Evaluation for Partial Event Coreference

Jun Araki, Eduard Hovy, and Teruko Mitamura

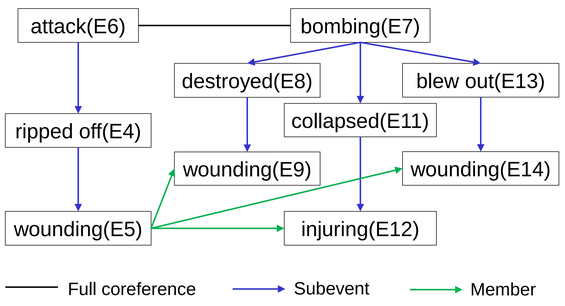
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Introduction

Hovy et al. (2013) introduced partial event coreference:

- **Subevent** relations represent parts of a sequence of stereotypical events, or a script.
- **Membership** relations represent instances of an event collection.

> A car bomb that police said was set by Shining Path guerrillas **ripped off**(E4) the front of a Lima police station before dawn Thursday, **wounding**(E5) 25 people. The **attack**(E6) marked the return to the spotlight of the feared Maoist group, recently overshadowed by a smaller rival band of rebels. The pre-dawn **bombing**(E7) **destroyed**(E8) part of the police station and a municipal office in Lima's industrial suburb of Ate-Vitarte, **wounding**(E9) 8 police officers, one seriously, Interior Minister Cesar Saucedo told reporters. The bomb **collapsed**(E11) the roof of a neighboring hospital, **injuring**(E12) 15, and **blew out**(E13) windows and doors in a public market, **wounding**(E14) two guards.



Partial coreference can form **hierarchical event structure**.

Problem

How should we evaluate the performance of partial coreference detection?

- There are no existing evaluation metrics for the task.
- Existing metrics for full coreference are not readily applicable for partial coreference.
  - It is unclear how to define a cluster for cluster-based metrics such as B-CUBED (Bagga and Baldwin, 1998) and CEAF (Luo, 2005).
  - It is unclear how link-based metrics such as MUC (Vilain et al., 1995) and BLANC (Recasens and Hovy, 2011) penalize incorrect directions of links.

Sub-problems

- What **metric** is suitable to what **evaluation scheme** under what **assumptions**?
- Are there any **existing algorithms or tools** applicable?

Evaluation Scheme

We make **three assumptions**.

A1. **Twinless mentions**

- Twinless mentions (Stoyanov et al., 2009) are the mentions that exist in the gold standard but not in a system response, or vice versa.
- Assumption: a metric for partial coreference should be able to handle twinless mentions.
  - We simultaneously show the performance of mention detection using precision, recall, and F1.

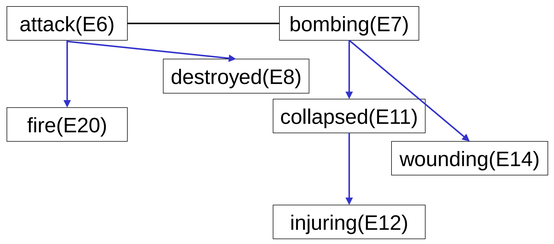
A2. **Intransitivity**

- Assumption: partial coreference is not transitive.
  - E.g., subevent link E7 →^{s} E14 is not a correct link as compared to the gold standard.

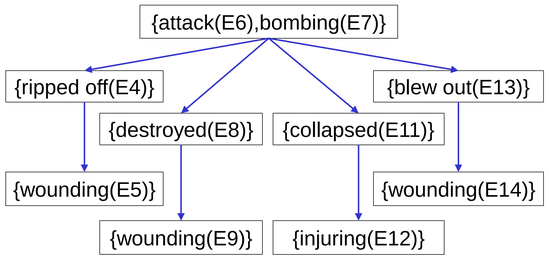
A3. **Link propagation**

- Assumption: partial coreference links can be propagated in combination with full coreference links.
  - E.g., subevent link E6 →^{s} E8 is a correct link from a combination of full coreference link E6 == E7 and subevent link E7→ E8 in the gold standard.

A system response (ignoring member relations):



A **conceptual event hierarchy** simplifies the evaluation for partial coreference.



Same score
- System A: E6 == E7 and E6 →^{s} E8
- System B: E6 == E7 and E7 →^{s} E8
- System C: E6 == E7, E7 →^{s} E8, and E7 →^{s} E8

Evaluation Metrics

Representation of partial coreference:

Partial coreference in a document is represented by a forest of unordered trees.

- F_g: the gold standard forest
- F_r: a system response forest

Five desired properties for a metric $sim(F_g, F_r)$:

P1. **Identity**: $sim(F_1, F_1) = 1$

P2. **Symmetricity**: $sim(F_1, F_2) = sim(F_2, F_1)$

P3. **Zero**: $sim(F_1, F_2) = 0$ if F_1 and F_2 are totally different.

P4. **Monotonicity**: $sim(F_g, F_r)$ should increase from 0 to 1 monotonically as F_r, which is totally different from F_g, approach F_g.

P5. **Linearity**: $sim(F_g, F_r)$ should increase linearly as each single individual correct piece of information is added to F_r.

We examine **three metrics**.

1. Extension to **MUC**: **MUC_p**

- Given a set of gold standard entities K and a set of response entities R, MUC is defined as:

$$Precision_{MUC} = \frac{\#\text{ common links between entities in } K \text{ and } R}{\#\text{ links in R}}$$

$$Recall_{MUC} = \frac{\#\text{ common links between entities in } K \text{ and } R}{\#\text{ links in K}}$$

- In MUC_p, a correct link is one matched with the gold standard including its direction.

2. Extension to **BLANC**: **$BLANC_p$**

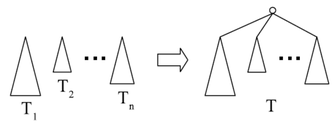
- BLANC averages F1 scores for positive and negative links.

$$F_{BLANC} = \frac{F_p + F_n}{2} = \frac{P_p R_p}{P_p + R_p} + \frac{P_n R_n}{P_n + R_n}$$

- For $BLANC_p$, we also change the definition of a correct link in the same way as MUC_p.

3. Extension to a normalized version of **Simple Tree Matching (STM)** (Yang, 1991): **$NSTM_p$**

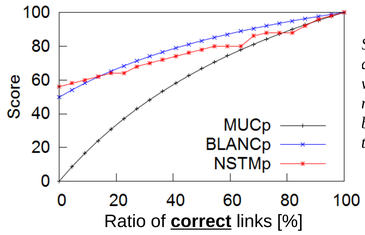
- STM uses dynamic programming to measure the maximum node-matches in a top-down fashion.
- We use greedy search instead, and merge a forest into a single tree.



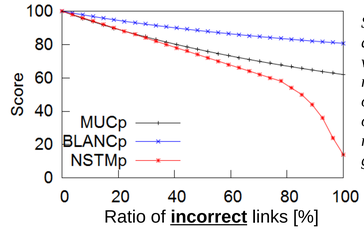
Results

- MUC_p and $BLANC_p$ are better than $NSTM_p$.

Property	MUC_p	$BLANC_p$	STM_p
Identity	Good	Good	Good
Symmetricity	Good	Good	Good
Zero	Good	Not good	Not good
Monotonicity	Good	Good	Not good
Linearity	Acceptable	Acceptable	Not good



Scores of MUC_p, $BLANC_p$, and $NSTM_p$ in the case where we increase the number of correct links one by one from all singletons to the perfect output.



Scores of MUC_p, $BLANC_p$, and $NSTM_p$ in the case where we increase the number of incorrect links one by one from the perfect output to a single tree merging all trees in the gold standard.

Conclusion

Summary of this work

- A conceptual event hierarchy simplifies the evaluation of partial event coreference.
- We extended MUC, BLANC, and STM for the evaluation of partial coreference.
  - The extended metrics are generic enough to be used in other evaluations involving data structures based on unordered trees.
- The extended MUC and BLANC are better than the extended STM for evaluating partial coreference.

Future work

- Incorporating structural consistency as an additional property
  - E.g., System D can be better than system E because system D finds a subevent sibling relation between {E8} and {E11}.

System D: {E6,E7} →^{s} {E8} and {E6,E7} →^{s} {E11}
System E: {E8} →^{s} {E9} and {E11} →^{s} {E12}