Cross-document Event Identity via Dense Annotation

Adithya Pratapa, Zhengzhong Liu, Kimihiro Hasegawa, Linwei Li,
Yukari Yamakawa, Shikun Zhang and Teruko Mitamura.

CoNLL 2021

Carnegie Mellon University
Language
Technologies
Institute

Coreference

- Recasens et al., 2011 defines as “identity of reference”
- Understanding coreference of entities and events is important to numerous NLP applications

Marta Recasens, E. Hovy, and M. A. Martí. *Identity, non-identity, and near-identity: Addressing the complexity of coreference*. Lingua 2011.

Identity of Reference

- This identity is often a *continuum* (Recasens et al., 2011, Hovy et al., 2013)
  - No-identity, near-identity (or partial-identity) and full-identity
- Partial identity is still understudied in cross-document event coreference

Marta Recasens, E. Hovy, and M. A. Martí. *Identity, non-identity, and near-identity: Addressing the complexity of coreference*. Lingua 2011.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. *Events are not simple: Identity, non-identity, and quasi-identity*. Workshop on Events: Definition, Detection, Coreference, and Representation 2013.

An illustration of partial identity of event mentions from proposed dataset

Nearly 200 dead in Haitian cholera outbreak

Nearly 200 people are confirmed dead and approximately 2600 are ill in a central Haitian cholera **outbreak**, according to the Centers for Disease Control (CDC), the U.S. Agency for International Development (USAID) and United Nations (UN). Haitian officials place the death toll at 194 deaths with 2,364 people being infected.

…

October 23, 2010

Over 250 dead in Haiti cholera outbreak, thousands infected

At least 259 people are dead and over 3000 people have been infected in the Haitian cholera **outbreak**. Officials from the United Nations have said that they fear that the disease will spread across the entire country.

…

October 26, 2010

'Explosive' Haitian cholera outbreak kills 292, neighboring countries prepare

The Haitian cholera **outbreak** has killed 292 people and infected over 4000, according to the Haitian government, although there are no new cases in the earthquake ravaged capital, Port-au-Prince.

…

October 28, 2010

Spatiotemporal Continuity

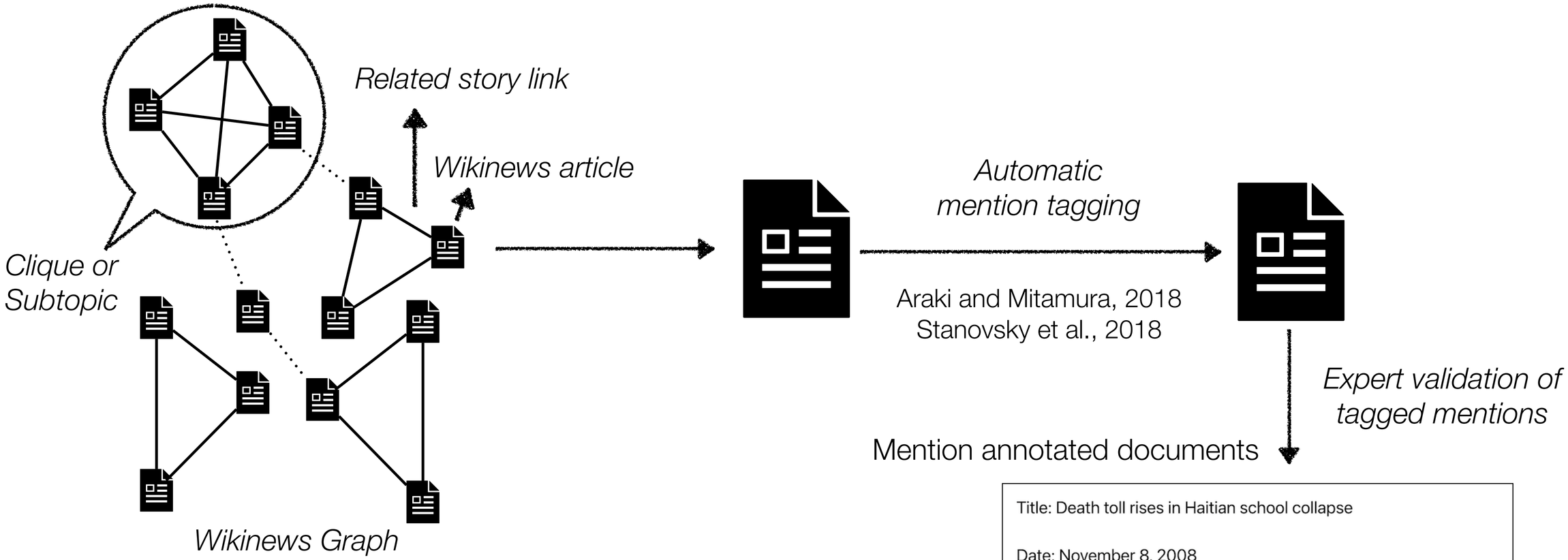
Limitations of Existing Datasets

- Do not account for the identity continuum during annotation
- Limits annotation to selected event types

Our Dataset: CDEC-WN

- A **C**ross-**D**ocument **E**vent **C**oreference dataset based on English **W**iki**n**ews (CDEC-WN)
- Our annotation framework presents two benefits,
    1. Facilitates dense annotation of coreference
    2. Collect evidence to identify cases of partial identity
- Annotations are crowd sourced on Mechanical Turk

Corpus Preparation



Second school in Haiti collapses

Thursday, November 13, 2008

Less than one week after the College La Promesse Evangelique in Pétionville, Haiti collapsed, a second school in the Haitian capital, Port-au-Prince, has partially collapsed injuring nine people.

The Grace Divine school partially collapsed while school was in session, but no one was trapped or killed. At least two people were transported to an area hospital with

Related news

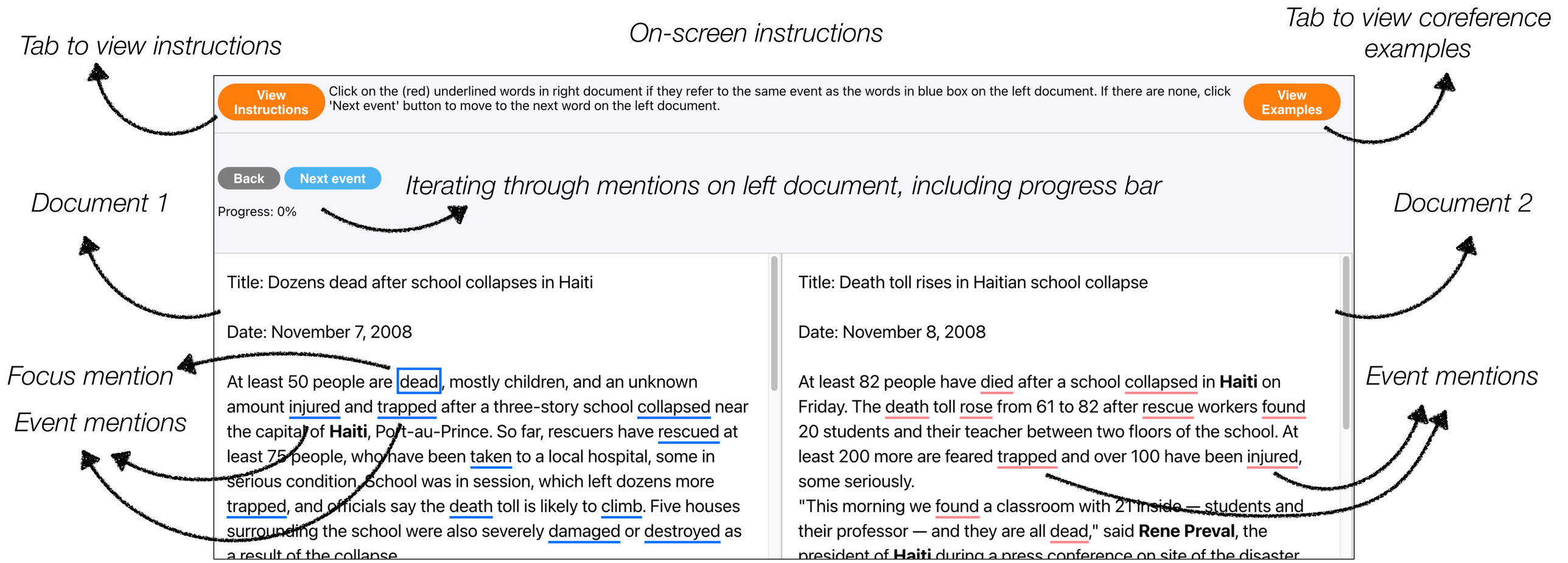
- "Death toll rises in Haitian school collapse" — *Wikinews*, November 8, 2008
- "Dozens dead after school collapses in Haiti" — *Wikinews*, November 7, 2008

Title: Death toll rises in Haitian school collapse

Date: November 8, 2008

At least 82 people have died after a school collapsed in **Haiti** on Friday. The death toll rose from 61 to 82 after rescue workers found 20 students and their teacher between two floors of the school. At least 200 more are feared trapped and over 100 have been injured, some seriously.
"This morning we found a classroom with 21 inside — students and their professor — and they are all dead," said **Rene Preval**, the president of **Haiti** during a press conference on site of the disaster

Annotation Interface



For the highlighted event mention on the left document, select all corefering mentions from the right document.

Annotation Task

- Annotate mentions pairs across every document pair ⇢ Dense annotation
- For coreference links, we ask the evidence questions
  - Mention share location/time/participants?
  - One mention is part-of another?

Dataset Summary

- 198 document pairs with over 4k links
- Dense set of mentions
  - 41 mentions/doc vs 15.3 in ECB+
- 3 annotators per document pair
  - Krippendorff's α of 0.46

# topics	1
# subtopics	55
# documents	176
# document pairs	198
# event mentions	7220
# CDEC links	4282
# CDEC links per doc.	21.6
# annotators	46

Baselines

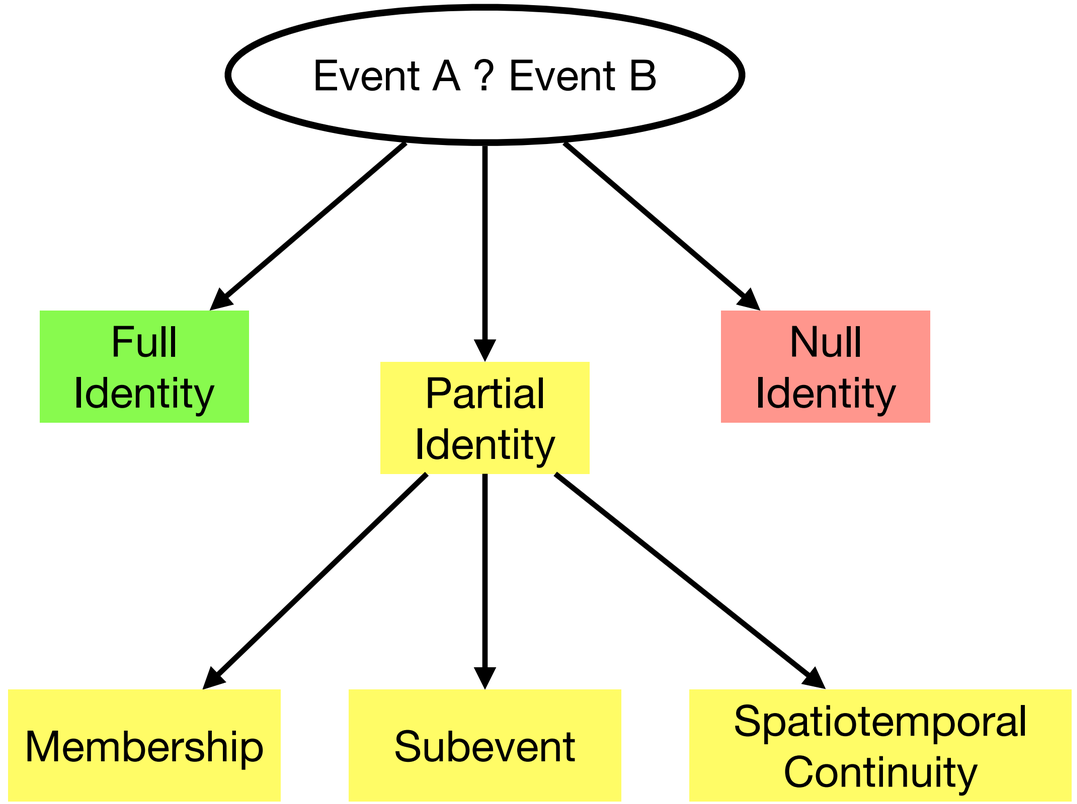
- Two baseline models to predict coreference vs non-coreference
  - Lemma match between mentions
  - BERT-based cross-encoder

Baseline	**Dev**			**Test**		
	P	R	F1	P	R	F1
Lemma-match	46.6	54.9	49.9	42.3	56.0	48.2
Cross encoder	43.1 ±0.6	75.4 ±0.5	54.3 ±0.5	45.9 ±0.8	77.3 ±1.1	57.6 ±0.6

The Case of Partial Identity

- Analyze the responses to evidence questions (location, time, participants and part-of)
- If majority of the 3 annotators say,
  - Location/time/participants overlap
  - There is a part-of relationship

A Taxonomy of Event Identity



Membership

October 22, 2007

The **fire** has burned about 4400 acres so far and 15 homes have been lost, however there have been no reported injuries or deaths.

October 24, 2007

Reports say that the amount of people fleeing from their homes in California located in the United States due to **wildfires** has reached the 1,000,000 mark as the fires continue to grow.

$$\text{fire} \subset \text{wildfires}$$

Subevent

July 18, 2007

A freight train in Lviv, Ukraine derailed, caught fire, and spilled a toxic chemical, releasing dangerous fumes into the air early Tuesday morning (local time), and people who live near the site of the **crash** are still becoming sick.

July 21, 2007

The available information about the phosphorous cloud following the railway **accident** in the Ukraine last Monday is becoming more and more cryptic.

crash $\subset$ accident

Spatiotemporal Continuity

October 24, 2010

Tropical **storm** Richard is nearing hurricane strength with winds of 70mph (115kph) as it lashes Honduras with heavy rains

October 25, 2010

Hurricane Richard made landfall in Belize about 20 mi (35 km) south-southeast of Belize City with winds of 90 mph (150 kph) at approximately 6:45 local time (0045 UTC) according to the National Hurricane Center (NHC)

Evolving identity of the event

Status	Storm → Hurricane
Wind speeds	70mph → 90mph
Location	Honduras → Belize

Summary

- CDEC-WN: a dataset for cross-document event coreference in English
- Annotation framework that accounts for partial identity of mentions
- Dense annotation of mentions: spanning open-domain events
- Evidence for a new type of partial identity, spatiotemporal continuity

github.com/adithya7/cdec-wikinews