Penn COGNITIVE COMPUTATION GROUP

Capturing the Content of a Document through Complex Event Identification

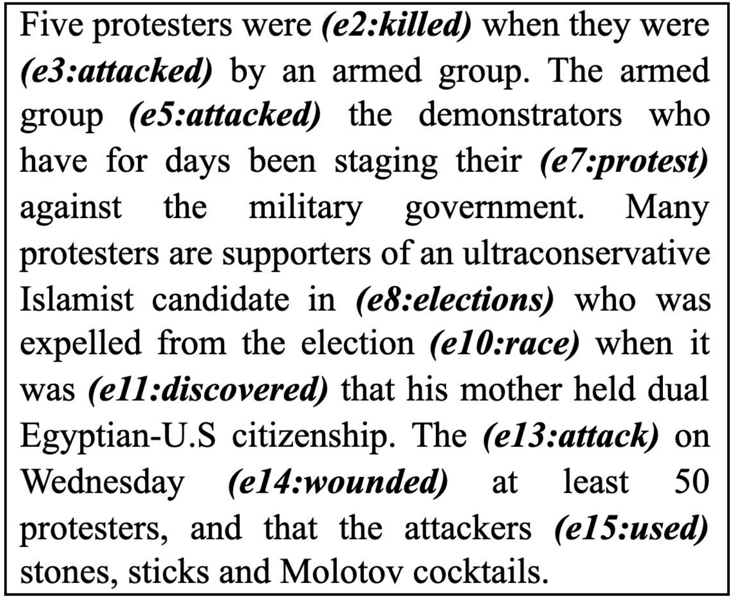
Zheng Qi, Elior Sulem, Haoyu Wang, Xiaodong Yu, Dan Roth

University of Pennsylvania

How to Capture the Content of a Document?

Considering each granular event, instantiated by predicates, in isolation is not sufficient

- Did “*e2:killed*” happen during “*e7:protest*” or “*e8:elections*”?
- Did “*e3:attacked*” and “*e14:wounded*” happen in the same “*e7:protest*”?



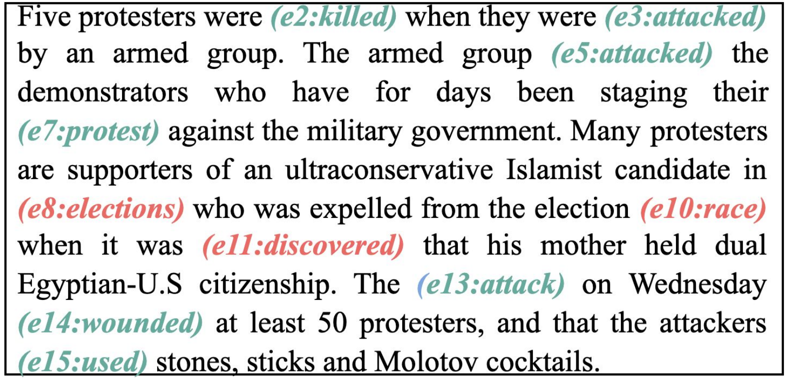
Five protesters were ***(e2:killed)*** when they were ***(e3:attacked)*** by an armed group. The armed group ***(e5:attacked)*** the demonstrators who have for days been staging their ***(e7:protest)*** against the military government. Many protesters are supporters of an ultraconservative Islamist candidate in ***(e8:elections)*** who was expelled from the election ***(e10:race)*** when it was ***(e11:discovered)*** that his mother held dual Egyptian-U.S citizenship. The ***(e13:attack)*** on Wednesday ***(e14:wounded)*** at least 50 protesters, and that the attackers ***(e15:used)*** stones, sticks and Molotov cocktails.

Granular events can be grouped into more general events

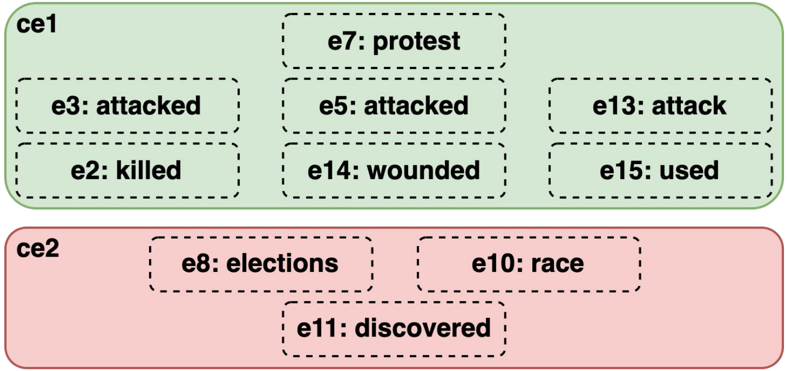
Together, these general events capture the major content of the document

Complex Event Identification: Solution to Capture General Events

- Complex Events are defined to be clusters of granular events that describe a more general event
- The sentences corresponding to the same complex event are not necessarily consecutive
- Granular events belonging to different complex events can appear in the same sentence



Five protesters were *(e2:killed)* when they were *(e3:attacked)* by an armed group. The armed group *(e5:attacked)* the demonstrators who have for days been staging their *(e7:protest)* against the military government. Many protesters are supporters of an ultraconservative Islamist candidate in *(e8:elections)* who was expelled from the election *(e10:race)* when it was *(e11:discovered)* that his mother held dual Egyptian-U.S citizenship. The *(e13:attack)* on Wednesday *(e14:wounded)* at least 50 protesters, and that the attackers *(e15:used)* stones, sticks and Molotov cocktails.



The Notions of Event-Event Relations and Event Region are Limited

- Temporal Relations
  - Only consider the ordering in time of granular events
- Causality
  - Only indicates if one granular event contributes to the occurrence of another one
- Event Coreference
  - Denotes two granular events refer to the same real-world event but at a much fine-grained level
  - "*e2:killed*" and "*e14:wounded*" belong to the same complex event but not co-referred
- Subevent Relations
  - The scope of the complex event depends on the existence of a high-level granular event
    - Billy Graham was *released* from the hospital after *recovering* from pneumonia
  - A pair of <event, subevent> may not belong to the same complex event
    - Simultaneous *raids* were conducted on company's offices. The *action* comes during the internal *turmoil* triggered by a *revolt* by his loyalists

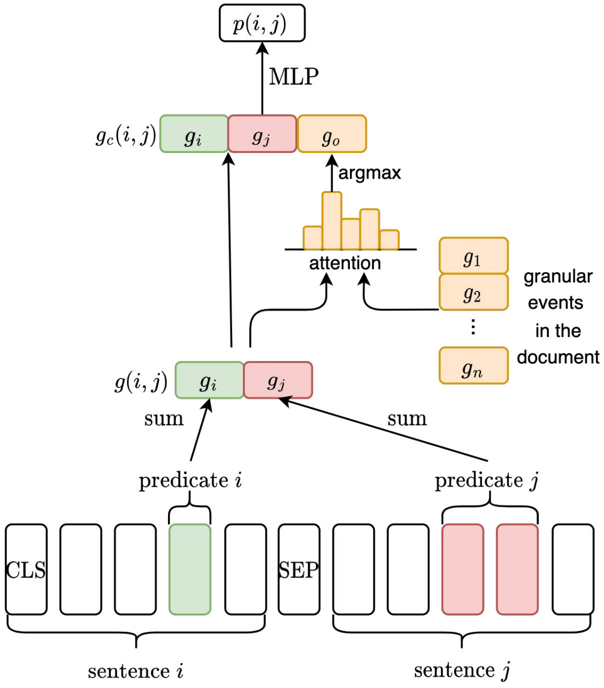
The Notions of Event-Event Relations and Event Region are Limited

- Event Region (Chen et al. 2020a)
  - A byproduct of document-level argument extraction, filtering out sentences irrelevant to the main content and then partitioning the text into several parts
  - Event regions have to be consecutive sentences that include relevant arguments

Context-augmented Predicate-based Approach

Step 1: Context-augmented Pairwise Complex Event Relation Prediction

- Issue: Whether two granular events belong to the same complex event also depends on the context of the document
- Solution: Adds an additional context granular event through the attention distribution over other granular events in the same document to model the pairwise complex event relation



Context-augmented Predicate-based Approach

Step 2: Agglomerative Clustering

- After predicting the pairwise complex event relation, we use agglomerative clustering to cluster granular events into complex events
- We define the distance between two granular events as the likelihood of not belonging to the same complex event
- We merge event clusters until no cluster pairs have a linkage distance lower than the threshold fine-tuned on the dev set

Experiments

Dataset:

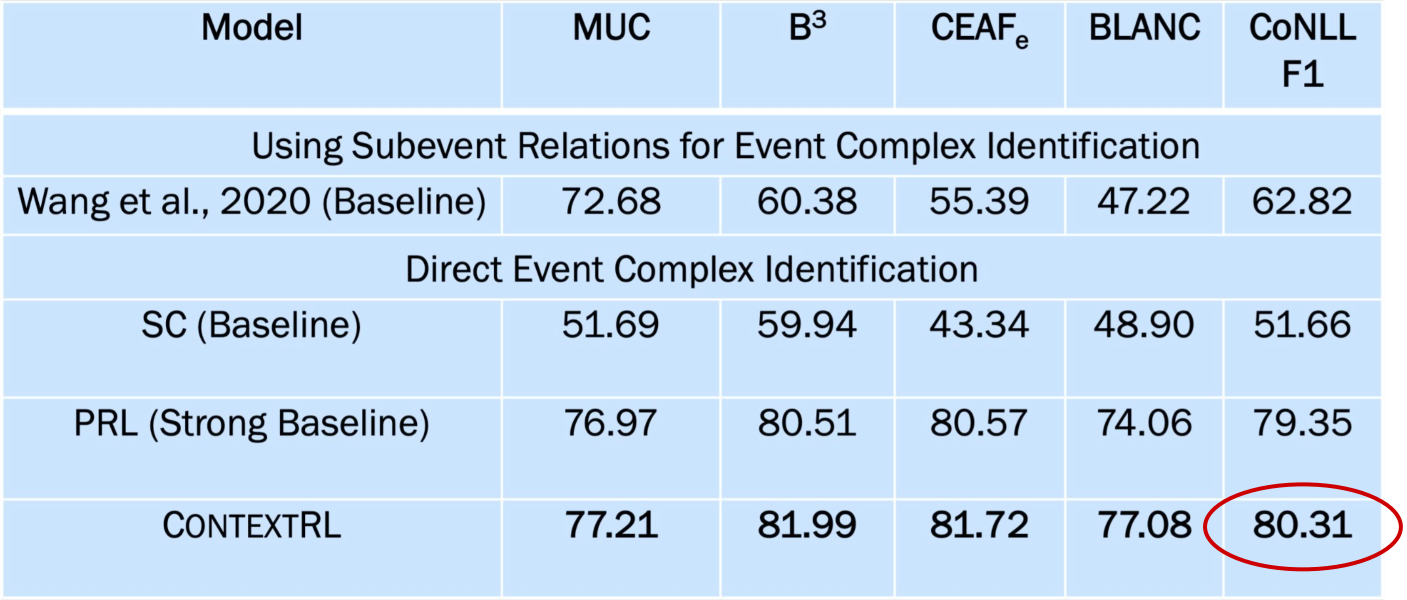
- We derive complex event annotation from HiEve Dataset (Glavaš et al., 2014) that annotates subevent relations and event coreference
- For each document, we build undirected acyclic graphs where vertices are granular events connected by subevent relations as edges and granular events in the same graph are in the same complex event

Baselines:

- **SC:** A simple Sequence Classification model plus the clustering step
- **PRL:** Paired Representation Learning (Yu, Yin and Roth, 2020) plus the clustering step
- **Wang et al., 2020:** predict subevent relations and follow the same graph-based clustering method as the dataset creation

	# Doc.	**# Pairs**	**# CE**	**# Events/CE**
Train	60	38124	121	7.01
Dev	20	13810	44	6.93
Test	20	16227	54	7.07

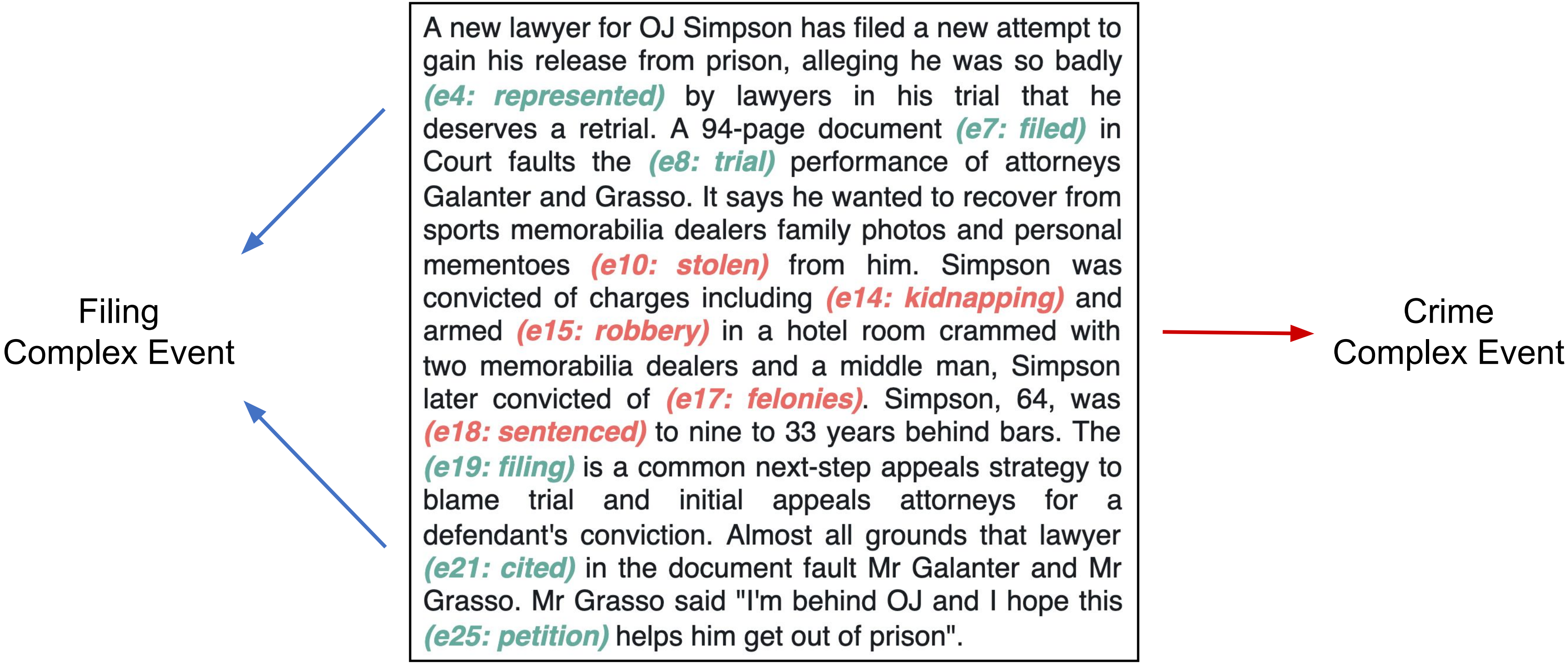
Experiment Results



Model	MUC	B^3	$CEAF_e$	BLANC	CoNLL F1
Using Subevent Relations for Event Complex Identification					
Wang et al., 2020 (Baseline)	72.68	60.38	55.39	47.22	62.82
Direct Event Complex Identification					
SC (Baseline)	51.69	59.94	43.34	48.90	51.66
PRL (Strong Baseline)	76.97	80.51	80.57	74.06	79.35
CONTEXTRL	**77.21**	**81.99**	**81.72**	**77.08**	**80.31**

Event complex identification performance on the event complex annotation derived from HiEve. The columns correspond to different cluster evaluation metrics. CoNLL F1 is the average of MUC, B^3 and $CEAF_e$. The highest score for each measure appears in bold.

Complex Event Prediction



A new lawyer for OJ Simpson has filed a new attempt to gain his release from prison, alleging he was so badly *(e4: represented)* by lawyers in his trial that he deserves a retrial. A 94-page document *(e7: filed)* in Court faults the *(e8: trial)* performance of attorneys Galanter and Grasso. It says he wanted to recover from sports memorabilia dealers family photos and personal mementoes *(e10: stolen)* from him. Simpson was convicted of charges including *(e14: kidnapping)* and armed *(e15: robbery)* in a hotel room crammed with two memorabilia dealers and a middle man, Simpson later convicted of *(e17: felonies)*. Simpson, 64, was *(e18: sentenced)* to nine to 33 years behind bars. The *(e19: filing)* is a common next-step appeals strategy to blame trial and initial appeals attorneys for a defendant's conviction. Almost all grounds that lawyer *(e21: cited)* in the document fault Mr Galanter and Mr Grasso. Mr Grasso said "I'm behind OJ and I hope this *(e25: petition)* helps him get out of prison".

Filing
Complex Event

Crime
Complex Event

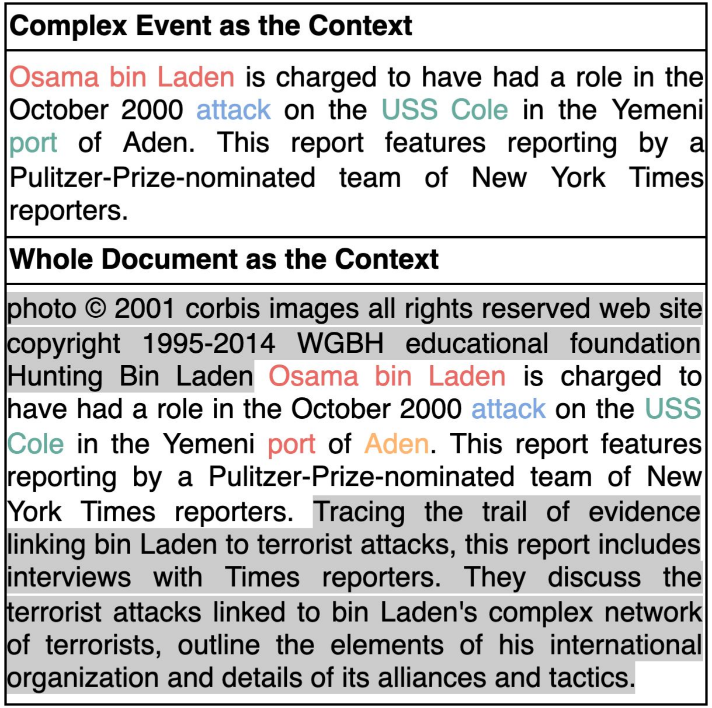
Complex Event as Input of Document-level Argument Extraction

- Using Complex Event as a starting point for document-level argument extraction.
- We applied our complex event identification system that was trained on HiEve, to the WikiEvents dataset (Li et al., 2021) and then used the BART-Gen model (Li et al., 2021) for argument identification and classification.

Context	Avg. Word Count	Argument Identification (F1)	Argument Classification (F1)
Entire document	787.90	71.21	66.55
Complex Event	539.25	71.07	66.25

Using Complex Events as the input

achieves a similar performance but a more concise context



Complex Event as the Context

Osama bin Laden is charged to have had a role in the October 2000 attack on the USS Cole in the Yemeni port of Aden. This report features reporting by a Pulitzer-Prize-nominated team of New York Times reporters.

Whole Document as the Context

photo © 2001 corbis images all rights reserved web site copyright 1995-2014 WGBH educational foundation Hunting Bin Laden Osama bin Laden is charged to have had a role in the October 2000 attack on the USS Cole in the Yemeni port of Aden. This report features reporting by a Pulitzer-Prize-nominated team of New York Times reporters. Tracing the trail of evidence linking bin Laden to terrorist attacks, this report includes interviews with Times reporters. They discuss the terrorist attacks linked to bin Laden's complex network of terrorists, outline the elements of his international organization and details of its alliances and tactics.

Conclusion & Future Work

- We introduce the complex event identification task that allows one to group related granular events into complex events that, together, capture the major content of the document
- We present a context-augmented representation learning approach ***ContextRL*** tailored to the task, showing that this approach outperforms strong baselines
- We conduct an exploratory case study on the ***WikiEvents*** dataset (Li et al., 2021), showing that using complex events as the input for document-level argument extraction obtains more concise context and achieves similar performance
- Future Work
  - Annotating a new Complex Event Dataset from scratch with fine-grained guidelines
  - Extending our approach towards an end-to-end system with granular event extraction

Thank you for listening

The data, code can be found at
http://cogcomp.org/page/publication_view/978