Incident Threading for News Passages

Gökçe Ayduğan – Onur Aydın

Content

- Introduction
- Incident Threading
- Previous Work
- Passage Threading
- Algorithms
- Evaluation
- Experiments & Results
- Conclusion

1. Introduction

1.1. Problem Definition

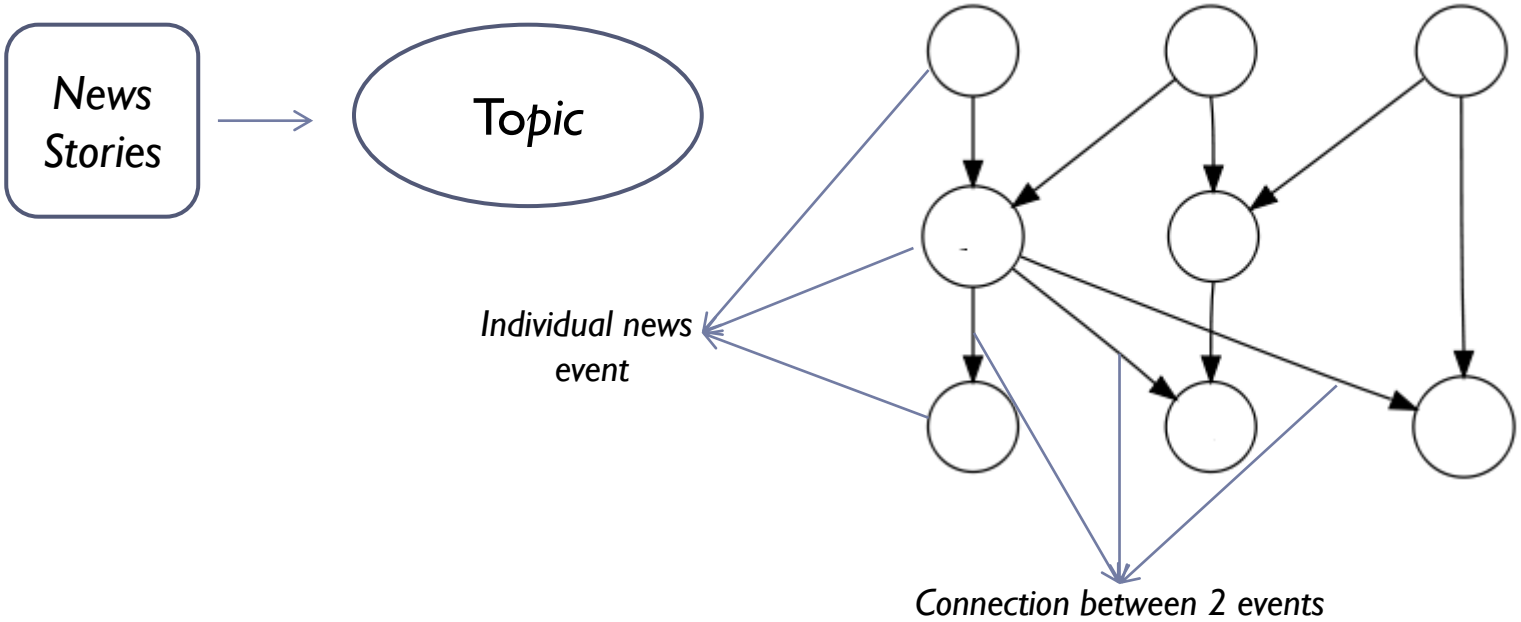
- Information Acquisition Task Tools
  - Search Engines
  - Question Answering Systems
  - Online Forums
  - Mail Lists
  - Serving Interesting News?

1.2. Motivation

- Grouped news according to the topic discussed
- Non-duplicate information
- Similar actions to link related event

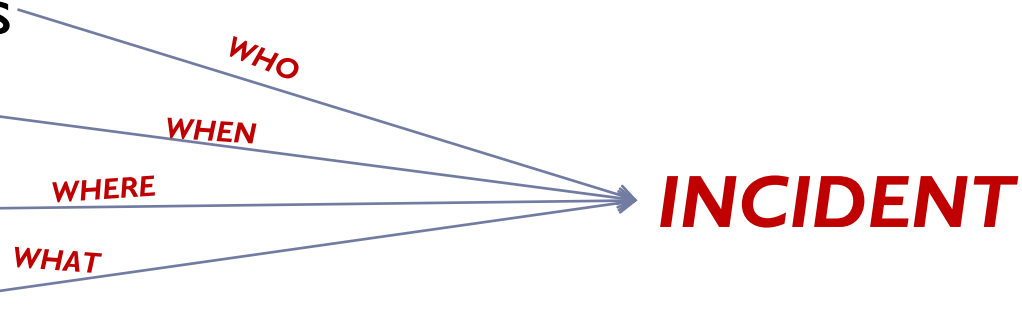
2.1. Incident Threading

- Topic Detection and Tracking(TDT)

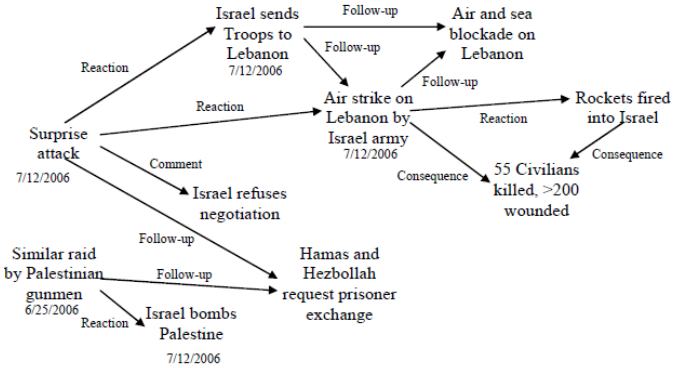


2.1.1 Incident

- Main Characters
- Time Stamps
- Location
- Action



2.1.2 Incident Network



an incident network that represents *some news reports about an Israel-Lebanon conflict form CNN archives*

Connection Types in an Incident Network

- Logical Relation

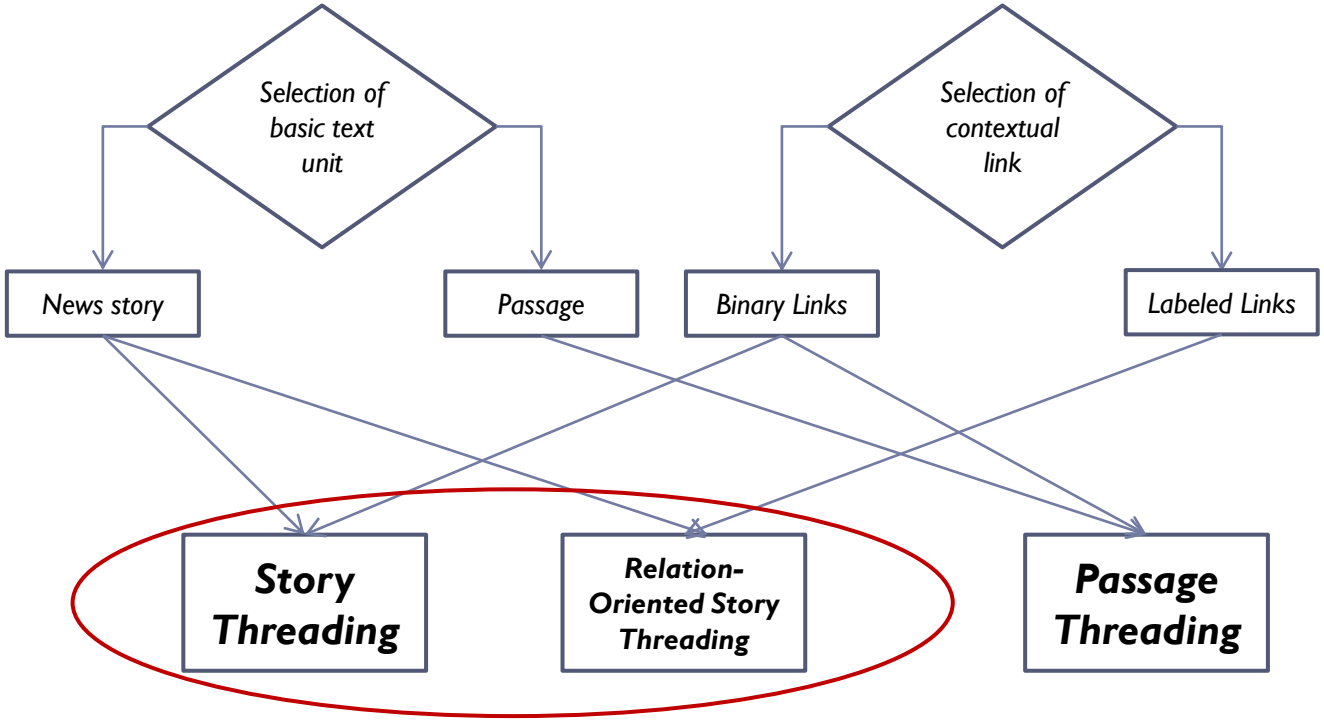
  Prediction, *Comment*, *Reaction*, *Analysis*, *Background*, and *Consequence*.

- Progression

  follow-up

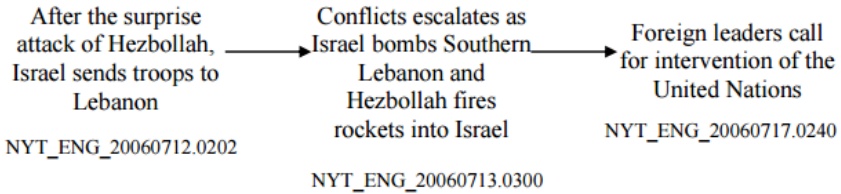
- Weak Relations

3. Previous Work

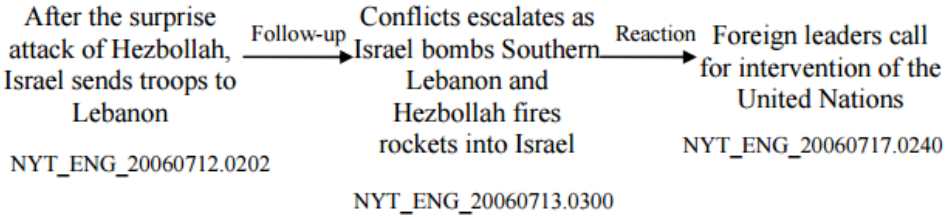


3. Previous Work(Cont.)

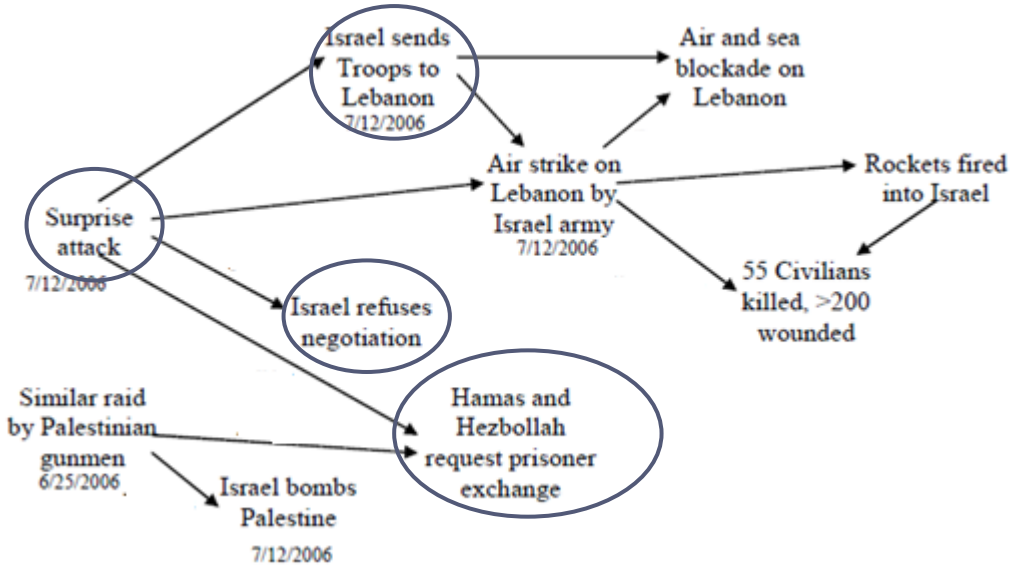
- Story Threading



- Relation-Oriented Story Threading



4. Passage Threading

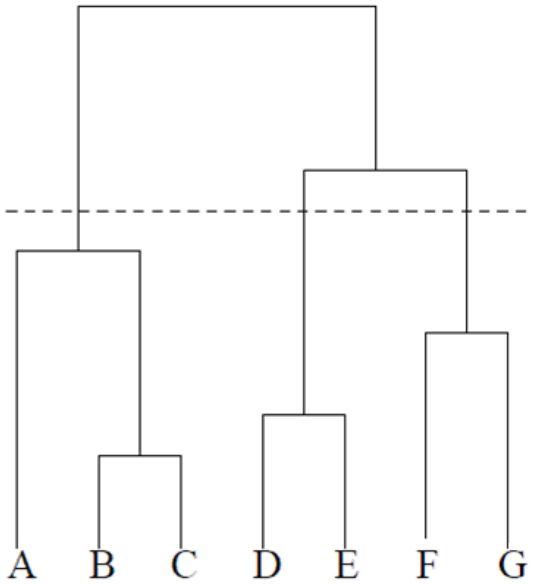


A passage is a continuous subset of a news story that contains a complete description of certain news information.

5. Algorithms

- Baseline Algorithm
  - Agglomerative Clustering
  - Linking Incidents

- Three-Stage Algorithm
  - Binary Classification (e.g.Violent)
  - Agglomerative Clustering
  - Linking Incidents

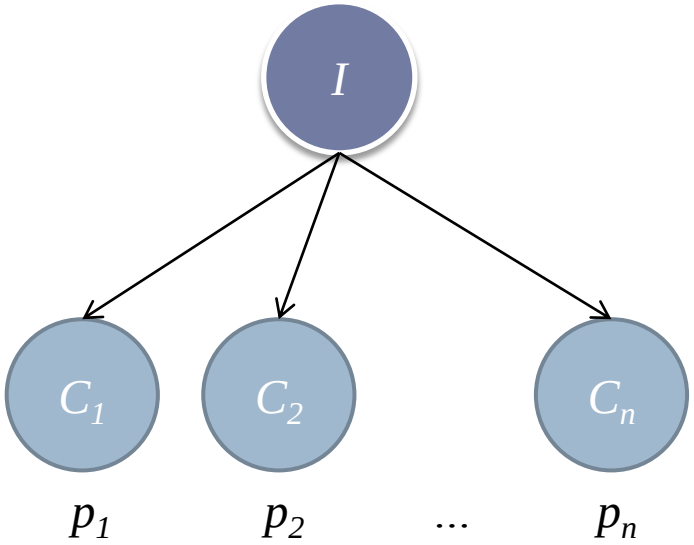


6. Evaluation

- Clustering Evaluation
  - Concentration (Distribution of clusters in an Incident)

$$Conc(I) = \frac{\sum_{i=1}^{n} p_i(p_i - 1)}{p(p-1)}$$

$$Concentration = \frac{\sum_i Conc(I_i)|I_i|}{\sum_i |I_i|}$$

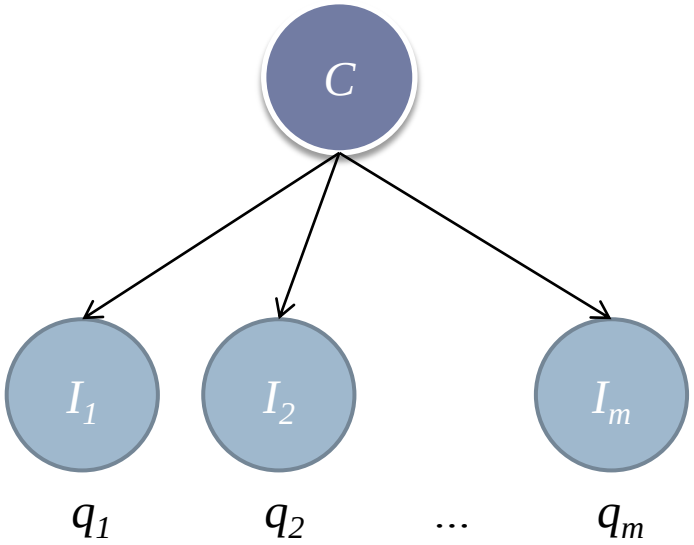


6. Evaluation

- Clustering Evaluation
  - Purity (Distribution of incidents in a Cluster)

$$Pur(C) = \frac{\sum_{i=1}^{m} q_i(q_i - 1)}{q(q-1)}$$

$$Purity = \frac{\sum_i Pur(C_i)|C_i|}{\sum_i |C_i|}$$



6. Evaluation

MT: Ground Truth
MS: System Output

- Linking Evaluation

$$P_{link} = \frac{\sum_{i,j} | MS_{ij} \times MT_{ij} |}{\sum_{i,j} MS_{ij} \times MS_{ij}}$$ = Precision

$$M_{ij} = \begin{cases} 1 & p_i \rightarrow p_j \\ -1 & p_j \rightarrow p_i \\ 0 & otherwise \end{cases}$$

$$R_{link} = \frac{\sum_{i,j} | MS_{ij} \times MT_{ij} |}{\sum_{i,j} MT_{ij} \times MT_{ij}}$$ = Recall

$$Err_{link} = (1 - \frac{\sum_{i,j} MS_{ij} \times MT_{ij}}{\sum_{i,j} | MS_{ij} \times MT_{ij} |}) / 2$$

6. Evaluation

- Combine Measurements

$$Mean_{cluster} = \frac{2 \times concentration \times purity}{concentration + purity}$$

$$Mean_{link} = \frac{2 \times P_{link} \times R_{link}}{P_{link} + R_{link}} (1 - Err_{link})$$

$$Mean_{all} = \frac{2 \times Mean_{cluster} \times Mean_{link}}{Mean_{cluster} + Mean_{link}}$$

6. Evaluation

- Complex Link Evaluation

$$SQ\_Sim(DT, DS) = \sqrt{\frac{\sum_{i,j} f(DT_{ij}, DS_{ij})}{\sum_{i,j} \max(f(DT_{ij}, DT_{ij}), f(DS_{ij}, DS_{ij}))}}$$

$$D_{ij} = \begin{cases} 0 & p_i \approx p_j \\ 1 & p_i \rightarrow p_j \\ -1 & p_i \leftarrow p_j \\ \infty & p_i \circ p_j \end{cases}$$

7. Experiments & Results

Table 5: Performance Comparison for Passage-Based Systems – $Mean_{all}$ Optimized

Binary Classification

Evaluation	Baseline	Three-stage	Change in %
Incident concentration	0.1985	0.2609	+31.4%
Cluster agreement	0.1494	0.2703	+80.8%*
Clustering precision	0.1427	0.2830	+98.3%*
Clustering recall	0.1445	0.2161	+49.4%*
Link precision	0.0345	0.1598	+362.5%*
Link recall	0.1574	0.1866	+18.5%
Link direction error	0.3995	0.4295	+7.4%
$Mean_{all}$	0.0361	0.0654	+80.1%*
SQ_SIM(DT,DS)	19.10%	26.40%	+38.2%*

7. Experiments & Results

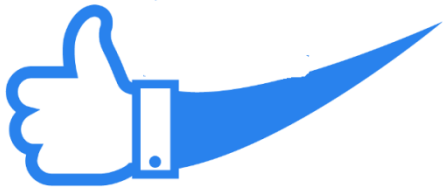
Table 6: Performance Comparison for Passage-Based Systems – *SQ_SIM(DT,DS)* Optimized

Evaluation	Baseline	Three-stage	Change in %
Incident concentration	0.3099	0.3864	+24.6%
Cluster agreement	0.1073	0.1855	+72.9%*
Clustering precision	0.1146	0.1807	+57.6%*
Clustering recall	0.2691	0.3472	+29.0%
Link precision	0.0380	0.0350	-7.8%
Link recall	0.0226	0.0113	-49.8%
Link direction error	0.2166	0.2678	+23.6%
$Mean_{all}$	0.0133	0.0110	-17.8%
SQ_SIM(DT,DS)	22.58%	25.05%	+10.9%

Complexity

8. Conclusion

- Three-Stage Algorithm is significantly better than Baseline Algorithm
- The application of incident threading is justifiable in a real system.
- This work has made contributions on both theoretical and technical aspects.



9. References

- Feng, Ao, and James Allan. "Incident threading for news passages."*Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.
- Feng, Ao, and James Allan. "Finding and linking incidents in news."*Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007.