ahalt@mit.edu
http://ahalt.io
github.com/ahalterman

Learning Political Events From Text

Andy Halterman, PhD Candidate

Department of Political Science, Massachusetts Institute of Technology

How can we use text to find out what happens in the world?

Many methods exist for learning from text what people ***say***, but fewer for what they ***do***.

“Local media reported that a Georgian court in Tblisi sentenced a man to prison on Thursday for his role in a terror attack.”

How can we convert descriptions of events like this one into structured data?

Didn’t Phil Schrodt/linguists already figure this out?

Existing political science approaches work on pre-specified event types [1], but don’t generalize to new domains and don’t allow for new event types to be automatically learned from text.

Existing CS/linguistics approaches to “semantic role labeling” are either too general [2], too specific for political science events [3], require dictionaries for actors [4], or don’t learn new event types well.

New framework for event extraction:

Create a single event extration framework to be applied across domains and that learns event types automatically.

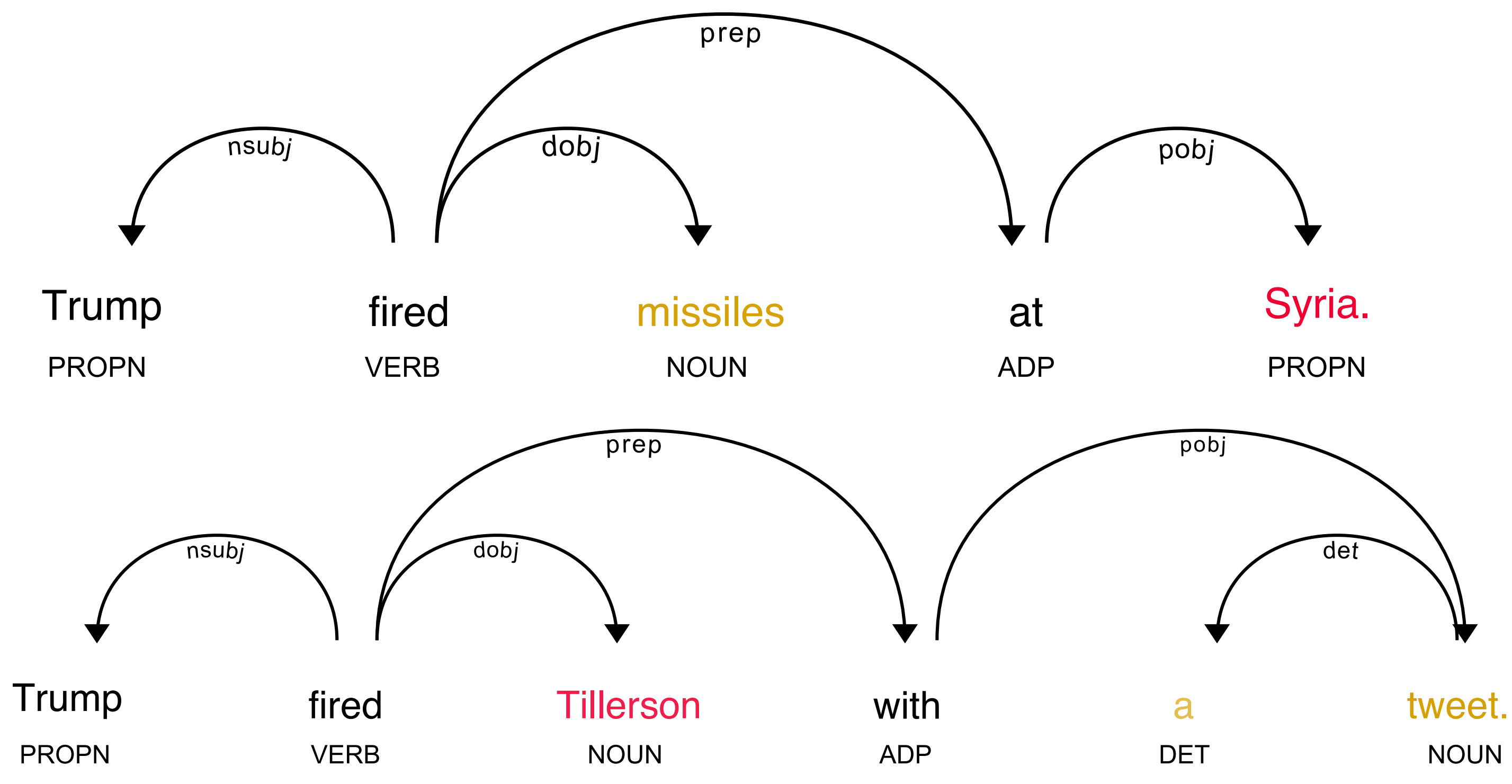
1. **Generalized Ontology**: What “slots” do all event types share?
2. **Slot filling**: Which words correspond to each event slot?
3. **Event Aggregation**: Which actions belong together as single event types?

Proposed event ontology

1. **actor**: the person/organization/role doing a political action
   “a Georgian court”
2. **action**: the verb and modifiers defining a political action
   “sentenced”
3. **recipient**: a political actor that the action is being done to
   “a man”
4. **means/instrument**: how or with what the action is being done.
   “to prison”
5. **reason/cause**: details on why the action was being carried out
   “for his role in a terror attack”
6. **location, date**: the location, date of the event
   “Tblisi”, “Thursday”
7. **reporter**: the reporter or source of the information in the event
   “Local media”

Slot filling algorithm

Dependency parse information gives *candidate* words for each slot: subjects are candidate actors, verbs are actions, objects are candidate recipients, passive sentences are reversed, etc.



The recipient and instrument words here demonstrate that syntactic information alone is not enough.

Use a convolutional neural network on pretrained embeddings, trained on 2,000 labeled spans, to disambiguate instruments and recipients with 85% accuracy.

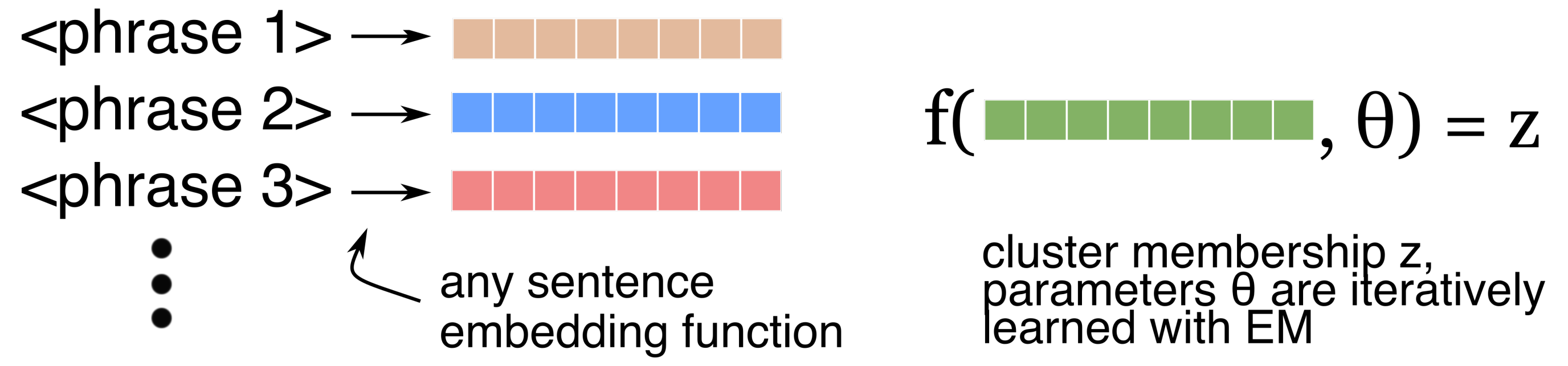
A separate model, trained on 500, recognizes reporter information. NER handles date, and [5] provides location.

Aggregating actions

How can we put extracted action+instrument spans into useful clusters of learned event types?

Incorporate prior information on semantics (word embeddings) and syntax (***w***(verbs) $\gg$ ***w***(adjectives)) to overcome the short document topic modeling problem.

Short document EM topic model:


<phrase 1> →
<phrase 2> →
<phrase 3> →
⋮
any sentence embedding function
f(, θ) = z
cluster membership z, parameters θ are iteratively learned with EM

Labels and labeling fuctions are iteratively learned. Using the SIF sentence embedding algorithm [6] with weighted logistic regression outperforms LDA in recovering synthetic span’s topics.

Application to human rights reporting

Extract event spans and aggregate actions from State Department reports to understand changing reporting.

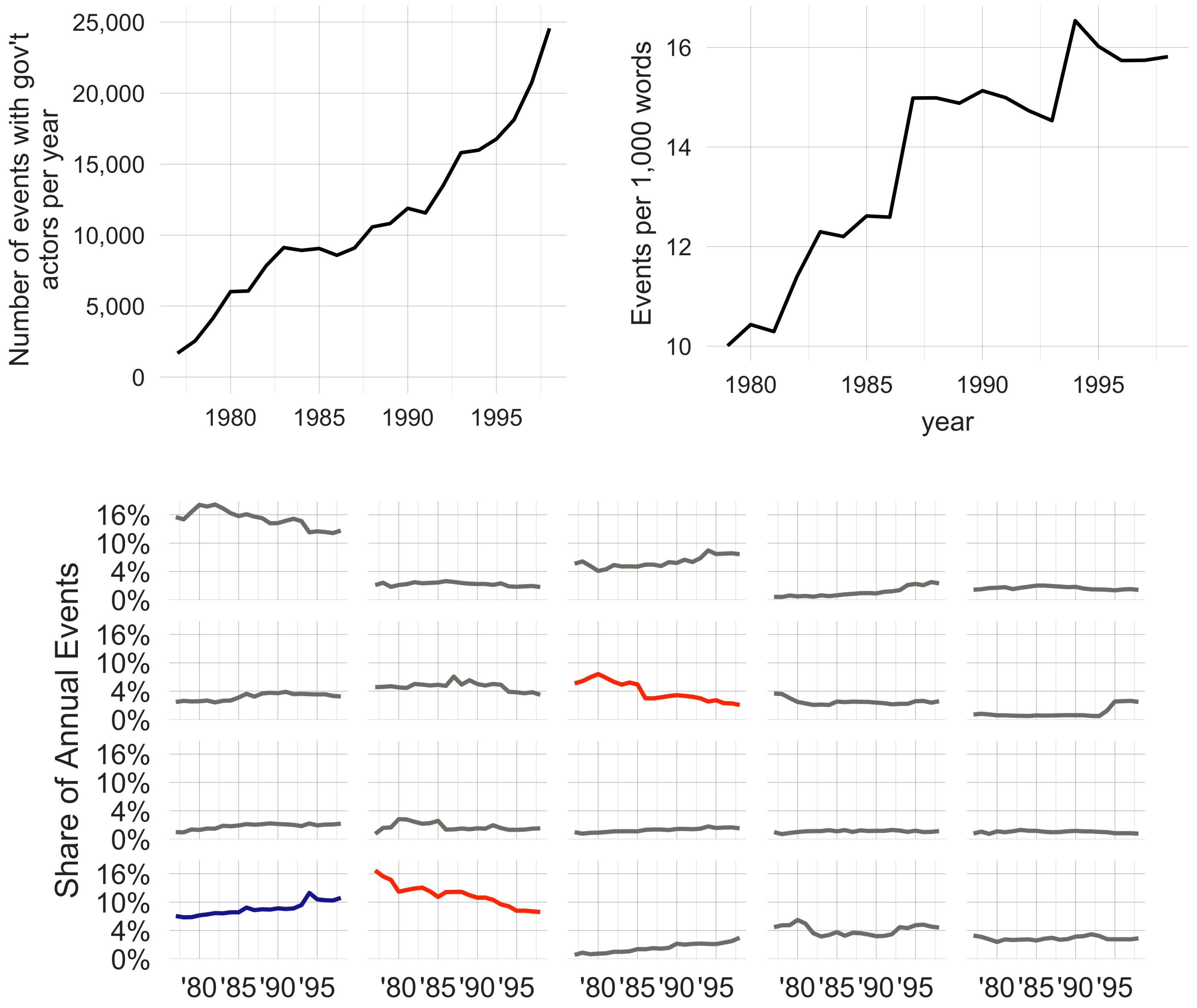


Figure: Total event counts are increasing (upper left) and reports are more event-focused (upper right). The trend is not uniform, however (lower). Positive/neutral/economic event types (red) are decreasing in proportion.

[“made internal economic reform and market stabilization”, “pursued a successful export-oriented agricultural growth strategy”, “own and run banking and insurance, air and rail transport, public utilities, and key industries”]

[’is a constitutional monarchy’, ’have lively and free, -, multiparty political systems’, ’is a multiparty democracy with mandatory universal suffrage’, ’is a representative democracy’]

[’provides for detention for an indefinite period without trial in national security cases’, ’not hold without a hearing before a magistrate’, ’when apprehend the accused during commission of a crime’, ’arrest persons without’, ’picked up and held on suspicion of robbery.’, ’remained in Zomba Central Prison’]

Acknowledgments & References

Thank you to Rich Nielsen, In Song Kim, Fotini Chrisia, and Blair Read for valuable feedback. I acknowledge support from an NSF Graduate Research Fellowship.

[1] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr, “Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions,” *International Studies Association, New Orleans*, 2002.

[2] M. Palmer, D. Gildea, and P. Kingsbury, “The proposition bank: An annotated corpus of semantic roles,” *Computational linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[3] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The berkeley FrameNet project,” in *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pp. 86–90, Association for Computational Linguistics, 1998.

[4] B. O’Connor, B. Stewart, and N. A. Smith, “Learning to extract international relations from political context,” *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. Vol. 1., 2013.

[5] A. Halterman, “Geolocating political events in text,” *NLP+CSS Workshop, NAACL*, 2019.

[6] S. Arora, Y. Liang, and T. Ma, “A simple but tough-to-beat baseline for sentence embeddings,” *ICLR*, 2017.