Corpus-Level Evaluation for Event QA

The IndiaPoliceEvents Corpus Covering the 2002 Gujarat Violence

Findings of ACL 2021

kaggle

Open Data Science Research Grant

We need real-world, full-corpus annotations and evaluation for automated event extraction

Police Responses to Communal Violence in 2002 India

- All *Times of India*
- Filter to March 2002 and “Ayodha” OR “Gujarat”
- Results in 1,257 articles (21,391 sentences)

Train fire kills Hindu Pilgrims, Feb. 27, 2002
Photo Credit: New York Times

Key Properties of Our Dataset

- Social science relevance
- Corpus-level full-recall
- Document-level context
- Natural language event specification
- High quality annotators

IndiaPoliceEvents Corpus

Semantic Event Class	Natural Language Question	Example	Num. Positive Sentences
Kill	“Did police kill someone?”	“In Vadodara, one person was killed in police firing on a mob in the Fatehganj area.”	96 (0.45%)
Arrest	“Did police arrest someone?”	“Police officials said nearly 2,537 people have so far been rounded up in the state.”	299 (1.40%)
Fail to Act	“Did police fail to intervene?”	“The news items [...] suggest inaction by the police force [...] to deal with this situation.”	207 (0.97%)
Force	“Did police use force or violence?”	“Trouble broke out in Halad [...] where the police had to open fire at a violent mob.”	222 (1.04%)
Any Action	“Did police do anything?”	“In the heart of the city’s Golwad area, the army is maintaining a vigil over mounting tension following [...]”	2,073 (9.69%)

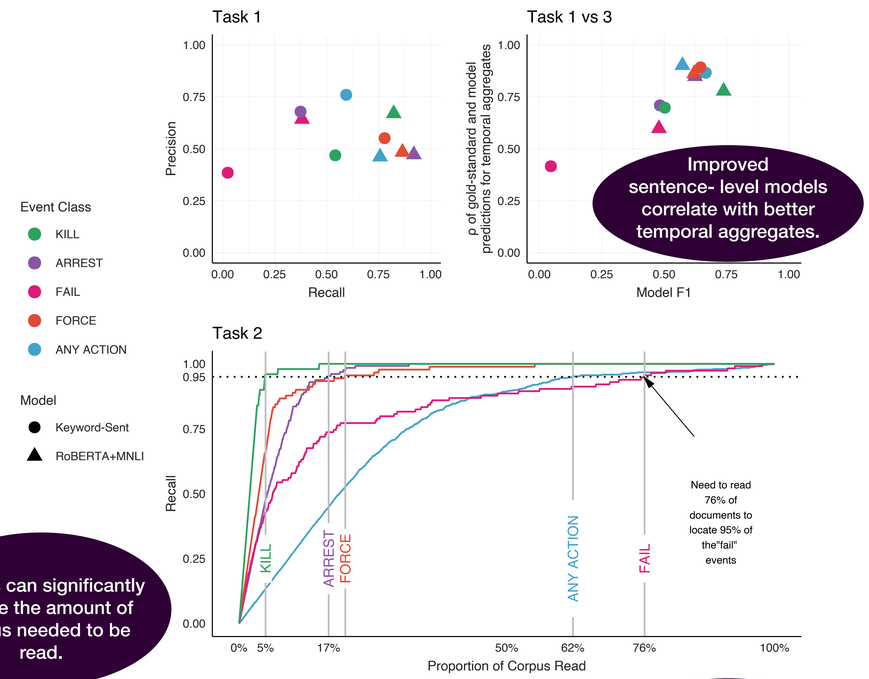
Models can significantly reduce the amount of corpus needed to be read.

Zero Shot Models

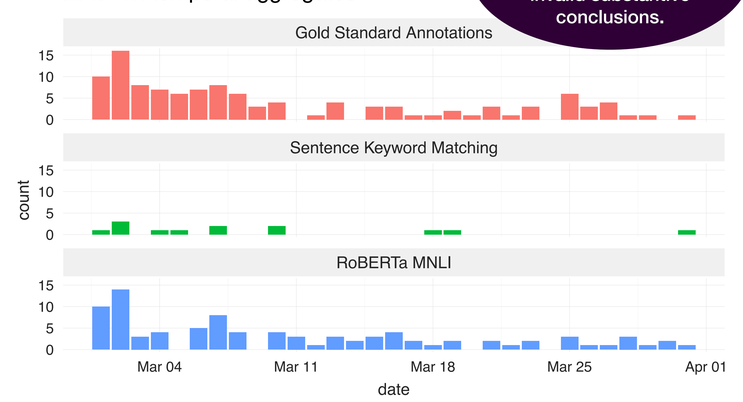
RoBERTa+MNLI	Uses a large-scale language model trained on MNLI (entailment, neutral, and contradiction classes) and declarative forms of the questions to classify sentences and documents.
Keywords	Uses Boolean queries on hand-constructed keywords matching police and events of interest to classify sentences and documents.
BM25+RM3	Automatically expands keywords in the natural language questions and uses the BM25 information retrieval model to rank documents
Electra+MS MARCO	Uses the ELECTRA variant of BERT fine tuned on the MS MARCO reading comprehension dataset to rank documents

Highlighted Results

- Task 1: Sentence classification
- Task 2: Document ranking
- Task 3: Substantive temporal aggregates



However, some models still undercount events, potentially leading to invalid substantive conclusions.



Andrew Halterman
Then: MIT, Political Science
Now: Faculty Fellow, NYU, Center for Data Science
andrew.halterman@nyu.edu

Katherine A. Keith
Then: University of Massachusetts Amherst, College of Information and Computer Sciences
Now: Postdoctoral researcher, AI2
kkeith@cs.umass.edu

Sheikh Muhammad Sarwar
University of Massachusetts Amherst, College of Information and Computer Sciences
smsarwar@cs.umass.edu

Brendan O’Connor
University of Massachusetts Amherst, College of Information and Computer Sciences
brenocon@cs.umass.edu

Data & Code: