ahalt@mit.edu
https://andrewhalterman.com
github.com/ahalterman

A Word-Order-Aware Method for Extracting Events From Text

Andy Halterman (MIT Political Science)

Method–Question Mismatch

Bag-of-words text analysis methods like topic models answer:

- What's this document about?
- Who's involved?

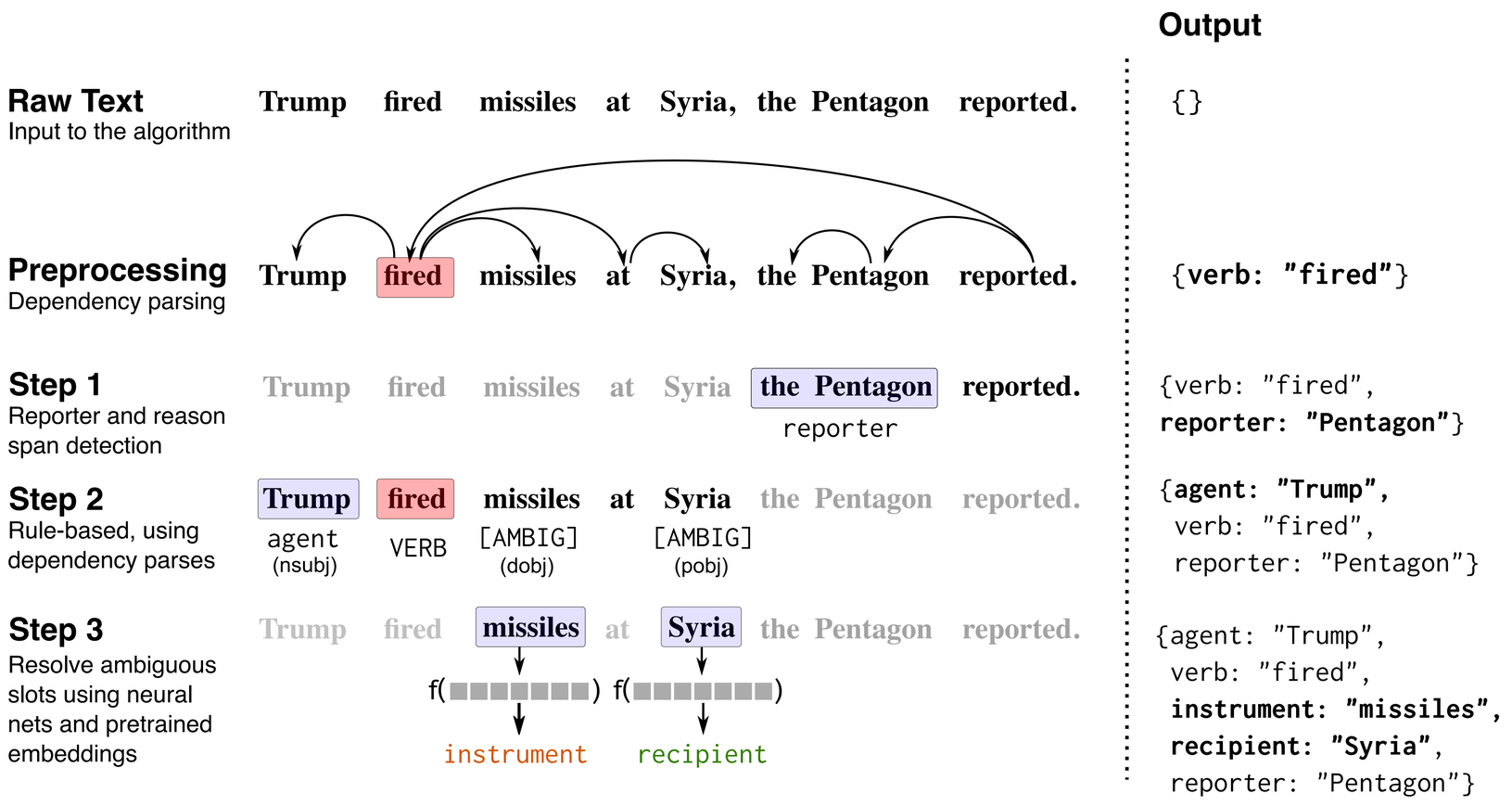
But empirical work in political science asks questions like:

- **Who** lobbied **whom** in Congress? (Kim 2017)
- **How**, or with what tactics, do resistance movements oppose the state (Chenoweth and Stephan 2011)?
- Against **whom** are laws enforced? (Holland 2015)
- **Where** and **how** did political violence occur? (Kalyvas 2006)

Word-order-aware methods, involving syntax, are needed to answer these questions with text.

Proposed Solution: Use Syntax and Semantics

I introduce a method for finding event properties using both **syntax** (the grammar of sentences) and **semantics** (the meaning of words).



Event Schema and Formalization

I frame the problem of event extraction as answering questions about properties using text.

- AGENT: Who did something?
- INSTRUMENT: How or with what was something done?
- RECIPIENT: To whom was something done?
- LOCATION: Where was something done?
- TIME: When was something done?
- REASON: Why was something done (reportedly)?
- REPORTER: According to whom?

- A corpus $\mathcal{X}$ is comprised of D documents $X_1 ... X_D$.
- Each document X_d is comprised of words: $X_d = \{x_1, ... x_{nd}\}$
- Each of J_d events e_{jd} in document d has one verb $v_{jd} \in X_d$.
- $A(v_d, S = s)$ is the set of words within X that correspond to event property s for verb $v \in V_d$.
  $A(v_d = \text{"fired"}, S = \text{AGENT}) = \text{"Trump"}$
  $A(v_d = \text{"fired"}, S = \text{RECIPIENT}) = \text{"Syria"}$

Why Both Syntax and Semantics?

Simple bag-of-words models won't capture the direction or properties of events ("Trump fired Tillerson" $\neq$ "Tillerson fired Trump"). Fully ML models require large amounts of (unavailable) training data. Instead, I use a hybrid approach:

- Use **rules** to get some spans using the grammatical dependency parse of the sentence (Step 2):
  - $A(v_d, S = \text{AGENT}) = \{$the "nsubj" words for $v_d\}$
- Use **neural networks** and pretrained embeddings to resolve ambiguous event properties: RECIPIENTS and INSTRUMENTS can be either direct objects or objects of prepositions ("fired Tillerson" vs. "fired missiles").
  - A classifier trained on 2,000 spans reaches F1 = 0.83 in distinguishing between them.
- Use other neural networks to identify REPORTERS, which are grammatically agents of a separate event, but should be included as a property of a first event.
  - Classifier reaches accuracy = 0.78 with n = 900 examples.)

Communal Violence and Police Response in India

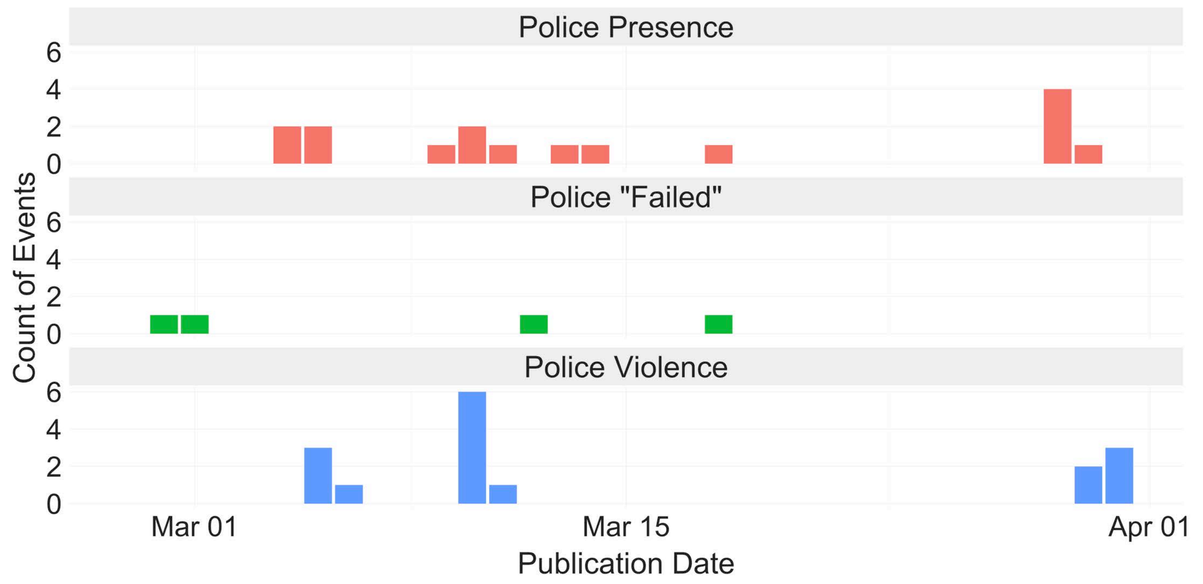
(Joint work with Katie Keith and Sheikh Muhammad Serwer, UMass Computer Science)

Wilkinson (2006) argues that **whether** police respond to communal violence in India determines how deadly it becomes. He draws on hand coded data on Hindu-Muslim violence in India from the *Times of India* (Varshney and Wilkinson 2006).

I create new data that records **how** security forces respond to communal violence in 2002.

- Scraped 8,600 articles from the *Times of India* in 2002 matching communal violence keywords.
- Applied the event extractor model, producing 222,000 events.
- Extracted 1,900 events with police as the agents.
- Clustered the extracted verbs + instruments using SIF embeddings (Arora et al. 2017) and k-means.

The findings reveal some heterogeneity in how police respond to communal violence in Gujarat, India beginning on 27 February. Initial police events consist of police *arresting* or *"failing to" act*. A week later, police engage in much more *patrolling*, *shooting* and other forms of violence.



{agent: "the task force, rapid action force, and the local police", verb: "have increased", instrument: "the patrolling"}

{agent: "the small posse of policemen", verb: "failed" instrument: "utterly to prevent the violence"}

{agent: "the police, which had remained inactive initially," verb: "beat up", recipient: "journalists and others"}