Events are Not Simple: Identity, Non-Identity, and Quasi-Identity

Eduard Hovy* Teruko Mitamura* Felisa Verdejo** Jun Araki* Andrew Philpot#

*Carnegie Mellon University, USA **UNED, Spain #University of Southern California, USA

Problems of Event Coreference

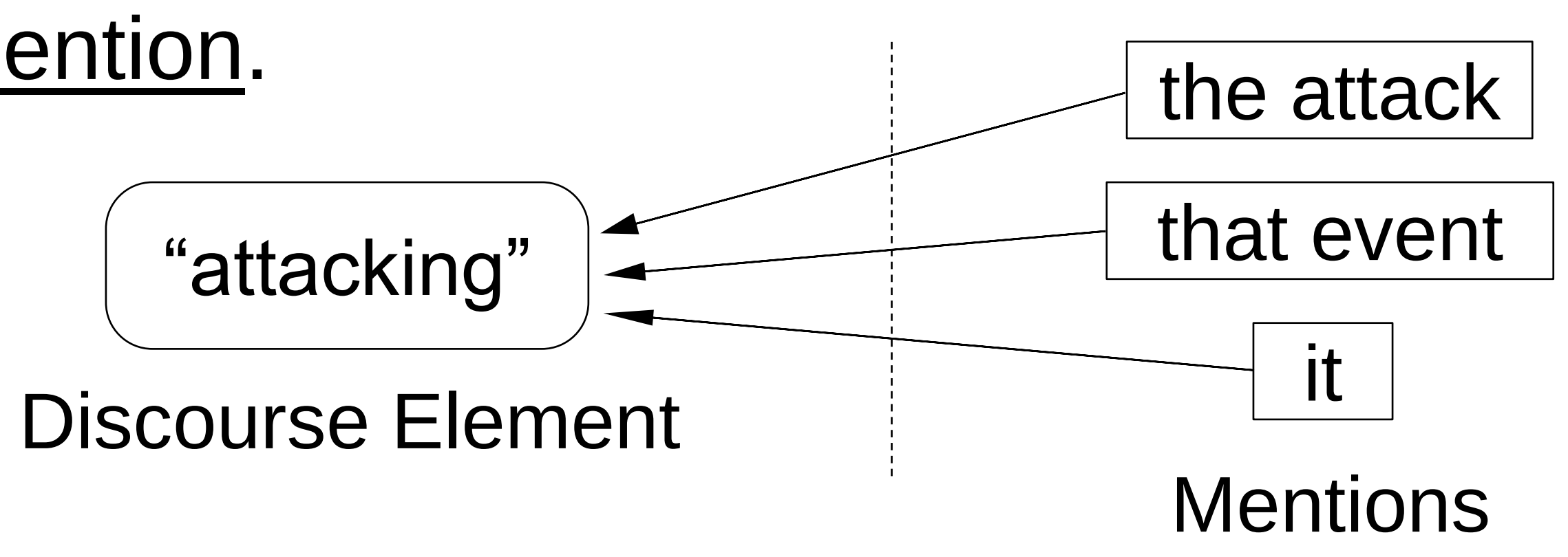
While Turkish troops have been fighting_E.1 a Kurdish faction in northern Iraq, two other Kurdish groups have been battling_E.2 each other.

A radio station operated_E.3 by the Kurdistan Democratic Party said_E.4 the party's forces attacked_E.5 positions of the Patriotic Union of Kurdistan on Monday in the Kurdish region's capital Irbil.
…
The fighting_E.10 was also reported_E.11 by a senior Patriotic Union official, Kusret Rasul Ali, who said_E.12 PUK forces repelled_E.13 a large KDP attack_E.14.
…
Ali claimed_E.16 that 300 KDP fighters were killed_E.17 or wounded_E.18 and only 11 Patriotic Union members died_E.19.

Legend: <mention>_E.n Domain event
<mention>_E.n Reporting event

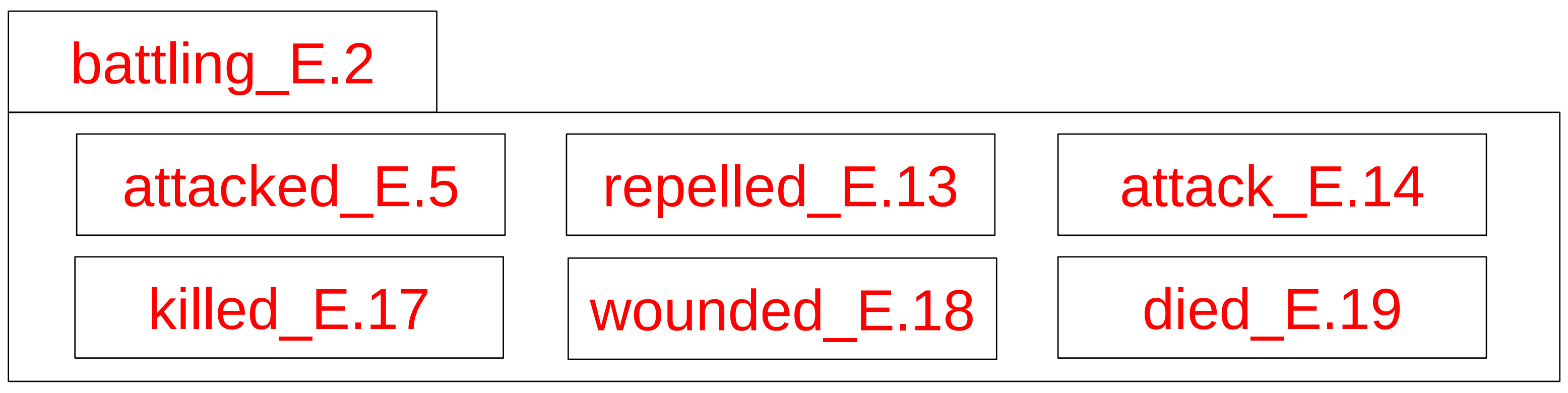
Premise: Event Representation

An event exists in text as a Discourse Element (DE), which is an abstract representation of the event, being referred to by a mention.



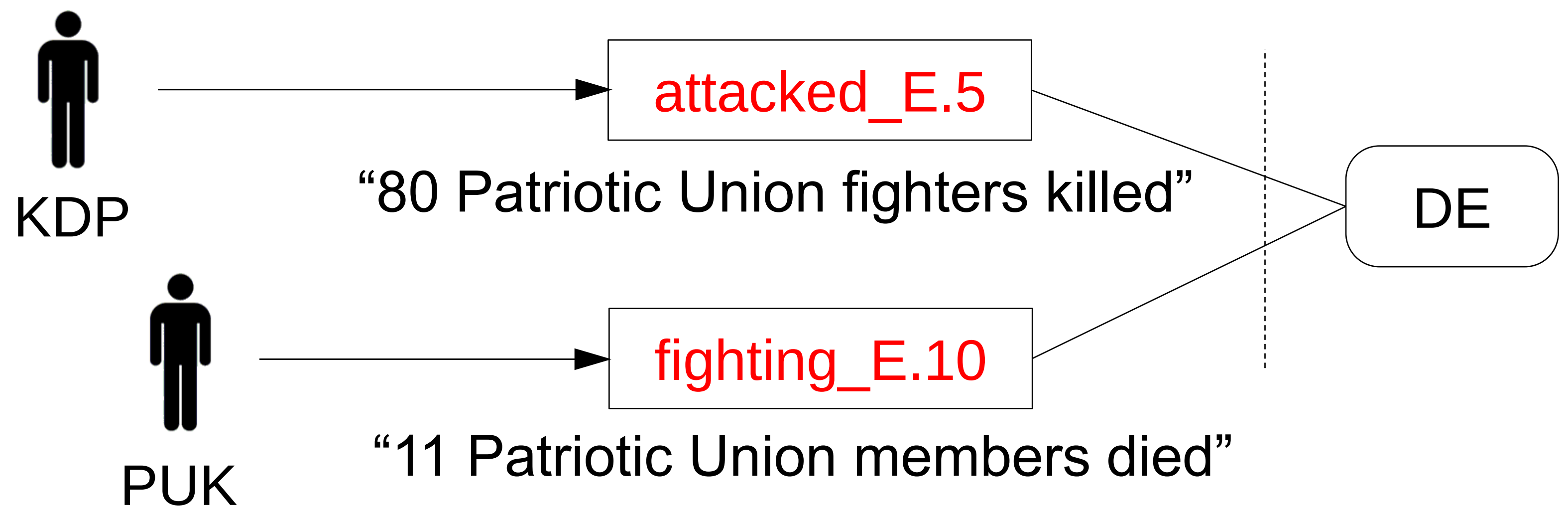
Problem 1: Partial Event Overlap

Some events corefer partially (not fully). For instance, E.2 refers to a series of skirmishes between KDP and PUK. Event E.5 is one of things that occurred in the battle, as shown below. In this case, E.2 and E.5 cannot fully corefer.



Problem 2: Inconsistent Reporting

Some events are reported by different speakers. For example, E.5 and E.10, being reported by KDP and PUK, respectively, refer to the same event. However, KDP and PUK reported different number of deaths, so it is not possible to figure out the coreference between E.5 and E.10 without considering such different perspectives.


KDP
attacked_E.5
"80 Patriotic Union fighters killed"
DE
PUK
fighting_E.10
"11 Patriotic Union members died"

Remaining Problems

Unclear Semantics of Events

Sometimes it is difficult to determine the relationship between events since their semantics is unclear. For instance, E.45 could be a member of E.44, but the decision is hard.

Amnesty International has accused both sides of violating_E.44 international humanitarian law by targeting_E.45 civilian areas, and ...

Approach

Full and Partial Identity

We define full and partial identity of two mentions in the table below. This definition gives a solution to Problem 1.

Identity Type		Full Identity	Partial Identity	
			Member	Subevent
Key Idea		Complete match	A set of multiple instances	Script (a stereotypical sequence)
Semantic Components	Agent	Identical	Identical or Not Identical	Identical
	Patient	Identical	Identical or Not Identical	Identical
	Location	Identical	Identical or Not Identical	Identical (more or less)
	Time	Identical	Identical or Not Identical	Identical (more or less)
	Lexical semantics	Identical	Identical	Not Identical
Example		The bombing (E1) happened early on the morning of July 15. It (E2) killed 3 people. → E1 and E2 corefer.	There were five explosions (E1) last night. The first one (E2) was at a local police station. The second one (E3) was at an airport. → E2 and E3 are members of E1.	The attack (E1) lasted all night. First the soldiers intruded (E2) the houses, and then set up (E3) some bombs there. → E2 and E3 are subevents of E1.

Domain and Reporting Events

We additionally annotate communication events, which we call Reportings. The link from a DE to a reporting event allows us to discount apparent contradictory aspects for more accurate decisions, giving a solution to Problem 2.

Epistemic Status

We also annotate epistemic statuses of an event: (1) actually occurred, (2) negated, (3) expected/desired/future event, and (4) negation of expected/desired/future event.

Annotation

We have been annotating the following two corpora:

Corpus	Typical events	Findings
The **Intelligence Community (IC) Corpus**	Bombing, killing, wars, etc.	This domain offers a manageable set of events (consisting of approximately 50 terms) with no more than three layers.
The **Biography (Bio) Corpus**	Born, dead, married, etc.	Temporal sequencing is more important than scriptal granularity.

The table below shows statistics and inter-annotator agreement for 65 articles in the IC domain corpus. For annotation, we used a modified version of AncoraPipe entity coreference annotation interface (Bertran et al., 2010).

Coreference Relations	Avg no per article	Agreement (Fleiss's kappa)
Full	19.5	0.620
Member	2.7	0.213
Subevent	7.2	0.467

(The avg no of domain and reporting events per article is 41.2)