** Finished in about one week (using crowdsourcing)
- **$10:** Costs roughly $10/doc
- **80%:** IAA improved from literature's 60% to 80%
- Re-thinking **identifying temporal relations** between events
  - Results in re-defining the temporal relations task, and the corresponding annotation scheme, in order to make it feasible
- **Outline of our approach (3 components)**
  - **Multi-axis:** types of events and their temporal structure
  - **Start & End points:** end-points are a source of confusion/ambiguity
  - **Crowdsourcing:** collect data more easily while maintaining a good quality

1. Temporal Structure Modeling: Existing Annotation Schemes

- *"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."*
- Task: to annotate the TempRels between the **bold** faced events (according to their start-points).

- Existing Scheme 1: General graph modeling *(e.g., TimeBank, ~2007)*
  - Annotators *freely* add TempRels between those events.
  - It's *inevitable* that some TempRels will be missed,
    - Pointed out in many works.
  - 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.

1. Temporal Structure Modeling: Existing Annotation Schemes

- *"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."*

- Existing Scheme 2: Chain modeling *(e.g., TimeBank-Dense ~2014)*
  - *All* event pairs are presented, one-by-one, and an annotator *must* provide a label for each of them.
  - *No* missing relations anymore.
  - *Rationale*: In the physical world, time is one dimensional, so we should be able to temporally compare any two events.
  - However, some pairs of events are very *confusing*, resulting in *low* agreement.
  - E.g., what's the relation between ***restore*** and ***killed***?

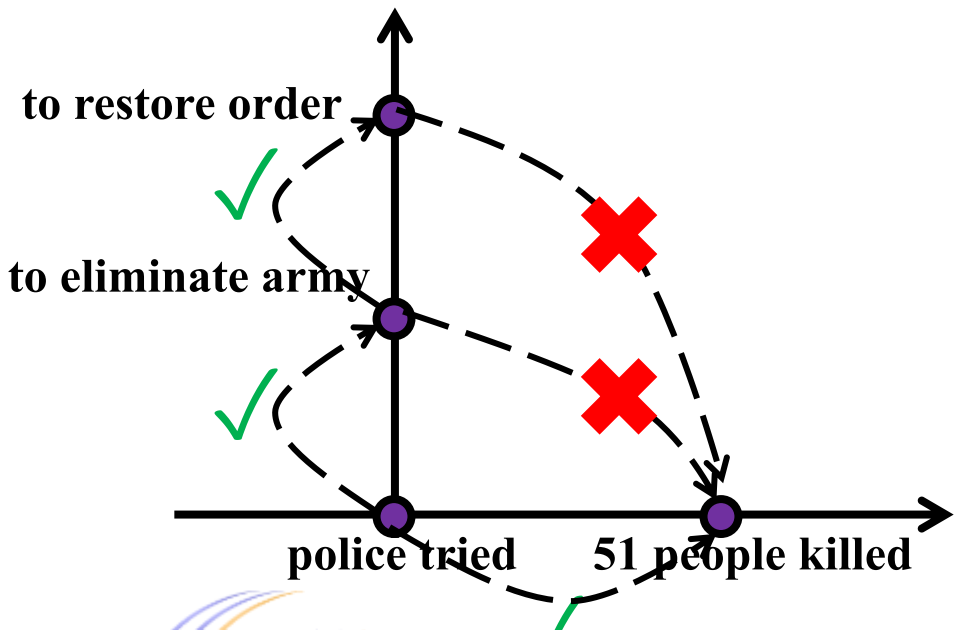
1. Temporal Structure Modeling: Difficulty

- *"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."*

- Why is ***restore*** vs ***killed*** confusing?
  - One possible explanation: the text doesn't provide evidence that the ***restore*** event actually happened, while ***killed*** actually happened
  - So, non-actual events don't have temporal relations?
- We don't think so:
  - ***tried*** is obviously before ***restore***: actual vs non-actual
  - ***eliminate*** is obviously before ***restore***: non-actual vs non-actual
  - So relations may exist between non-actual events.

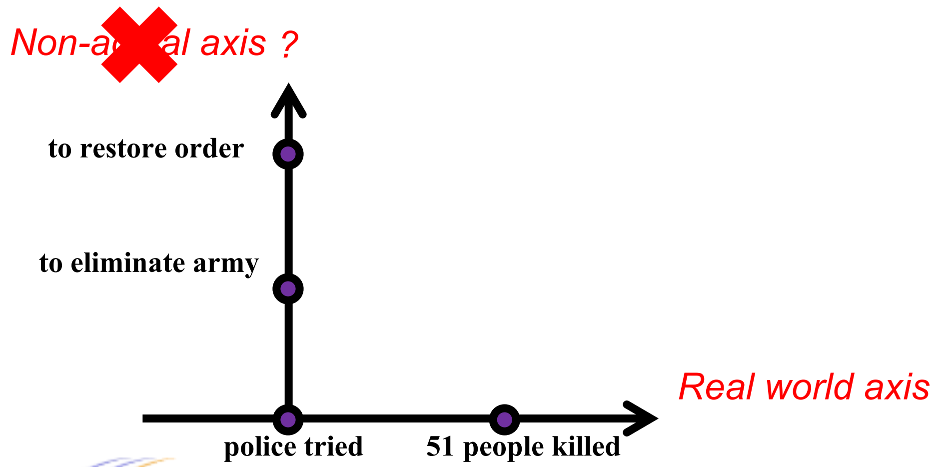
1. Temporal Structure Modeling: Multi-Axis

- *"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 while time is 1-dimensional in the physical world, **multiple temporal axes may exist in natural language**.



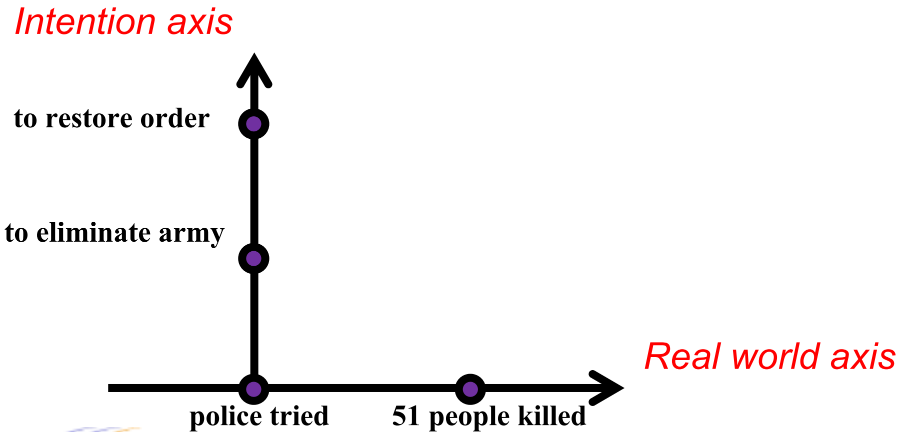
1. Multi-Axis Modeling: NOT Simply Actual Vs Non-Actual

- *"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."*
- Is it a "non-actual" event axis?—We think no.
  - First, ***tried, an actual event,*** is on both axes.
  - Second, whether ***restore*** is non-actual is questionable. It's very likely that order was indeed ***restored*** in the end.



1. Multi-Axis Modeling

- *"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."*
- Instead, we argue that it's an *Intention Axis*
- It contains events that are intentions: ***restore*** and ***eliminate***
  - and intersects with the real world axis at the event that invokes these intentions: ***tried***



Intention Vs Actuality

- Identifying "intention" can be done ***locally***, while identifying "actuality" often ***depends on other events***.

Text	Intention?	Actual?
I called the police to ***report*** *the body.*	Yes	Yes

1. Multi-Axis Modeling

- So far, we introduced the *intention* axis and distinguished it from (non-) *actuality* axis.

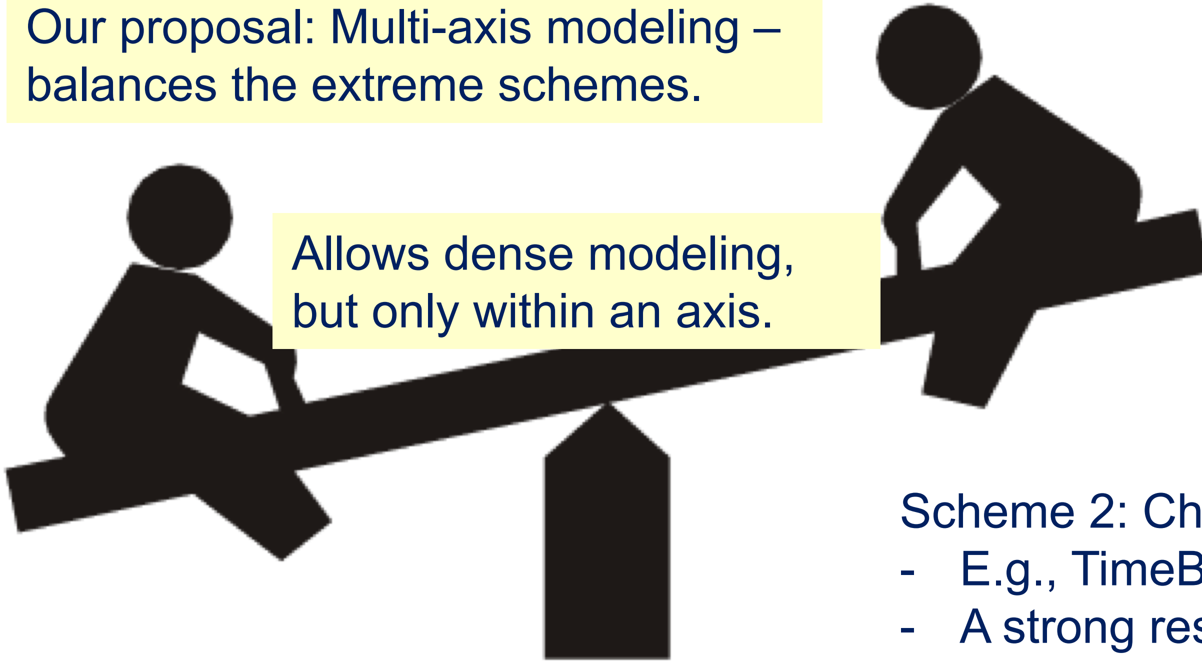
- The paper extends these ideas to more axes and discusses their difference form (non-)actuality axes
  - Sec. 2.2 & Appendix A; Sec. 2.3.3 & Appendix B.

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

1. Multi-Axis Modeling: A Balance Between Two Schemes

Our proposal: Multi-axis modeling – balances the extreme schemes.

Allows dense modeling, but only within an axis.



Scheme 2: Chain modeling
- E.g., TimeBank-Dense
- A strong restriction on modeling
- Any pair is comparable
- But many are confusing

Scheme 1: General graph modeling
- E.g., TimeBank
- No restrictions on modeling
- Relations are inevitably missed

OVERVIEW: MULTI-AXIS ANNOTATION SCHEME

- Step 0: Given a document in raw text
- Step 1: Annotate all the events
- Step 2: Assign axis to each event (intention, hypothesis, …)
- Step 3: On each axis, perform a "dense annotation" scheme

- In this paper, we use events provided by TempEval3, so we skipped Step 1.

- Our second contribution is successfully using crowdsourcing for Step 2 and Step 3, while maintaining a good quality.

2. Crowdsourcing

- Platform: CrowdFlower https://www.crowdflower.com/
- Annotation guidelines: Find at
  http://cogcomp.org/page/publication_view/834
- Quality control: A gold set is annotated by experts beforehand.
  - **Qualification**: Before working on this task, one has to pass with 70% accuracy on sample gold questions.
  - Important: with the older task definition, annotators did not pass the qualification test.
  - **Survival**: During annotation, gold questions will be given to annotators without notice, and one has to maintain 70% accuracy; otherwise, one will be kicked out and all his/her annotations will be discarded.
  - **Majority vote:** At least 5 different annotators are required for every judgement and by default, the majority vote will be the final decision.

3. An Interesting Observation: Ambiguity In End-Points

- Given two time intervals: $[t^1_{start}, t^1_{end}], [t^2_{start}, t^2_{end}]$

Metric	Pilot Task 1 t^1_{start} vs t^2_{start}	Pilot Task 2 t^1_{end} vs t^2_{end}	Interpretation
Qualification pass rate	50%	11%	Comparing the end-points is significantly harder than comparing the start-points.
Survival rate	74%	56%	
Accuracy on gold	67%	37%	
Avg. response time	33 sec	52 sec	Task 2 is also significantly slower.

- How durative events are expressed (by authors) and perceived (by readers):
  - Readers usually take longer to perceive durative events than punctual events, e.g., "***restore*** *order*" vs. "***try*** *to restore order*".
  - Writers usually assume that readers have a prior knowledge of durations (e.g., college takes 4 years and watching an NBA game takes a few hours)
- We only annotate start-points because duration annotation should be a different task and follow special guidelines.

Overview: Multi-Axis Annotation Scheme

- Step 0: Given a document in raw text
- Step 1: Annotate all the events
- Step 2: Assign axis to each event (intention, hypothesis, …)
- Step 3: On each axis, perform a “dense annotation” scheme according to events’ start-points

Quality Metrics Of Our New Dataset

		Step 2: Axis	Step 3: TempRel
Expert (~400 random relations)		$\kappa = 85\%$	$\kappa = 84\%, F_1 = 90\%$
Crowdsourcing (same docs in TBDense)	Accuracy	86%	88%
	Agreement (WAWA)	79%	81%

- Remember: Literature expert κ/F_1 values are around 60%
- For interested readers, please refer to our paper for more analysis regarding <u>each individual label</u>.
- <u>Worker Agreement With Aggregate (WAWA)</u>: assumes that the aggregated annotations are gold and then compute the accuracy.

Result On Our New Dataset

- We implemented a baseline system, using conventional features and the sparse averaged perceptron algorithm
- The overall performance on the proposed dataset is much better than those in the literature for TempRel extraction, which used to be in the low 50's (Chambers et al., 2014; Ning et al., 2017).
  - We do NOT mean that the proposed baseline is better than other existing algorithms
  - Rather, the proposed annotation scheme better defines the machine learning task.

Annotation	Training Set	Test Set	Training			Test		
			P	R	F	P	R	F
TBDense	Same-axis & Cross-axis	Same-axis	44	67	53	40	60	48
Proposed	Same-axis	Same-axis	**73**	**81**	**77**	**66**	**72**	**69**

CONCLUSION

Thank you!

- We proposed to re-think the important tasks **of identifying temporal relations**, resulting in a new annotation scheme it.
- Three components:
  - Multi-axis modeling: a balance between general graphs and chains
  - Identified that “end-point” is a major source of confusion
  - Showed that the new scheme is well-defined even for non-experts and crowdsourcing can be used.
- The proposed scheme significantly improves the inter-annotator agreement level, by ~20%.
- The resulting dataset defines an easier machine learning task.
- We hope that this work can be a good start for further investigation in this important area.