

Being Lazy and Preemptive at Learning toward Information Extraction

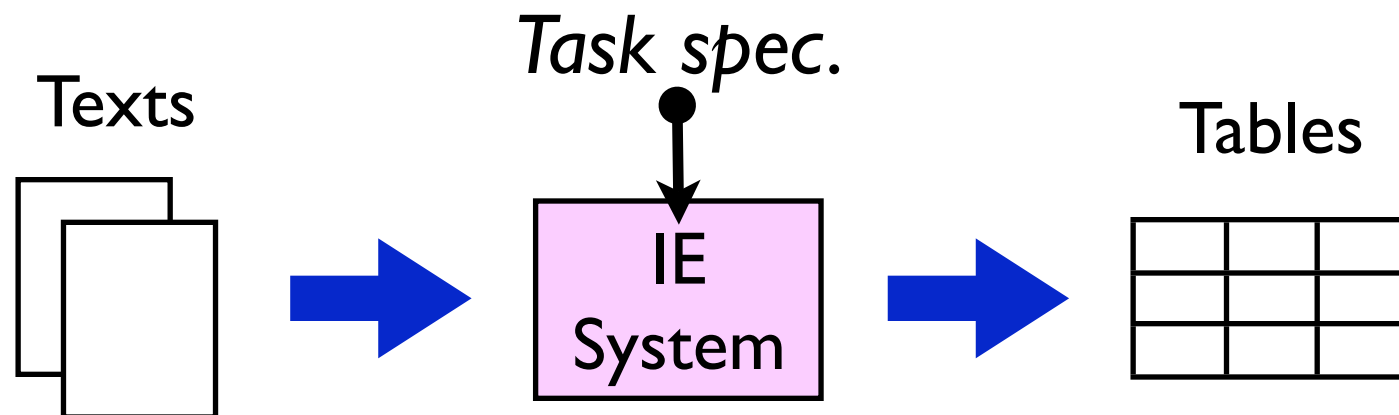
**Yusuke Shinyama
New York University
Thesis Defense
Aug. 2, 2007**



Introduction and Problem Description

What is IE? (not)

- To convert unstructured information into structured information.
 - Input: Natural language texts
 - Output: Tables (relations)
 - Parameter: Task specification (*scenario*)



What is IE?

- Task: *Murder in news articles.*
 - Input: **"Alice killed Bob."**
"John killed Fred."
 - IE system (simplistic):

([A-Za-z]+) killed ([A-Za-z]+)

- Output:

<i>Murderer (\$1)</i>	<i>Victim (\$2)</i>
Alice	Bob
John	Fred



What is IE?

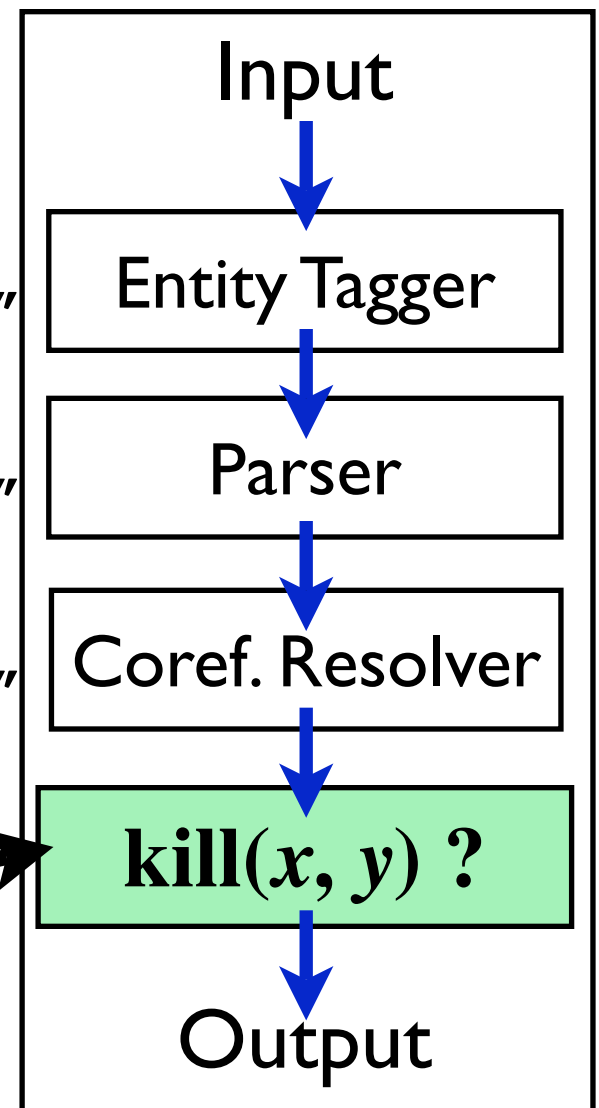
- Reduce linguistic variations.

"Alice killed John C. Smith, Jr."

"Alice allegedly killed Bob."

"Alice killed him."

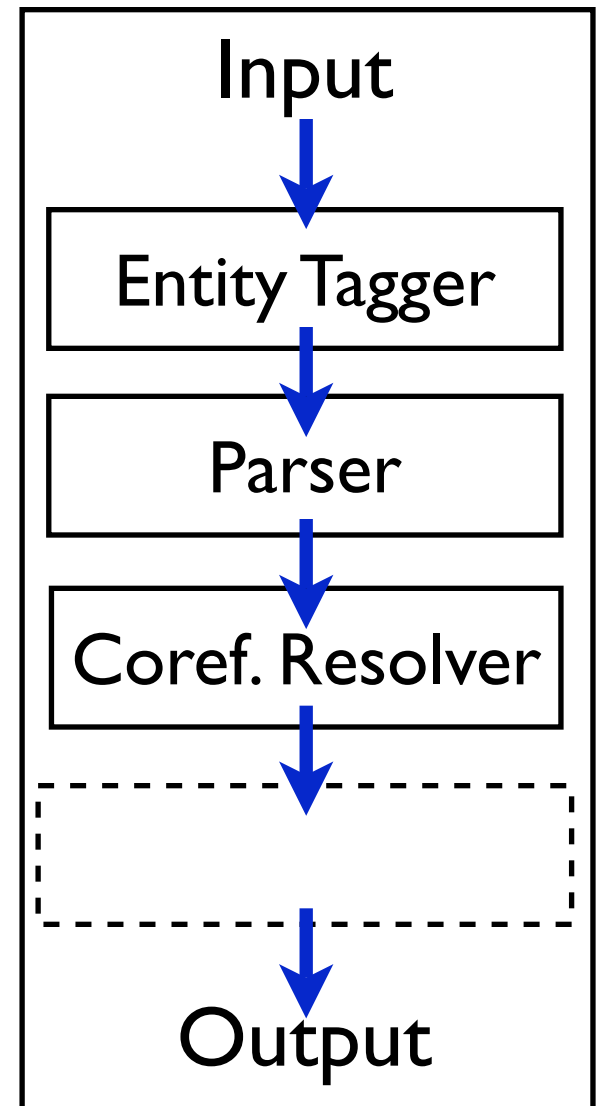
Pattern matching



Challenges in IE

- How to reduce the cost of building a system?

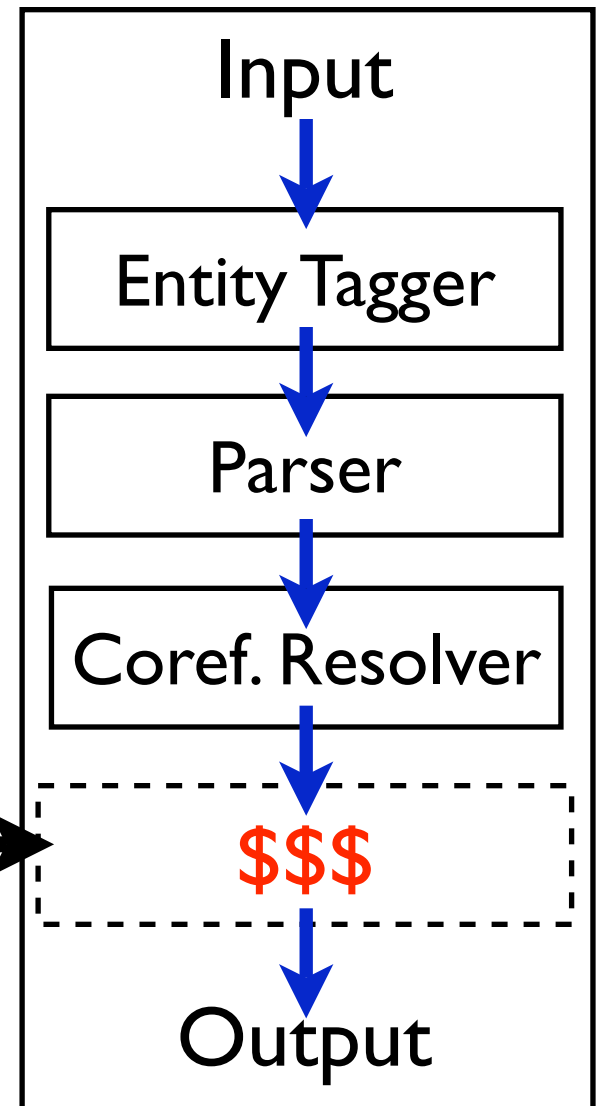
Reusable for every task.



Challenges in IE

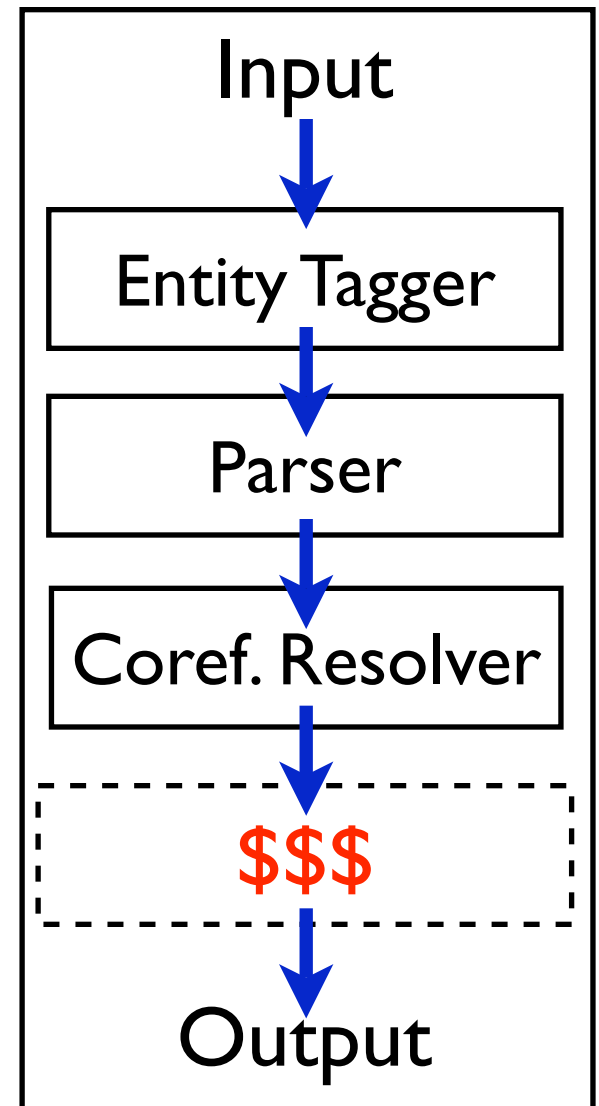
- The “pattern” part still remains expensive.
 - Task dependent.
 - Varied expressions:
 - “x killed y”
 - “x shot y to death”
 - ...

Need to create for every task.



Challenges in IE

- How to get the patterns?
 - Craft manually (expensive).
 - Learn from hand-tagged corpus (expensive).
- Attempts to reduce cost:
 - [Riloff, 96] [Brin, 98]
[Agichtein, 00] [Sudo, 03]



Challenges in IE

- Every time a user changes the task, a certain part of the system needs to be tuned or rebuilt.
- Some tasks are appropriate for IE, and some are not.
 - What kind of tasks are appropriate for IE in the first place?



Challenges in IE

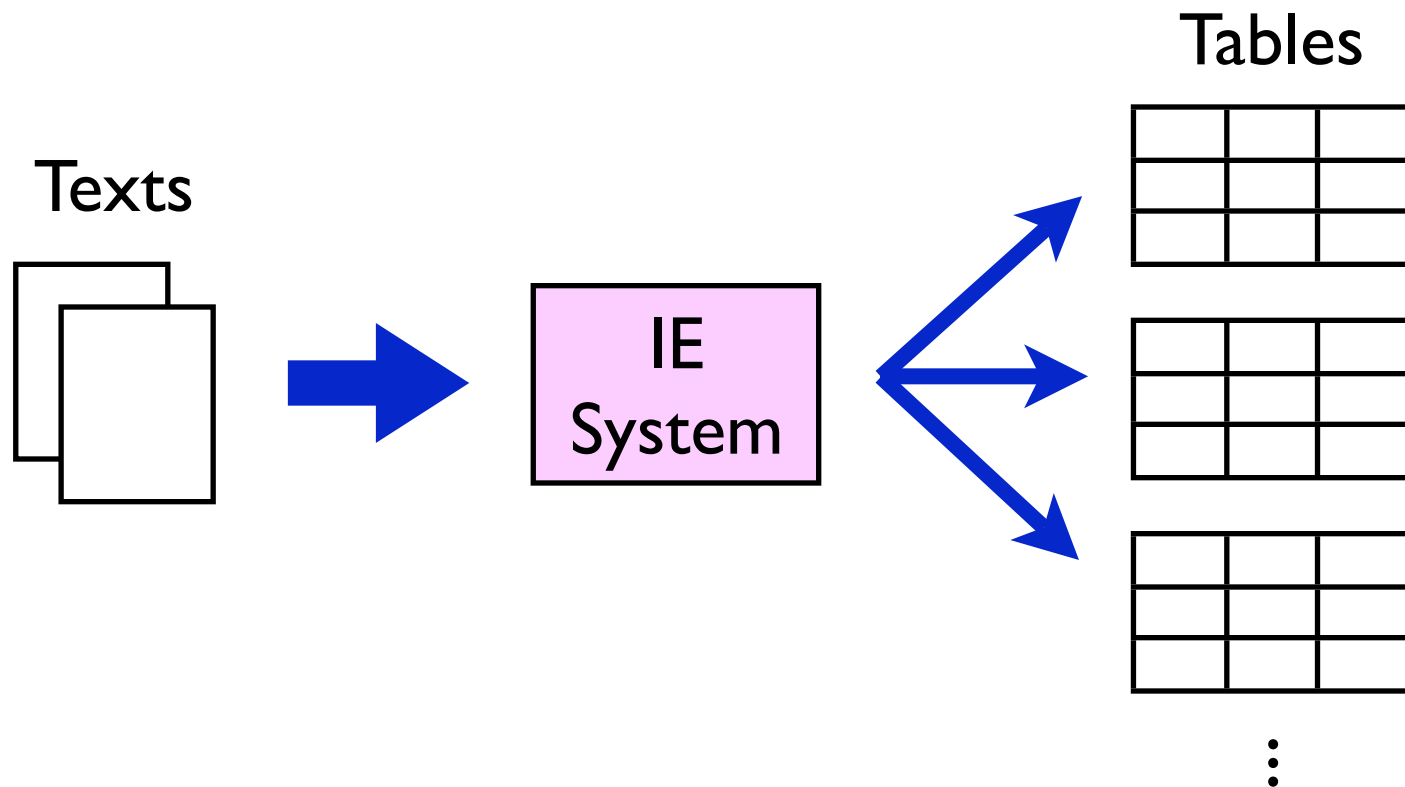
- Some tasks are not appropriate:
 - Events that need description:
 - President's agenda on foreign policy.
 - Book review.
 - Infrequent events:
 - "Her body was found in the apartment of a man authorities said planned to eat the corpse."
 - Cannibalism or just a murder?
 - Depends on how to generalize the event.

Assume we agree on this.



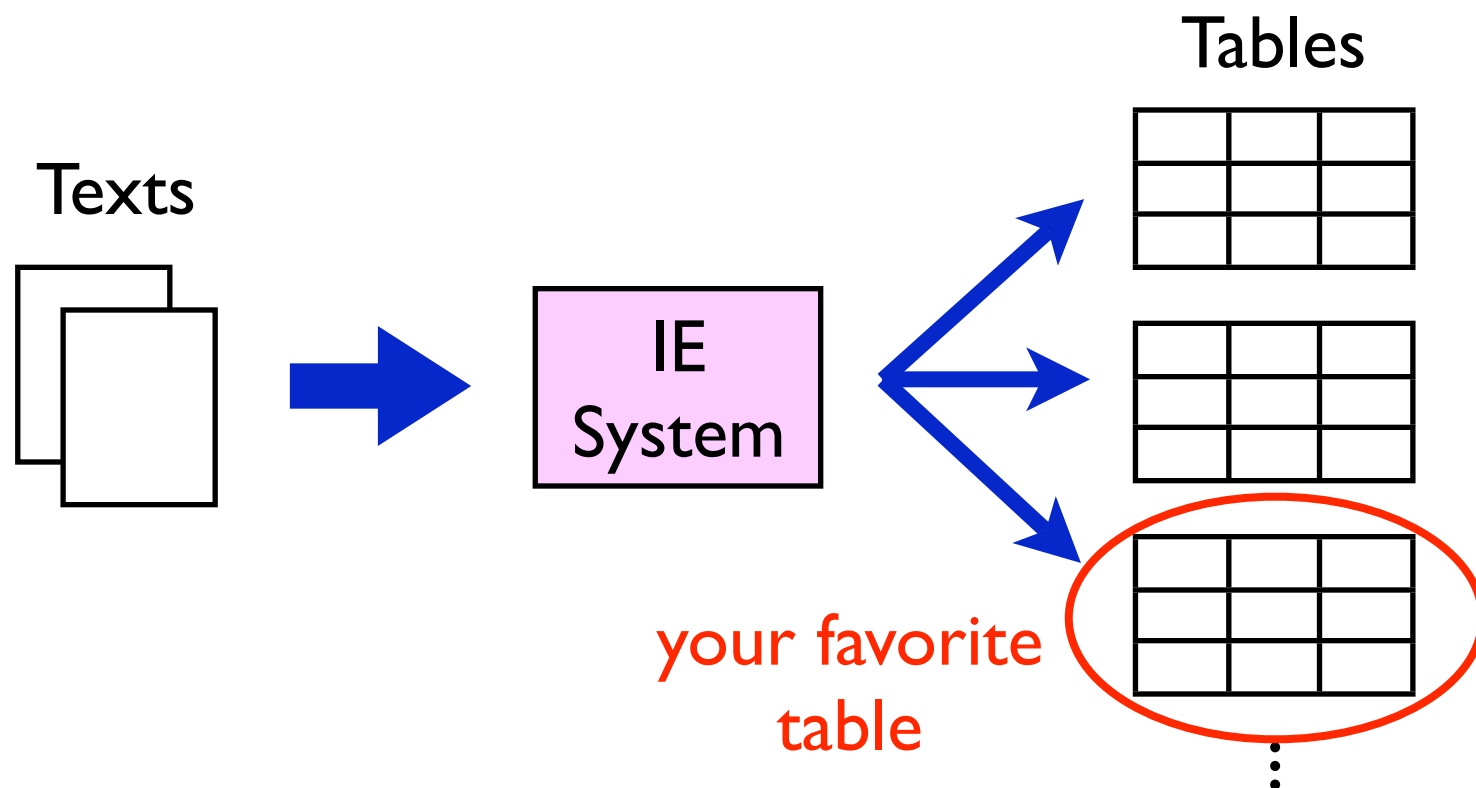
Our Question

- Is it possible to find *all* the tasks that are feasible in IE?



Our Question

- “Preemptive IE”
 - IE is now turned into a search problem.



**Idea
and
Methodology**

Preemptive IE?

- Traditional IE:
 1. Obtain the patterns.
 2. Extract a piece of information.
- Preemptive IE:
 1. Extract every piece of information.
 2. Figure out the relations among the extracted data.

How to do that?



Problem

- We can't use pattern matching to extract a piece of information:

"Alice killed Bob."

"John killed Fred."

If we use patterns...

`([A-Za-z]+) killed ([A-Za-z]+)`

`(Alice, Bob)`

`(John, Fred)`

Limit the type
of relations.



Solution

- Instead, we extract **every** tuple of entities that *may* have some relation:

"Alice killed Bob."

"Alice met David in NY."

"Fred called John."

"John killed Fred."



(Alice, Bob)

(Alice, David, NY)

(Fred, John)

(John, Fred)



Solution

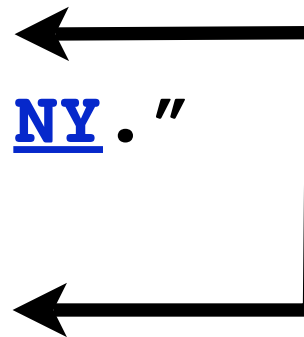
- Then, find out the tuples with the same relation based on their expressions.

"Alice killed Bob."

"Alice met David in NY."

"Fred called John."

"John killed Fred."



Must have the same relation.



Solution

- Use a **clustering technique** to group similar events into a cluster.

"Alice killed Bob."

"Alice met David in NY."

"Fred called John."

"John killed Fred."

Group into
one cluster



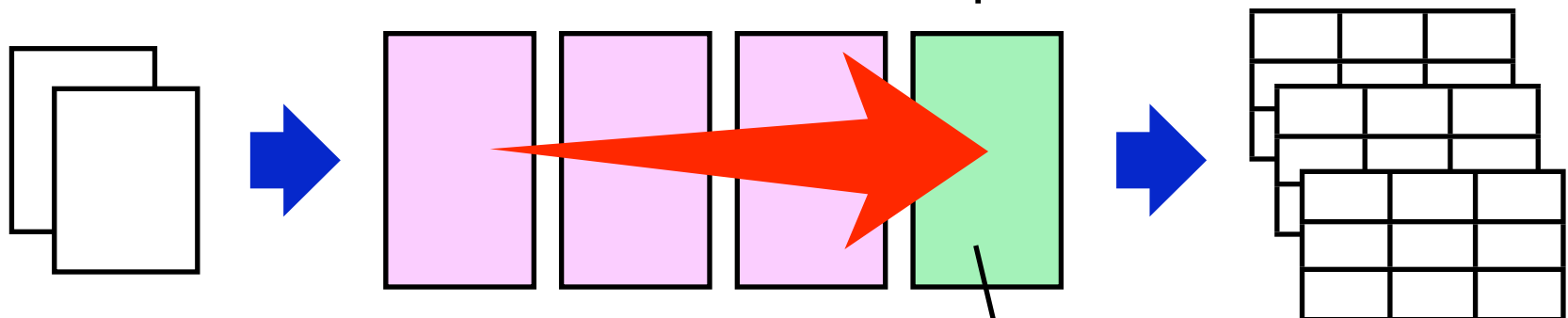
Alice	Bob
John	Fred



Inspired by...

- End-to-End principle [Saltzer, 84]
 - System performance should be measured and tuned at the end layers.
 - Intermediate layers should not introduce arbitrary bias.

1. Generate redundant, non task-specific data.



2. Delay the decision until the end.

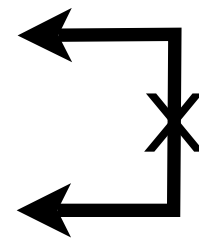


Practical Problems

1. Number of all the possible entity combinations is exponential.
 - Choose 5 entities out of 10.
 - 9,864,091 tuples!
2. What if different expressions are used?

"Alice killed Bob."

"John murdered Fred."



Solution

- Use “comparable articles”:
 - Articles that report the same event.

Newspaper A (Aug. 2):

“Alice killed Bob Sunday...”

Newspaper B (Aug. 2):

“Bob was fatally wounded...”

Newspaper C (Aug. 2):

“Alice stabbed Bob with ...”



Solution

- Treat a set of comparable articles as a single event.
- Weight frequent entities:

"Alice killed Bob, not Paul..."

Single
event

"Bob was fatally wounded..."

"Alice stabbed Bob with ..."

(Alice, Bob) > (Alice, Paul)
(Bob, Paul)



Solution

- Find the same expressions among a set of comparable articles.

Aug. 2
(Alice, Bob)

"Alice killed Bob, ..."

"Bob was fatally wounded..."

"Alice stabbed Bob with ..."

Oct. 10
(John, Fred)

"John murdered Fred..."

"John killed Fred..."

"Fred was shot to death..."



Overall Direction

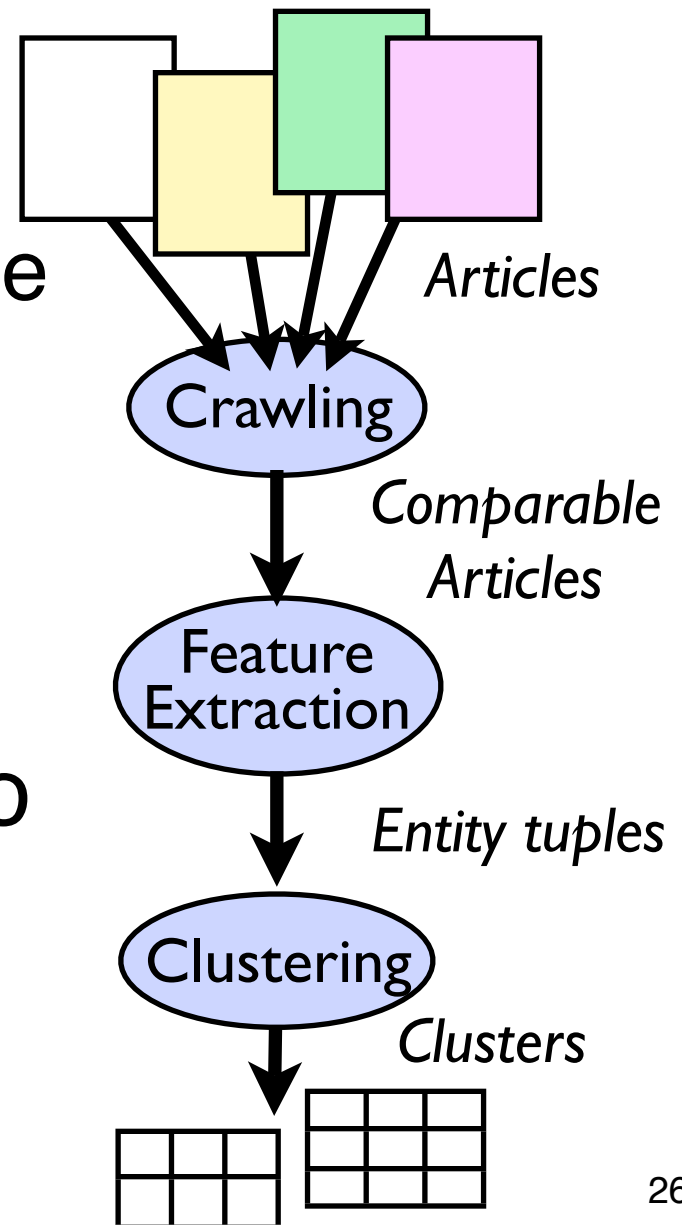
- Input: a lot of comparable articles.
- Output: a lot of tables.
- Procedure:
 - For each set of comparable articles:
 1. Generate possible entity tuples.
 - e.g. (**Alice**, **Bob**)
 2. Extract the expressions (**features**).
 - e.g. “**killed**”, “**murdered**”, ...
 - Group the similar tuples to form clusters.



Implementation and Experiments

System Overview

- Input:
 - News articles from multiple news sites on the Web.
- Output:
 - Clusters (tables)
- Use existing packages to extract various features.



Web Crawling

- How to get the right part of text from various web pages?

The New York Times
APRIL 20, 2005, SUNDAY, MARCH 20, 2005 7:47 AM ET
SEARCH

Congress Ready to Approve Bill in Schiavo Case

By Helen Thomas and Cole Miller
Congressional leaders reached a compromise on legislation to force the case of Terri Schiavo into federal court.

- Election Day, Shutdown, Don't Ask, Don't Tell
- Iraq's Situation, Unsettling, as U.S. Pulls Out
- Iran, The Schiavo Case (Photo-essay)

States and Communities Battling Another Round of Base Closings

By Eric Lipton
Thousands of jobs and billions of dollars in local and state economies are at stake as the Military evaluates 425 domestic bases.

- Japan: The Road Ahead for Military Bases

Trading Knees Base, Rice Sends Forceful Reminder to

Sports
WEST VIRGINIA - Wake Forest - Wake Forest - Wake Forest

West Virginia's Persistence Wears Down Wake Forest

West Virginia's Mike Gandy scored 11 of his 20 points after the two timeouts.

By JOE LAMARCA
Published March 20, 2005

CLEVELAND, March 19 - Lesser teams have won it with greater aplomb than West Virginia to top in the N.C.A.A. basketball tournament, but none has had a more dramatic and satisfying two-game start.

The Mountaineers advanced to the Round of 16 Saturday night by overcoming a 19-point halftime deficit and defeating Wake Forest, 111-100, in double overtime. Mike Gandy led the Mountaineers with 23

NEW YORK
© USA Today
© Sports Illustrated
© World's Largest Outdoor Sports Equipment Store
© Sports Illustrated

Washington
WASHINGTON, March 19 - Congressional leaders reached a compromise Saturday on legislation to force the case of Terri Schiavo into federal court, an extraordinary intervention intended to protect the life of the brain-damaged woman whose condition has triggered a painful national debate over when medical treatment should be withdrawn.

Top lawmakers in both the House and the Senate said they hoped to pass the compromise bill as early as Sunday. They said it would allow Mr. Schiavo's parents to ask a federal judge to restore her feeding tube on the ground that their daughter's constitutional rights were being violated by the withholding of nutrition needed to keep her alive.

The White House announced late Saturday that President Bush, who was vacationing at his ranch in Crawford, Tex., would make an unannounced return on Sunday to Washington, where he would remain until early Monday in anticipation of signing the measure.

NEW YORK
© USA Today
© Sports Illustrated
© World's Largest Outdoor Sports Equipment Store
© Sports Illustrated



Web Crawling

- Correctly identified 90% of the pages in 20 news sites automatically.

New York Times

Washington Post

ABC News

Los Angeles Times

CBS News

NY Daily News

Channel News Asia

Voice of America

Financial Times

NYI

Newsday

Boston Globe

BBC

Reuters

Seattle Times

International Herald Tribune

CNN

Independent

USA Today

1010 Wins



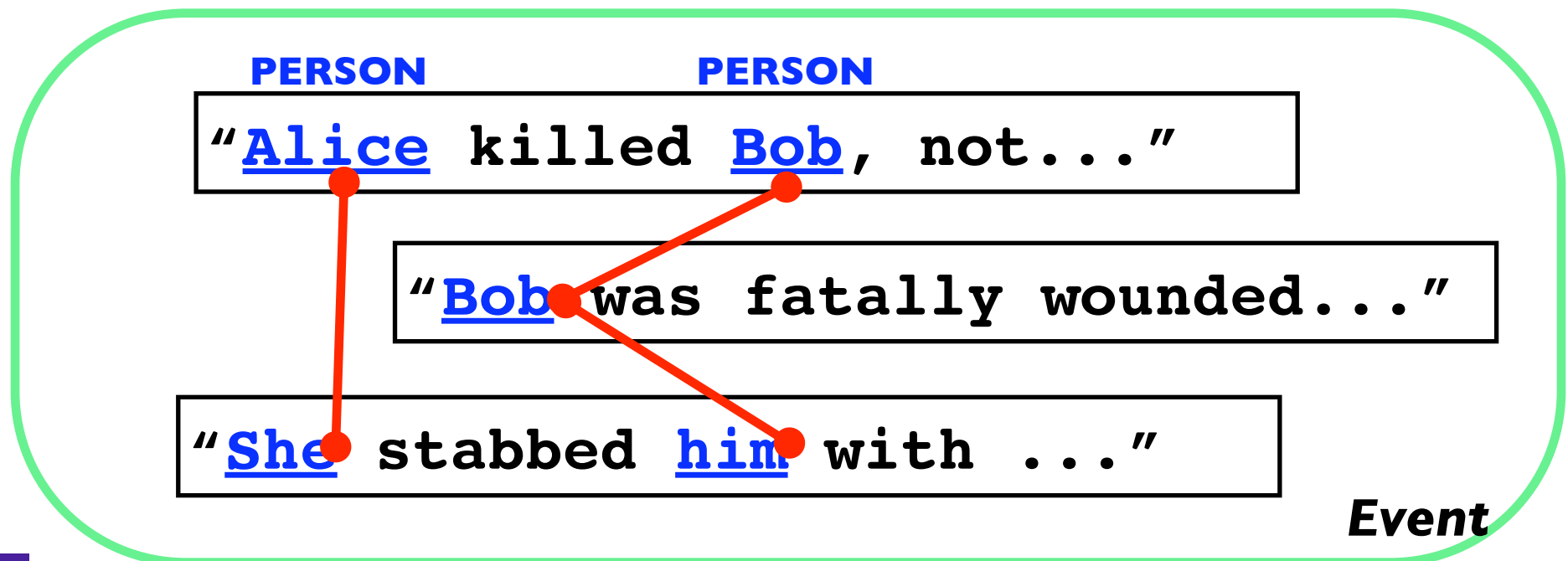
Web Crawling

- Use a clustering method to collect comparable articles.
 - A bag-of-words + Vector Space Model.
- Run the crawler daily:
 - About a year (349 days)
 - 1.1 million articles (3,800 articles/day)
 - 35,398 comparable article sets (size \geq 5)



Entity Recognition

- Use a Named Entity tagger and coreference resolver to recognize the entities within an article set.



Tuple Generation

- Generate possible combinations of entities based on the weight of entity.

"Alice killed Bob, not ..."

"Bob was fatally wounded..."

"She stabbed him with ..."

Event

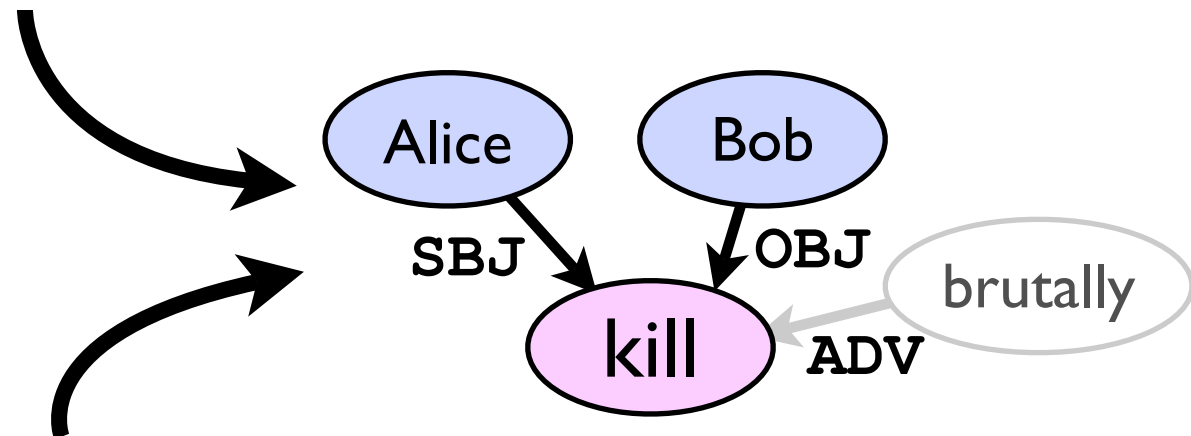
(PER:Alice, PER:Bob)



Feature Extraction

- Local features: GLARF [Meyers, 02]
 - Representation that captures various linguistic phenomena in a uniform way.

“Alice killed Bob brutally.”



“Bob was killed by Alice.”



Feature Extraction

- Global feature:
 - Differentiate topics:
 - “Bloomberg beat Ferrer in NYC.” (Politics)
 - “Sharapova beat Henin in Queens.” (Sports)
 - Bag of words in the article set.

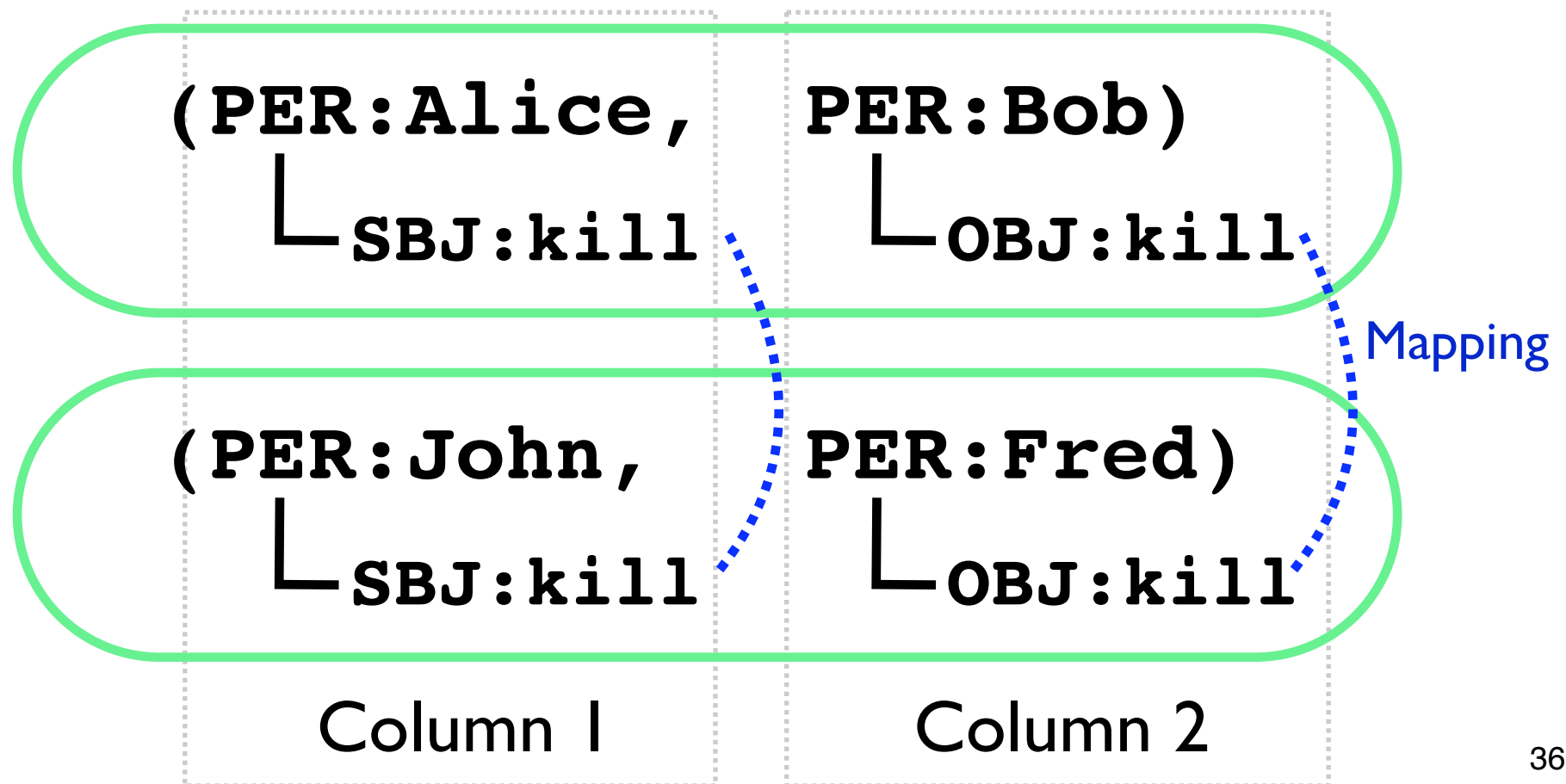
vote campaign party leader
(PER:Bloomberg, PER:Ferrer)

tennis tournament title champion
(PER:Sharapova, PER:Henin)



Clustering

- Find a pair of entity tuples that share some features.



Clustering

- Group a pair of tuples if the similarity between two tuples A and B exceeds a certain threshold.

$$A = (e_{A1}, e_{A2}, \dots)$$

$$B = (e_{B1}, e_{B2}, \dots) \quad \text{Vector of [freq.} \times \text{weight]}$$

$$\mathbf{Sim}(A, B) = \sum \mathbf{f}(\underline{e_{A.local}}, e_{B.local}) +$$

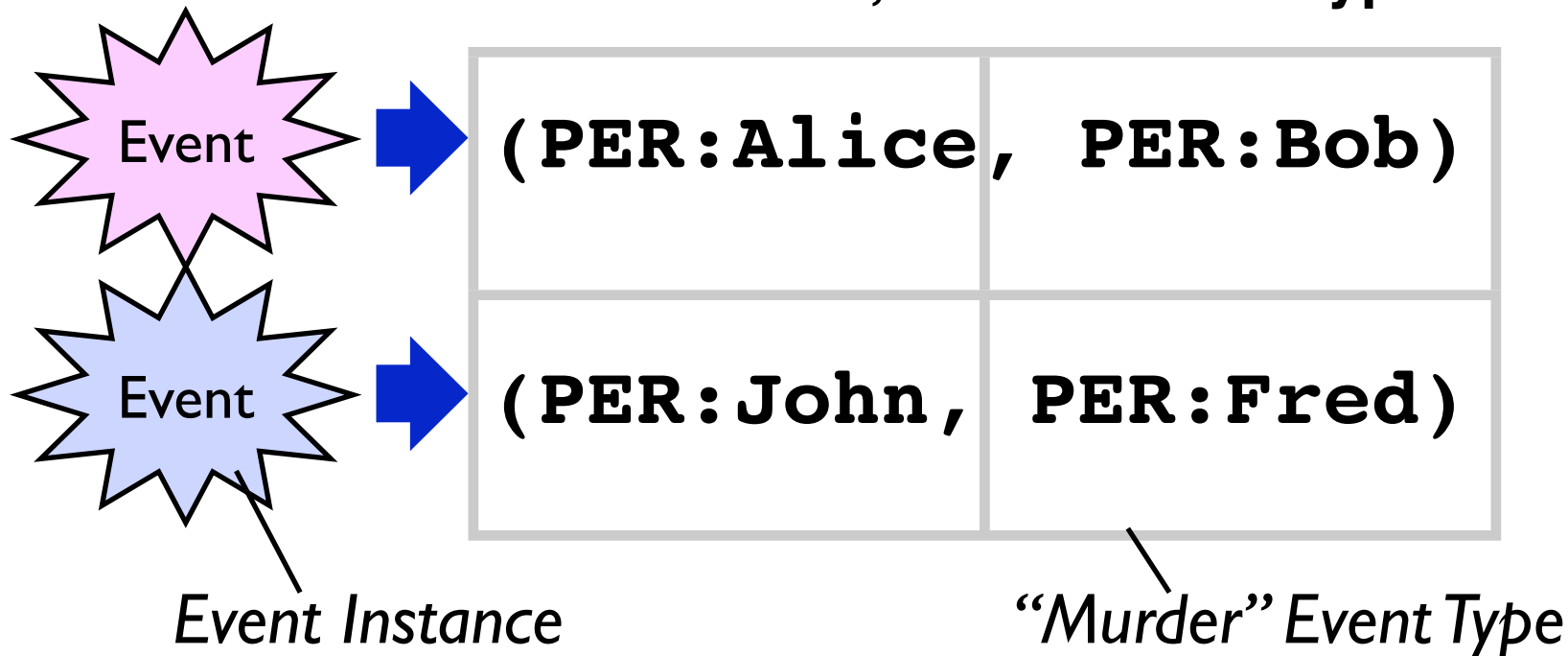
$$\mathbf{f}(e_{A.global}, e_{B.global})$$

Cosine distance



Clustering

- Hierarchical Clustering
 - Gradually form a table by adding entity tuples.
 - Row = Event **Instance**, Table = Event **Type**.



User Interface

- We obtain a lot of tables, but don't know what they actually represent.
- Provide an interface to browse the obtained tables:
 - A user can search entities, expressions and topic keywords (global features).
 - Show support articles that provide the actual context for each event.



Results

- Events: 35,398 (article sets)
- Local features (from 391,384 articles):
 - Token: 28,544,893
 - Type: 4,417,109
- Obtained clusters: 2,193 (tables)

Columns=2	1,878
Columns=3	282
Columns=4	33

10< rows	1,730
10~100 rows	444
100≥ rows	19



Results

- Obtained tables: (interpreted)

GPE + PER	PER visited GPE .
PER + PER	PER ₁ and PER ₂ got married.
ORG + PER	ORG acquired PER .
GPE + ORG	ORG put sanctions on GPE .
NAN + ORG	ORG reported a death by NAN .
ORG + PER + PER	ORG won, with PER ₁ and PER ₂ .
GPE + ORG + PER	ORG won in GPE election with PER .



Computational Cost

- Collect articles every day:
 - Crawling & Text extraction: 6 hours.
 - Feature extraction: 2 hours.
- Clustering:
 - About 3 hours for 35,398 events.
 - Worst time: $O(n^2)$
 - There are various speedup techniques.



Evaluation and Discussion

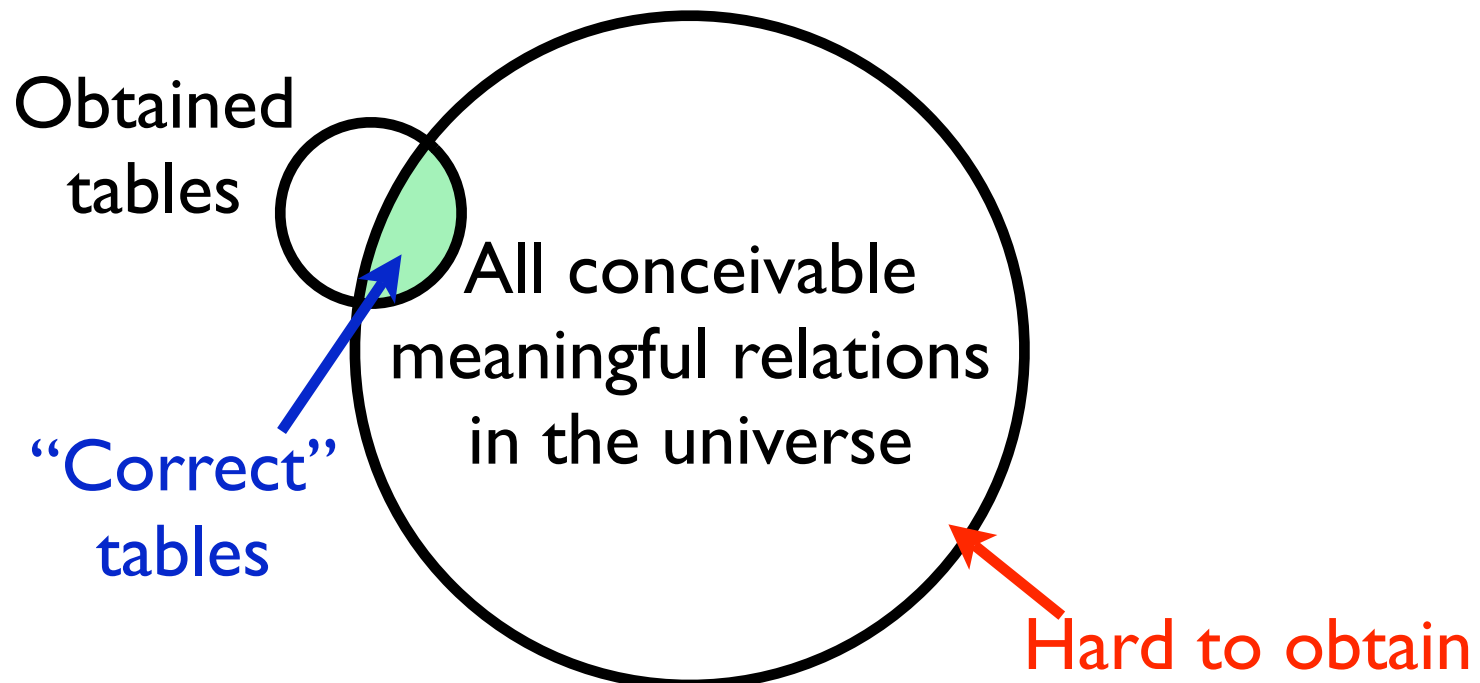
How Well Did It Work?

- Evaluate obtained clusters (tables).
- Two metrics:
 - Performance at *table* level.
 - How many good tables (relations) were obtained?
 - Performance at *event* level.
 - For each table, how many good event instances (rows) were extracted?
- **Warning:** this part is tricky!



At *Table* Level

- Performance at *table* level.
 - **Precision:** (Correct) / (Obtained)
 - **Recall:** (Correct) / (**All meaningful**)



At *Table* Level

- We use *ACE 2005 Event Corpus* as a representative set.
 - 332 news articles (bn + wn)
 - 2,104 event instances are manually annotated.
 - 33 event types are defined with a guideline.



At *Table* Level

- Performance at *table* level.
 - **Precision:** (Correct) / (Obtained)
 - **Recall:** (Correct) / (**ACE event types**)



At *Table* Level

- Evaluation method:
 - **Precision: 16/20 tables (80%)**
 - Pick 20 tables randomly. (not using ACE definition)
 - Review each table manually.
 - Try to find a reasonable interpretation for the table. (e.g. “**PER₁** killed **PER₂**”).
 - **Recall: 28/33 tables (84%)**
 - Try to find 33 tables with search interface.
 - Check if they are compatible with the ACE definitions.



Discovered Tables

- What kind of tables we discovered?
 - Many tables (tasks) that traditional IE systems have been tackling:
 - Event types defined in ACE.
 - Personal (birth, death, marriage, trip, arrest)
 - Military operations (attack, bombing)
 - Legal events (lawsuit, convict, sentence)
 - Business events (merger, share)
 - Natural disasters (hurricane, storm, virii)
 - Sports results



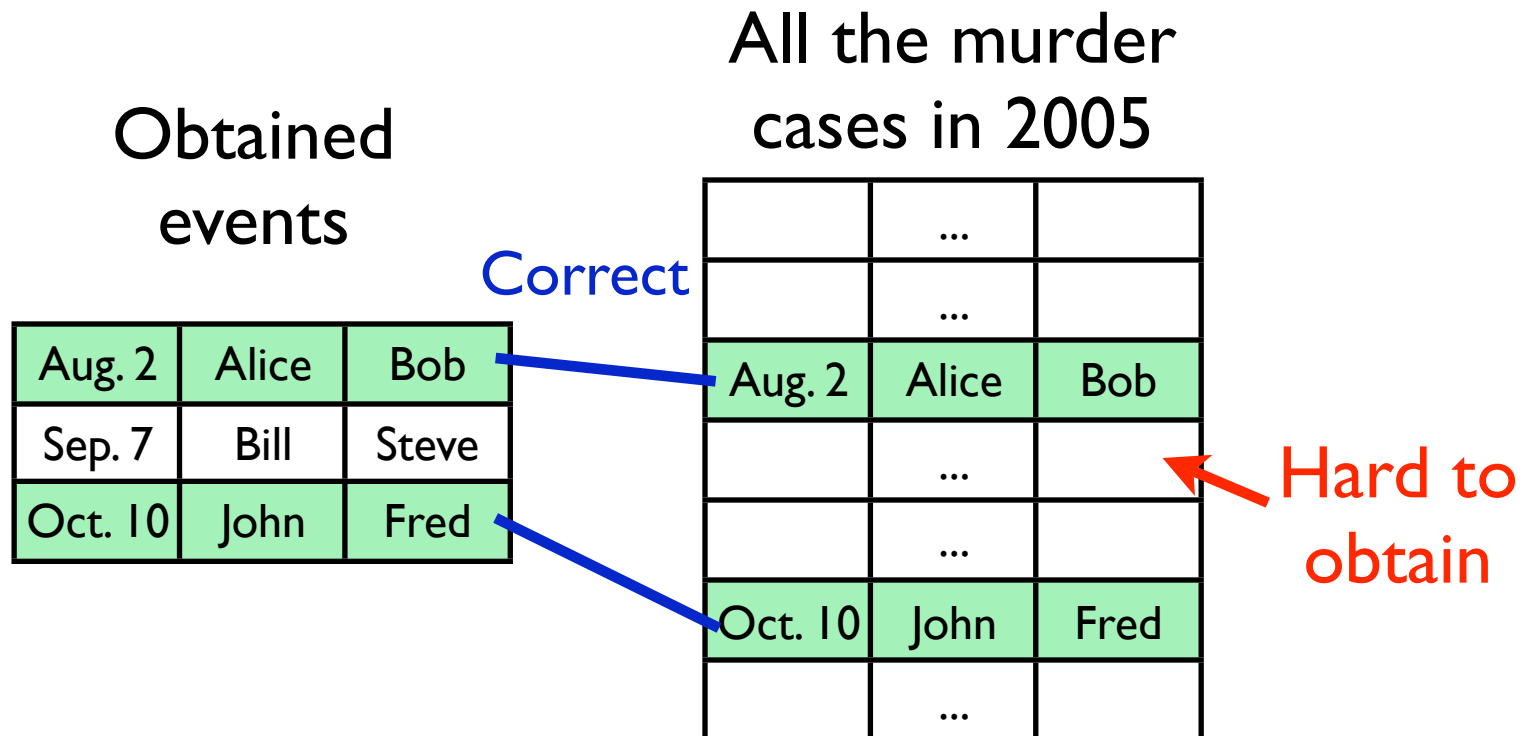
Undiscovered Tables

- What kind of tables we *couldn't* discover?
 - To answer this question, we looked into isolated events that were not grouped into any cluster.
 - “A ship capsized and sank in a lake.”
 - “Researchers discovered a treatment for virus.”
 - “Government’s plan for pandemics.”
 - “Book review.”
 - “Airplane ran off runways.”



At *Event* Level

- Performance at *event* level.
 - **Precision:** (Correct) / (Obtained)
 - **Recall:** (Correct) / (**All cases**)



At *Event* Level

- Performance at *event* level.
 - **Precision:** (Correct) / (Obtained)
 - **Recall:** (Correct) / (**ACE events**)

Obtained
events from ACE
corpus

Aug. 2	Alice	Bob
Sep. 7	Bill	Steve
Oct. 10	John	Fred

Correct

Annotated events
in ACE corpus

Aug. 2	Alice	Bob
	...	
Oct. 10	John	Fred
	...	



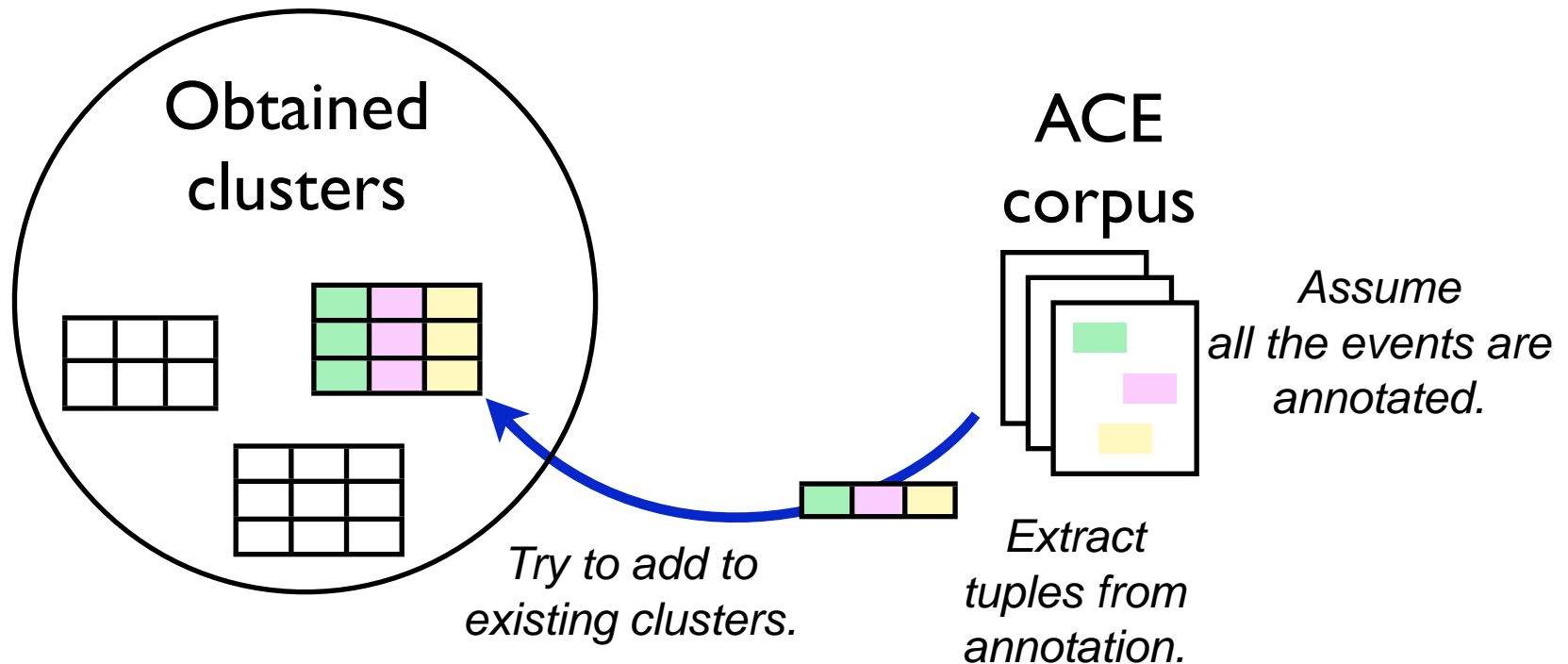
At *Event* Level

- Evaluation method:
 - **Precision: 105/161 rows (69%)**
 - Pick 20 tables whose relation is identified. (not ACE)
 - Review each row manually:
 - **Correct / Wrong relation / Wrong value**
 - Count how many rows are correct.
 - **Maximum Recall: 225/529 rows (45%)**
 - Try to add event instances in the ACE corpus to existing clusters.
 - Count how many rows were clustered.



At *Event* Level

- Evaluation method:



At *Event* Level

- Error analysis: Manually review 20 incorrect rows.
 - **Coreference error:**
 - Wrong expressions were associated to an entity.
 - **Irrelevant feature weights:**
 - The weight of expression “**die**” should have been much lower when they were used for creating the “***Murder***” table.
 - Current weighting schema:
 - $\text{Weight}(\mathbf{x}) = \log(1 / \text{frequency of } \mathbf{x})$



Overall Performance

- The system is useful!
- Possible improvements:
 - More Named Entity types.
 - Dates, Numeric expressions, Currencies, ...
 - Improve the coreference resolution.
 - Precision-oriented is preferred.
 - Better way of comparing of feature vectors.
 - Use a supervised method to learn the weights of features.
 - Alternative user interface.



Using By-products

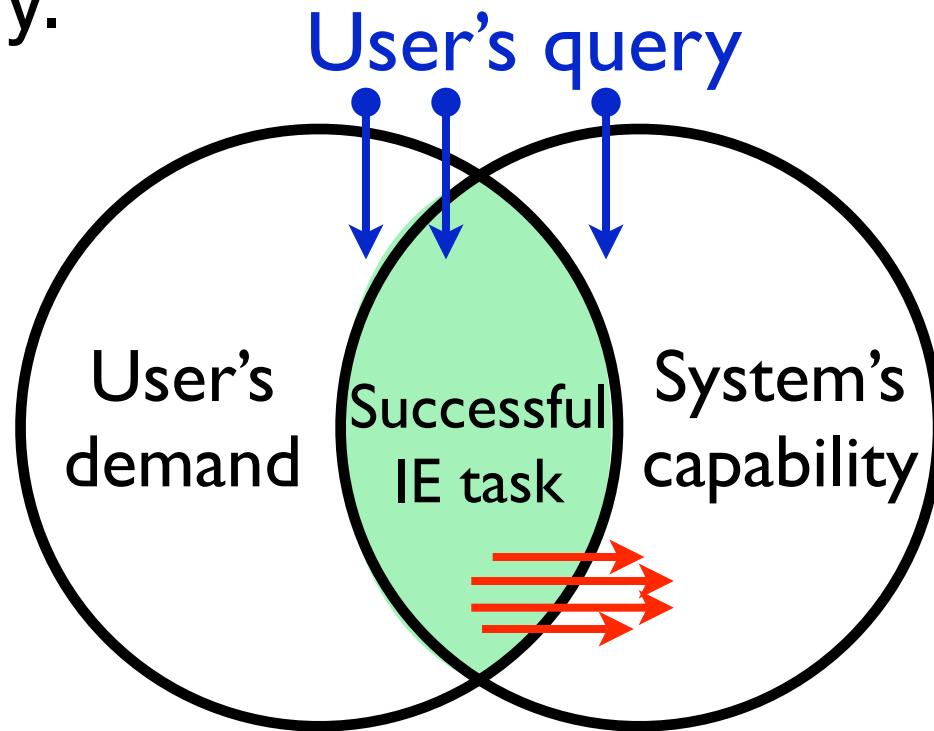
- Local features used in “*Murder*” table.
 - We could use them as IE patterns.

Murderer column	Victim column
PERSON's lawyer	PERSON's body
PERSON's attorney	PERSON is killed
PERSON is charged	PERSON's death
PERSON is convicted	PERSON's family



Related Work

- On-Demand IE [Sekine, 06]
 - Generate IE patterns based on a user's query.



Preemptive IE



Related Work

- Relation Discovery [Hasegawa, 04] [Banko, 07]
 - Use surface clues.
 - Word sequence, shallow parser, etc.
 - Better at collecting well-defined relations that appear again and again.
 - e.g. (“**Bill Gates**”, “**Microsoft**”)
 - Static knowledge.



Wrap-up

Conclusion

- Proposed a new approach of IE.
- Implemented and evaluated a system.
 - Discovered a lot of meaningful tables.
 - Traditional tasks: Murder, Natural Disaster, ...
 - Non-traditional tasks: Sports, Product launch, ...
- Hope this can be used to explore potential IE tasks.



Future Work

- Broader directions:
 - How to handle the *generality* of relations objectively?
 - Currently “death by murder,” “death by war,” “death by illness” and “death by natural disaster” are separated. But some people might want a more general category.
 - The degree of generality can be adjusted by changing the clustering threshold, but is the degree of generality one dimensional?



Thanks!

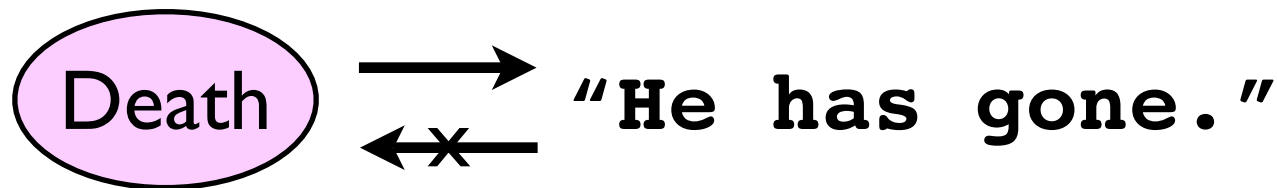


Related Work

- Difference between paraphrase discovery and IE pattern discovery?

- Paraphrase discovery:

- Collect expression s such that $e \Rightarrow s$.



- IE pattern discovery:

- Collect expression s such that $s \Rightarrow e$.

