

# **Preemptive Information Extraction**

**using Unrestricted Relation Discovery**

---

**Yusuke Shinyama  
Satoshi Sekine**

**New York University**



# Quick Recipe for IE

- Specify the information you want.
  - e.g. “hiring and firing event”

*Manual*

- Create a system to extract the information.

*Semi-automatic*

- Get the results.

<b>Person</b>	<b>Position</b>	<b>Company</b>
Smith	President	PQR
Jones	Vice President	XYZ

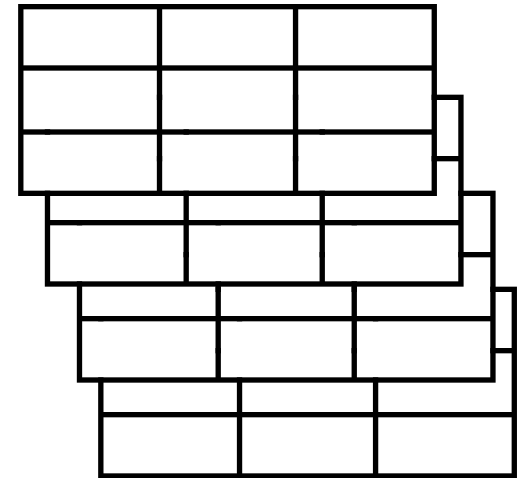
*Automatic*



# IE Gone Wild

---

- We want to discover all the IE tasks that are potentially feasible.
- Create an IE system for every scenario *preemptively*.
- Specifying the information  
⇒ Searching a relevant IE system (table).



# Prior Work

---

- Pattern acquisition for predefined relation:
  - (Riloff, 96; Brin, 98; Yangarber et al, 00; Agichtein et al, 00)
- Unsupervised relation discovery:
  - (Hasegawa, 04; Chen et al, 05; Zhang et al, 05)



# Preemptive IE - How?

---

- Overall procedure:
  1. Find a pair (or triple) of entities that have a similar relation in articles.
    - ex. Relation between “**Katrina**”-“**New Orleans**” in **Article A** is similar to relation between “**Longwang**”-“**Taiwan**” in **Article B**.
  2. Cluster relations based on their similarity.
  3. A cluster of similar relations = table.

<b>Article A</b>	Katrina	New Orleans
<b>Article B</b>	Longwang	Taiwan



# Find Similar Relations

---

- Consider a simple case.
  - **Named Entities (NE)** are tagged.

## Article A

**Katrina** headed for **New Orleans**.

Common expression

## Article B

**Longwang** headed for **Taiwan**.



# Find Similar Relations

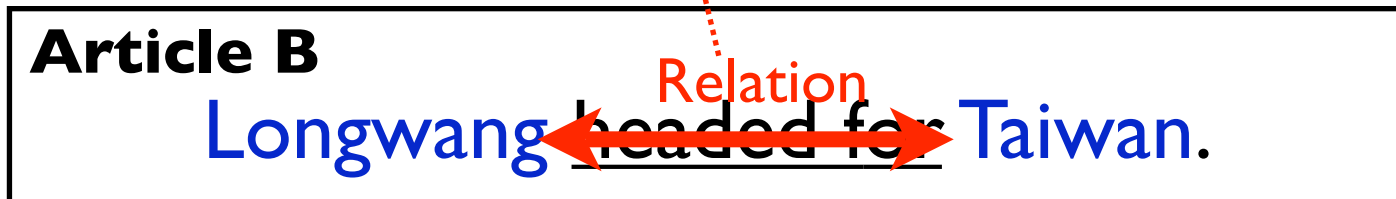
---

- Consider a simple case.
  - Named Entities (NE) are tagged.

**Article A**  
Katrina headed for New Orleans.

A diagram showing the sentence "Katrina headed for New Orleans." with "Katrina" and "New Orleans" in blue. A red double-headed arrow labeled "Relation" is drawn over the underlined phrase "headed for".

**Article B**  
Longwang headed for Taiwan.

A diagram showing the sentence "Longwang headed for Taiwan." with "Longwang" and "Taiwan" in blue. A red double-headed arrow labeled "Relation" is drawn over the underlined phrase "headed for". A dotted red line connects the "Relation" label in Article A to the "Relation" label in Article B.

Common expression between NEs  
= They have similar relation.



# Find Similar Relations

---

- Actual expressions are varied, but...

## Article A

Katrina headed for ... New Orleans was hit ...  
Katrina threatened ... New Orleans residents ...  
Katrina is category 5 ... evacuated New Orleans

## Article B

Longwang hit ... Taiwan 's coast ...  
Longwang headed for ... Taiwan was pounded...  
Longwang is swirling ... Taiwan was hit...



# Find Similar Relations

---

- There might be common expressions.

## Article A

**Katrina** headed for ...    **New Orleans** was hit ...

Katrina threatened ...    New Orleans residents ...

Katrina is category 5 ...    evacuated New Orleans

Common expression  
(Disjoint, but close)

## Article B

Longwang hit ...    Taiwan 's coast ...

**Longwang** headed for ...    Taiwan was pounded...

Longwang is swirling ...    **Taiwan** was hit...



# Find Similar Relations

- Parallel correspondence of expressions = similar relations.

## Article A

**Katrina** headed for ... **New Orleans** was hit ...

Katrina threatened ... New Orleans residents ...

Katrina is category 5 ... evacuated New Orleans

## Article B

Longwang hit ... **Longwang** headed for ... **Taiwan** 's coast ...

Longwang is swirling ... **Taiwan** was hit...

Relation

Relation



# Find Similar Relations

---

- Unrestricted Relation Discovery:
  1. Collect all the expressions (*patterns*) that modify each NE.
  2. For every pair of articles:
    - Try to find a parallel correspondence of patterns between multiple NEs.
    - Exist? → These NEs must have similar relations.
    - Cluster similar relations.
- We have a bunch of similar relations:
  - Don't know what they are, but show them as a table.



# Multiplying Patterns

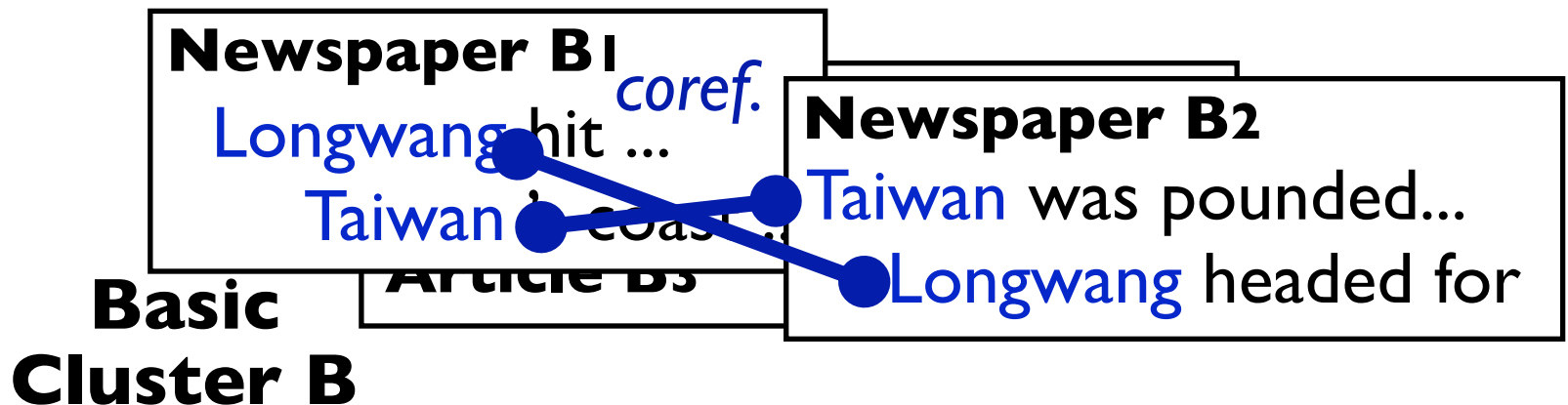
---

- Reality:
  - The number of patterns is not enough.
- Solution:
  1. Use a cluster of articles reporting the same event instead of a single article. (*basic cluster*)
  2. Cluster basic clusters that have similar relations. (*metacluster*)
    - Metaclusters = IE tables.



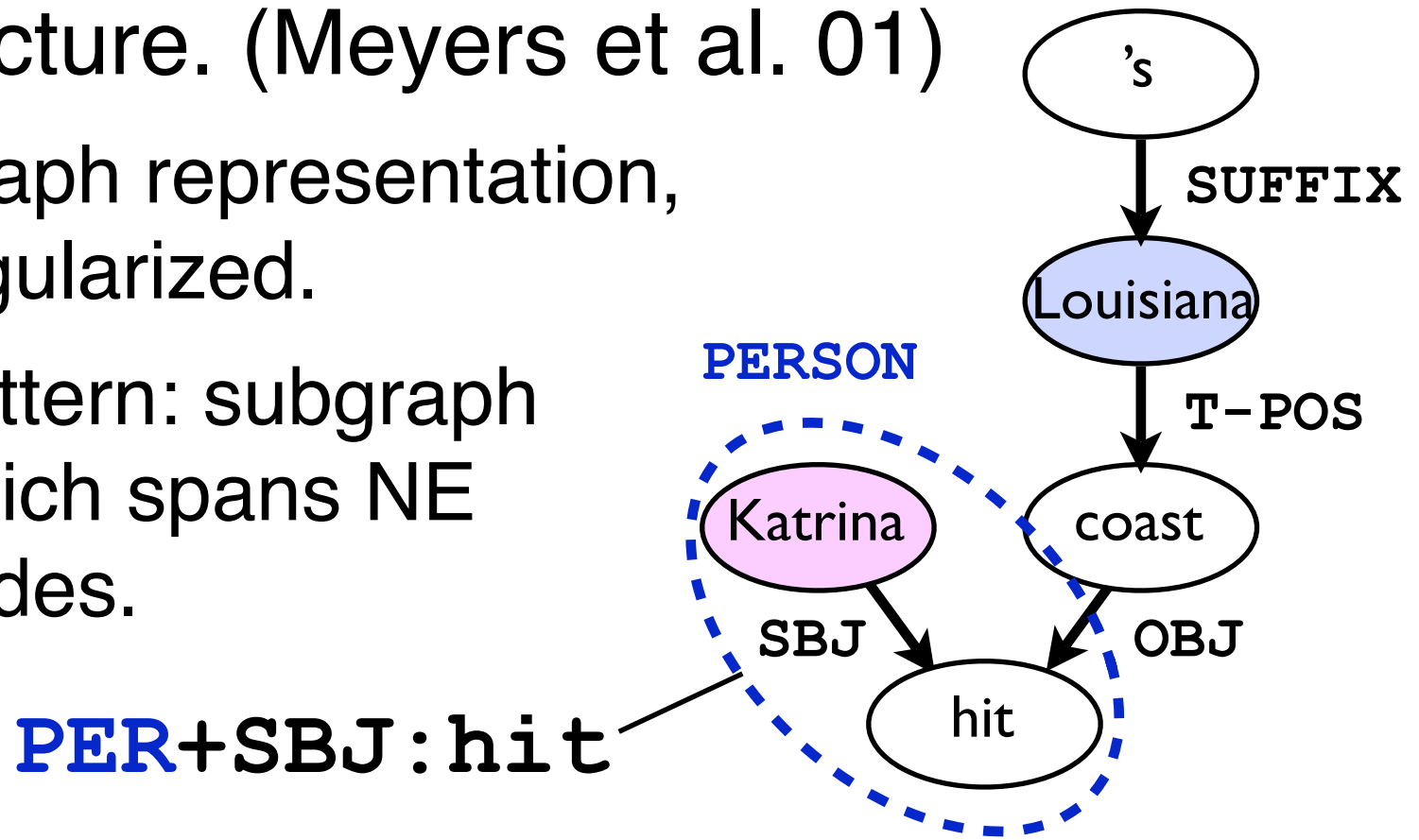
# Multiplying Patterns

- Use multiple news sources and coreference resolution.



# Pattern Format

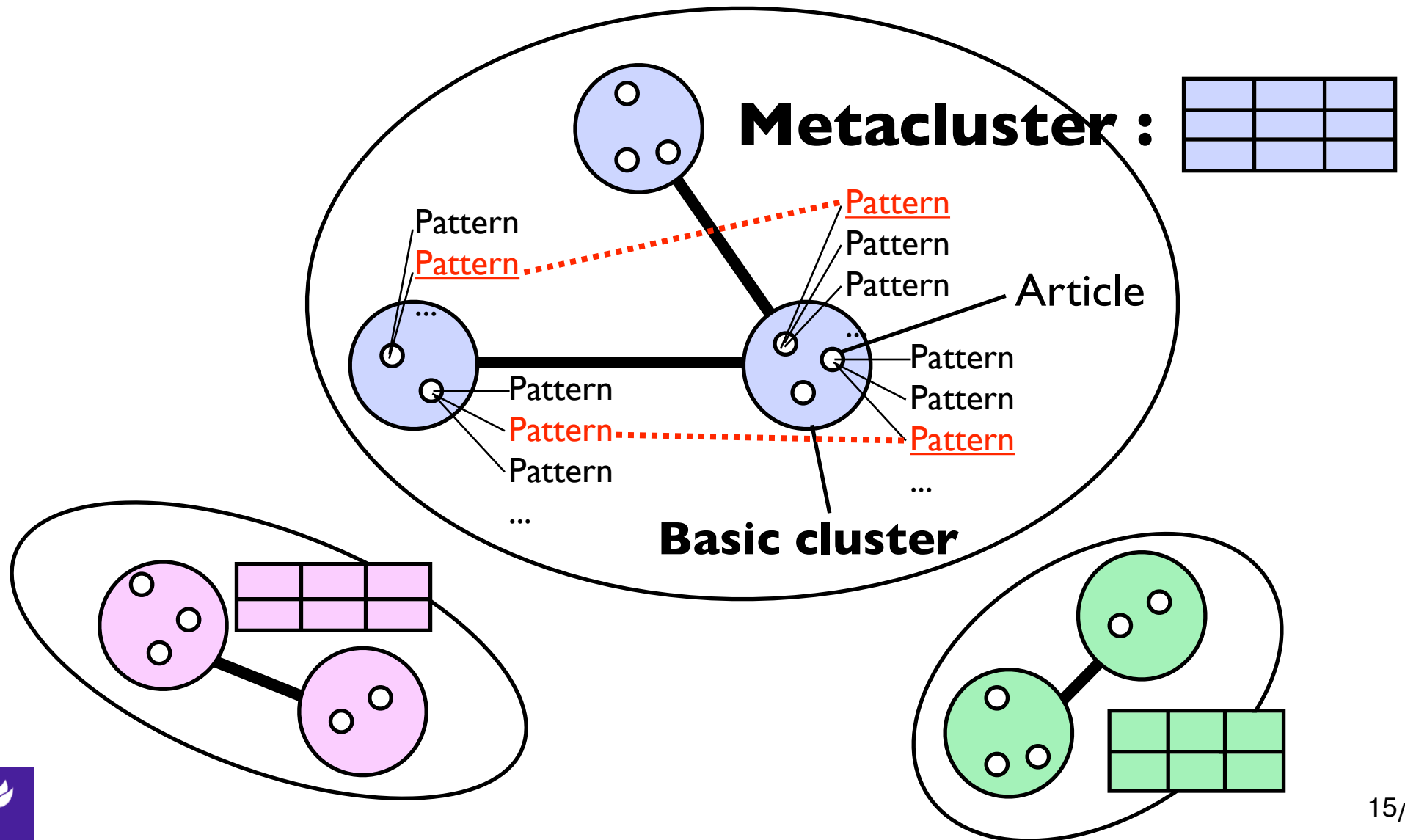
- GLARF: predicate argument structure. (Meyers et al. 01)
  - Graph representation, regularized.
  - Pattern: subgraph which spans NE nodes.



*"Katrina hit Louisiana's coast."*



# Metaclustering



# Implementation

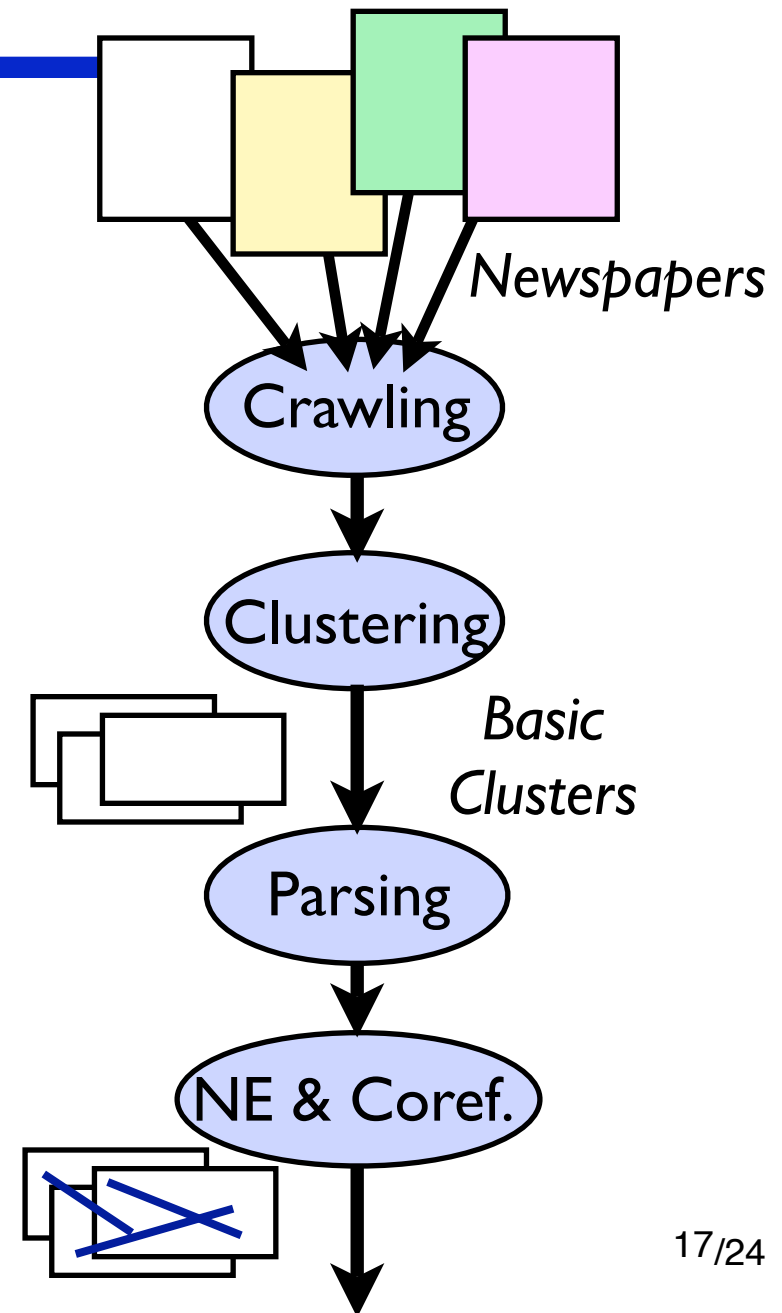
---

- Obtain news articles from multiple news sites on the Web.
- Basic Cluster: Vector space model.
- NE tagging & Coref. Resolution: Use an in-house NE tagger and coreference resolver.
- Generate GLARF structures; weight obtained patterns with ICF.



# Implementation

- System Overview
  1. Crawling the Web
  2. Clustering articles (*basic clusters*)
  3. Parsing & NE tagging
  4. Coreference Resolution
  - ...



# Implementation

- System Overview (*cont.*)

5. GLARF generation  
(*patterns*)

6. Metaclustering

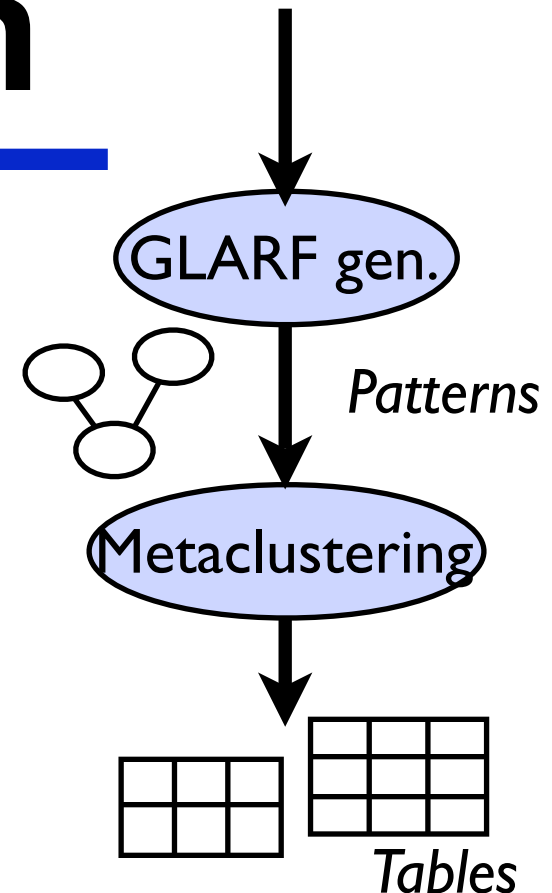
- Use patterns as features of a basic cluster.

Katrina hit Louisiana's coast.

**PER+SBJ:hit** (**PER=Katrina**)

Longwang hit Taiwan.

**PER+SBJ:hit** (**PER=Longwang**)



# Experiments

---

Source articles (2 months)	28,009
Basic clusters ( $\geq 3$ articles)	5,543
Patterns - token	643,767
Patterns - type	7,990
Metaclusters	302
Metaclusters ( $\geq 3$ events)	101

We used 12 news sites in U.S.



# Evaluation

---

- We evaluated 48 tables manually.
  - Try to find a sentence that explains at least 2/3 of the rows in the table ("a consistent table").
  - For consistent tables, count how many rows fit the explanation.

Consistent	36 (75%)
Inconsistent	12
Total	48

Tables

Rows fitted	118 (73%)
Rows not fitted	43
Total	161

Rows



# Evaluation

---

<b>Table Description</b>	<b>Rows</b>
Storm <i>PERSON</i> probably affected <i>GPE</i> .	8/16
Nominee <i>PERSON</i> must be confirmed by <i>ORG</i> .	4/7
<i>PERSON</i> urges <i>GPE</i> to make changes.	4/6
<i>GPE</i> launched an attack in <i>GPE</i> .	3/5
<i>PERSON</i> ran against <i>PERSON</i> in an election.	4/5
<i>PERSON</i> visited <i>GPE</i> on a diplomatic mission.	2/4
<i>PERSON</i> beat <i>PERSON</i> in golf.	4/4
<i>GPE</i> soldier(s) were killed in <i>GPE</i> .	3/3
<i>PERSON</i> ran for governor of <i>GPE</i> .	2/3
Boxer <i>PERSON</i> fought boxer <i>PERSON</i> .	3/3



# Error Analysis

---

- Incorrect rows (out of 10)
  - 4 rows: coreference resolution error (referring to wrong NEs.)
  - 4 rows: patterns are distant to each other.
    - “ Hamas ’ most senior leader, Mahmoud al-Zahar, announced ...  Sharon  faced a vote in his party ...”
  - 1 row: parse error.
  - 1 row: obscure.



# Conclusion

---

- We proposed Preemptive Information Extraction.
  - Find parallel correspondence of patterns between multiple entities.
  - Use clustering to build tables.
- Obtained a reasonable result.
  - Possible to prepare many tables offline: Elections, Sports results, M&As, ...



# Future Work

---

- Questions:
  - How NE types and pattern format could limit types of news articles?
  - How much does each stage affect the overall performance?
- Improvements:
  - Use various NE types:
    - DATE, CURRENCY, NUMBER, ...
  - Decent evaluation (ACE event corpus).

