Are *Triggers* Needed for Document-Level Event Extraction?

Shaden Shaar*, Wayne Chen*, Maitreyi Chatterjee, Barry Wang, Wenting Zhao and Claire Cardie
{ss2753, zc272}@cornell.edu

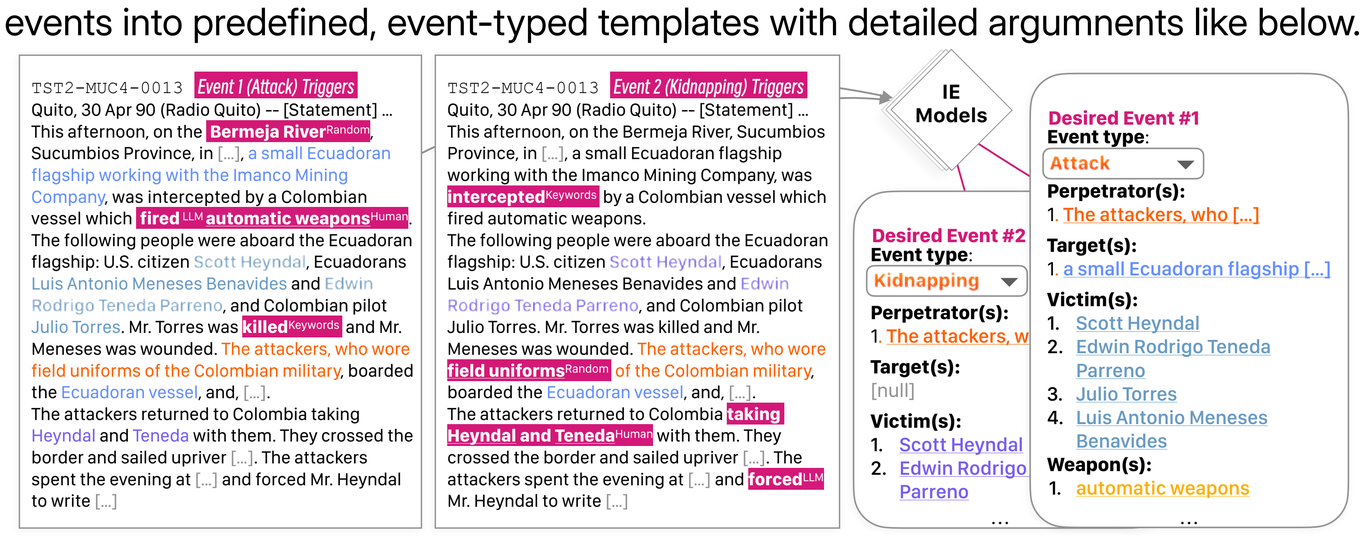
The Question

How important are pre-labeled *Triggers* today for accurate, auto event extraction?
Spoiler: Not so much.

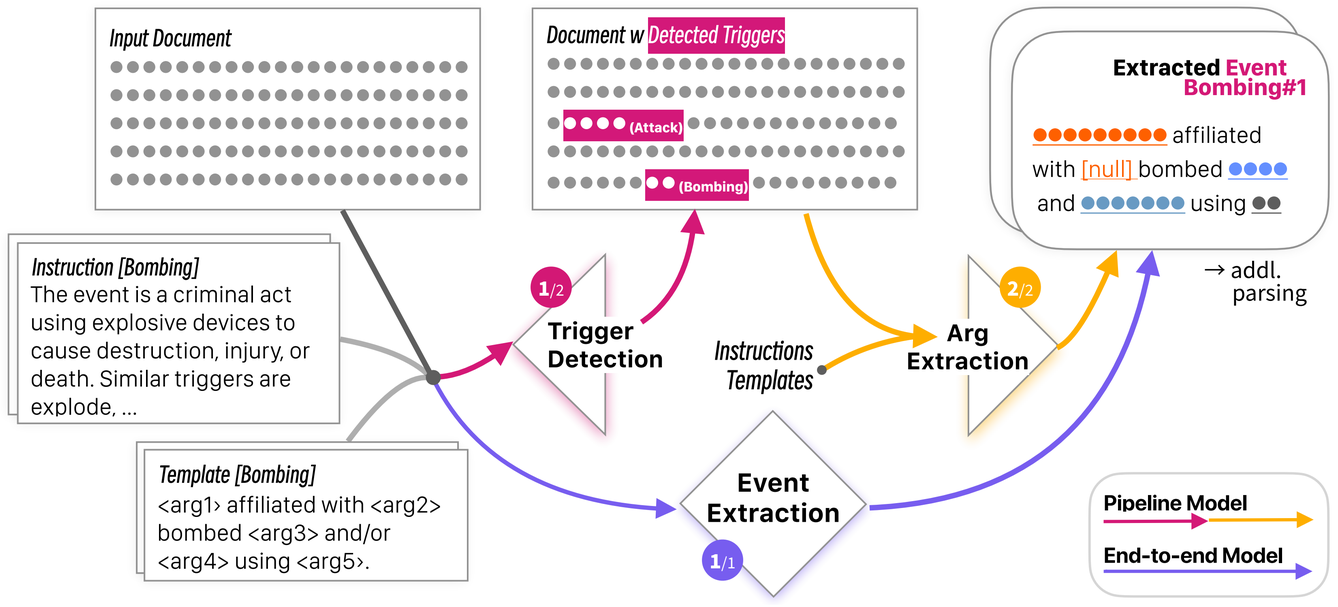
NEWS: At 12.23pm on Wednesday, as [...] Charlie Kirk was *struck fatally* by a bullet in the neck, sending thousands of screaming [...]

The Task: Document Level Event Extraction

One free text, one or more events. The task is to correctly identify and extract the events into predefined, event-typed templates with detailed argumnents like below.



Event Extraction Methods: E2E and Pipelined



The Study: Triggers v.s. Accuracy

Experiment	*MUC*	*WikiEvents*	*CMNEE*
		no *Triggers*	
GPT-4o-mini$_{prompting}$	$.434_{\pm.01}$	$.325_{\pm.02}$	$.565_{\pm.00}$
GPT-4o$_{prompting}$	$.391_{\pm.01}$	$.339_{\pm.02}$	$.635_{\pm.00}$
TANL$_{e2e}$	$.559_{\pm.01}$	$.361_{\pm.02}$	$.438_{\pm.00}$
DEGREE$_{e2e}$	$.597_{\pm.00}$	$.248_{\pm.01}$	$.789_{\pm.00}$
GTT$_{e2e}$	$.454_{\pm.01}$	—	$.318_{\pm.00}$
		Human *Triggers*	
GPT-4o-mini$_{prompting}$	$.438_{\pm.02}$	$.364_{\pm.02}$	$.564_{\pm.00}$
GPT-4o$_{prompting}$	$.405_{\pm.01}$	$.379_{\pm.02}$	$.667_{\pm.00}$
TANL$_{pipeline}$	$.460_{\pm.01}$	$.471_{\pm.02}$	$.596_{\pm.01}$
DEGREE$_{pipeline}$	$.524_{\pm.01}$	$.357_{\pm.01}$	$.723_{\pm.00}$
TANL+GENIE$_{pipeline}$	$.464_{\pm.01}$	$.382_{\pm.01}$	$.599_{\pm.00}$
TANL$_{e2e}$	$.530_{\pm.01}$	$.337_{\pm.02}$	$.410_{\pm.00}$
DEGREE$_{e2e}$	$.566_{\pm.01}$	$.339_{\pm.02}$	$.782_{\pm.00}$
GTT$_{e2e}$	$.429_{\pm.01}$	—	$.326_{\pm.00}$
		LLM *Triggers*	
GPT-4o-mini$_{prompting}$	$.416_{\pm.02}$	$.368_{\pm.02}$	$.580_{\pm.00}$
GPT-4o$_{prompting}$	$.414_{\pm.01}$	$.383_{\pm.02}$	$.659_{\pm.00}$
TANL$_{pipeline}$	$.305_{\pm.01}$	$.207_{\pm.01}$	$.543_{\pm.00}$
DEGREE$_{pipeline}$	$.567_{\pm.01}$	$.335_{\pm.01}$	$.715_{\pm.00}$
TANL+GENIE$_{pipeline}$	$.490_{\pm.01}$	$.388_{\pm.01}$	$.558_{\pm.00}$
TANL$_{e2e}$	$.507_{\pm.00}$	$.267_{\pm.02}$	$.440_{\pm.00}$
DEGREE$_{e2e}$	$.557_{\pm.00}$	$.313_{\pm.02}$	$.703_{\pm.00}$
GTT$_{e2e}$	$.439_{\pm.02}$	—	$.332_{\pm.00}$
		Keyword *Triggers*	
GPT-4o-mini$_{prompting}$	$.429_{\pm.02}$	$.361_{\pm.02}$	$.582_{\pm.00}$
GPT-4o$_{prompting}$	$.420_{\pm.01}$	$.389_{\pm.02}$	$.659_{\pm.00}$
TANL$_{pipeline}$	$.525_{\pm.02}$	$.306_{\pm.02}$	$.640_{\pm.00}$
DEGREE$_{pipeline}$	$.532_{\pm.01}$	$.241_{\pm.01}$	$.532_{\pm.00}$
TANL+GENIE$_{pipeline}$	$.470_{\pm.01}$	$.316_{\pm.01}$	$.628_{\pm.00}$
TANL$_{e2e}$	$.495_{\pm.00}$	$.292_{\pm.02}$	$.443_{\pm.00}$
DEGREE$_{e2e}$	$.523_{\pm.01}$	$.273_{\pm.01}$	$.601_{\pm.00}$
GTT$_{e2e}$	$.455_{\pm.01}$	—	$.324_{\pm.00}$
		Random *Triggers*	
GPT-4o-mini$_{prompting}$	$.412_{\pm.01}$	$.369_{\pm.02}$	$.533_{\pm.01}$
GPT-4o$_{prompting}$	$.402_{\pm.00}$	$.359_{\pm.02}$	$.533_{\pm.01}$
TANL$_{pipeline}$	$.116_{\pm.01}$	$.059_{\pm.01}$	$.223_{\pm.00}$
DEGREE$_{pipeline}$	$.538_{\pm.01}$	$.255_{\pm.01}$	$.589_{\pm.00}$
TANL+GENIE$_{pipeline}$	$.322_{\pm.01}$	$.120_{\pm.01}$	$.487_{\pm.00}$
TANL$_{e2e}$	$.473_{\pm.01}$	$.343_{\pm.02}$	$.394_{\pm.00}$
DEGREE$_{e2e}$	$.514_{\pm.01}$	$.270_{\pm.02}$	$.595_{\pm.01}$
GTT$_{e2e}$	$.304_{\pm.01}$	—	$.310_{\pm.00}$

Quick observations:

- ICL (prompting) < smaller-scale fine-tuned models
- The presence of triggers, even random ones, is generally helpful for prompting eapproaches.

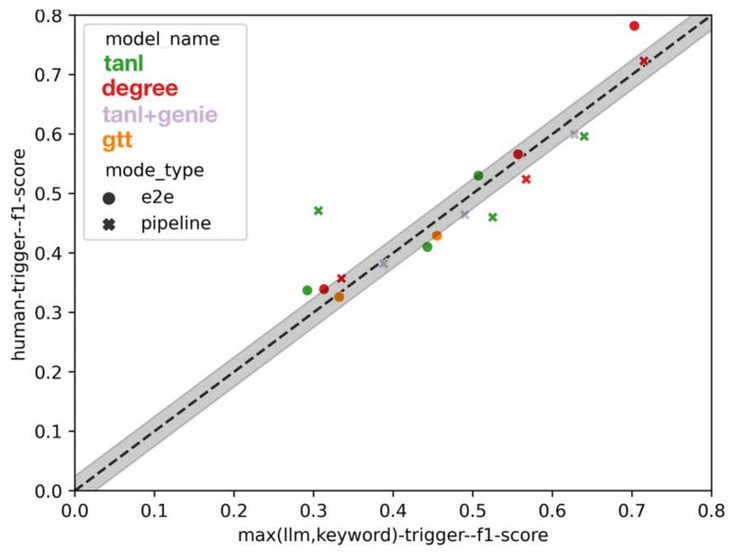
The Event Trigger: Types

NONE:	Passing `null` as triggers
HUMAN:	Author annotated
LLM:	GPT-4-Turbo identified
KEYWORD:	Hardcoded keywords
RANDOM:	Random 1~3 words in doc

The Event Extract Datasets

MUC is an event dataset of 1700 news articles reporting Latin American terrorist incidents.
WikiEvents has 246 event-rich documents
CMNEE is an annotated dataset on Defense news. We translate it from Chinese to English.

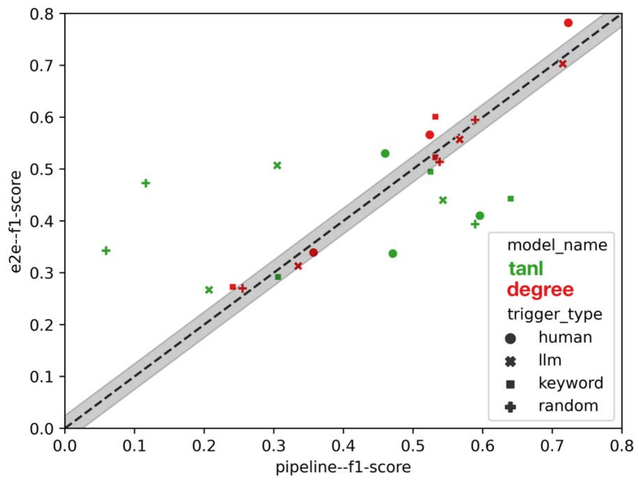
Accuracy: HUMAN Triggers?



For fine-tuning methods:

- Trigger quality does not significantly impact overall performance.
- **Lower quality trigger annotations that are less expensive to obtain can be used as an alternative to HUMAN triggers.**

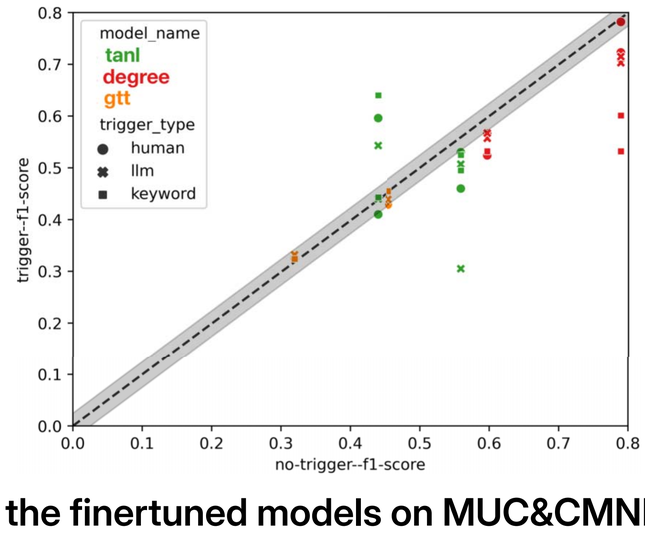
Accuracy: E2E v.s. Pipeline



For the finertuned models:

- Upper left points mean E2E > Pipeline
- We found that **E2E ≈ pipeline models** when trained and tested using the same model types

Accuracy: Triggers v.s. `null`



For the finertuned models on MUC&CMNEE:

- Upper left points mean triggers help.
- **Triggers are not needed for corpora that have few per-documentevents to extract.**
- For *Wikievent*, the pattern is less clear.

Published @TACL
Presenting @EMNLP 2025 Suzhou, China 中国苏州
Presenter &Poster Credit
Barry Wang 苏州er
barryw@cs.cmu.edu