

NYU

# Refining Event Extraction (Old-Fashion 'Traditional' IE) Through Cross-document Inference

Heng Ji and Ralph Grishman

(hengji,grishman)[@cs.nyu.edu](mailto:cs.nyu.edu)

Computer Science Department  
New York University  
June, 2008



# Outline

---

- Background: Event Extraction and Its Performance Limitation
- Motivation: Event Extraction Beyond Document Boundary
- Approach Overview
  - System Pipeline
  - Baseline Within-Sentence Event Extraction
  - Information Retrieval and Query Construction
- Global Confidence Estimation and Inference
- Experimental Results
  - Confidence Metric Thresholding
  - Overall Performance
- Conclusion and Future Work



# Event Extraction: 'Traditional' IE

Barry Diller on Wednesday *quit* as chief of Vivendi Universal Entertainment t.

Trigger	Quit (a "Personnel/End-Position" event)	
Arguments	Role = Person	Barry Diller
	Role = Organization	Vivendi Universal Entertainment
	Role = Position	Chief
	Role = Time-within	Wednesday



vivendi

- Target: 33 different types of Automatic Content Extraction (ACE) events



# IE Beyond Document Boundary

---

- Most event extraction systems operate a sentence a time; MUC-style Event Extraction hit the 60% 'performance ceiling'
- Look back at the initial goal of IE
  - Create a database of relations and events from the ***entire corpus***
  - Within-doc/Within-Sent IE was an artificial constraint to simplify the task and evaluation
- Many events will be reported multiple time in different forms
  - Get ***background knowledge*** from ***a cluster of topically-related documents***
  - Favor ***interpretation consistency*** within each cluster
- Hypotheses
  - One Trigger Sense Per Cluster
  - One Argument Role Per Cluster



# One Trigger Sense Per Cluster

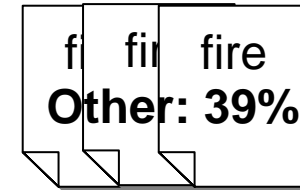
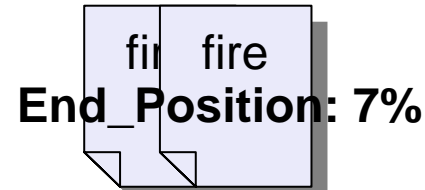
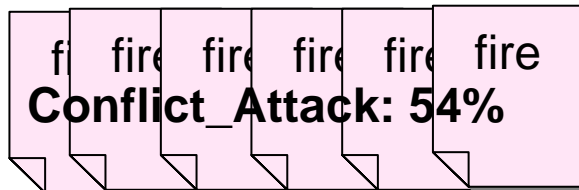
## Test Doc

It took a federal act -- by the Canadian government --to give Martha Stewart permission to paddle a hollowed-out, 600-pound pumpkin across a lake in Windsor, Nova Scotia. Martha Stewart was planning to **fire** the pumpkin. write it a nice letter and let it appear on her daytime show...



**Conflict\_Attack or End\_Position?**

## Training Corpora

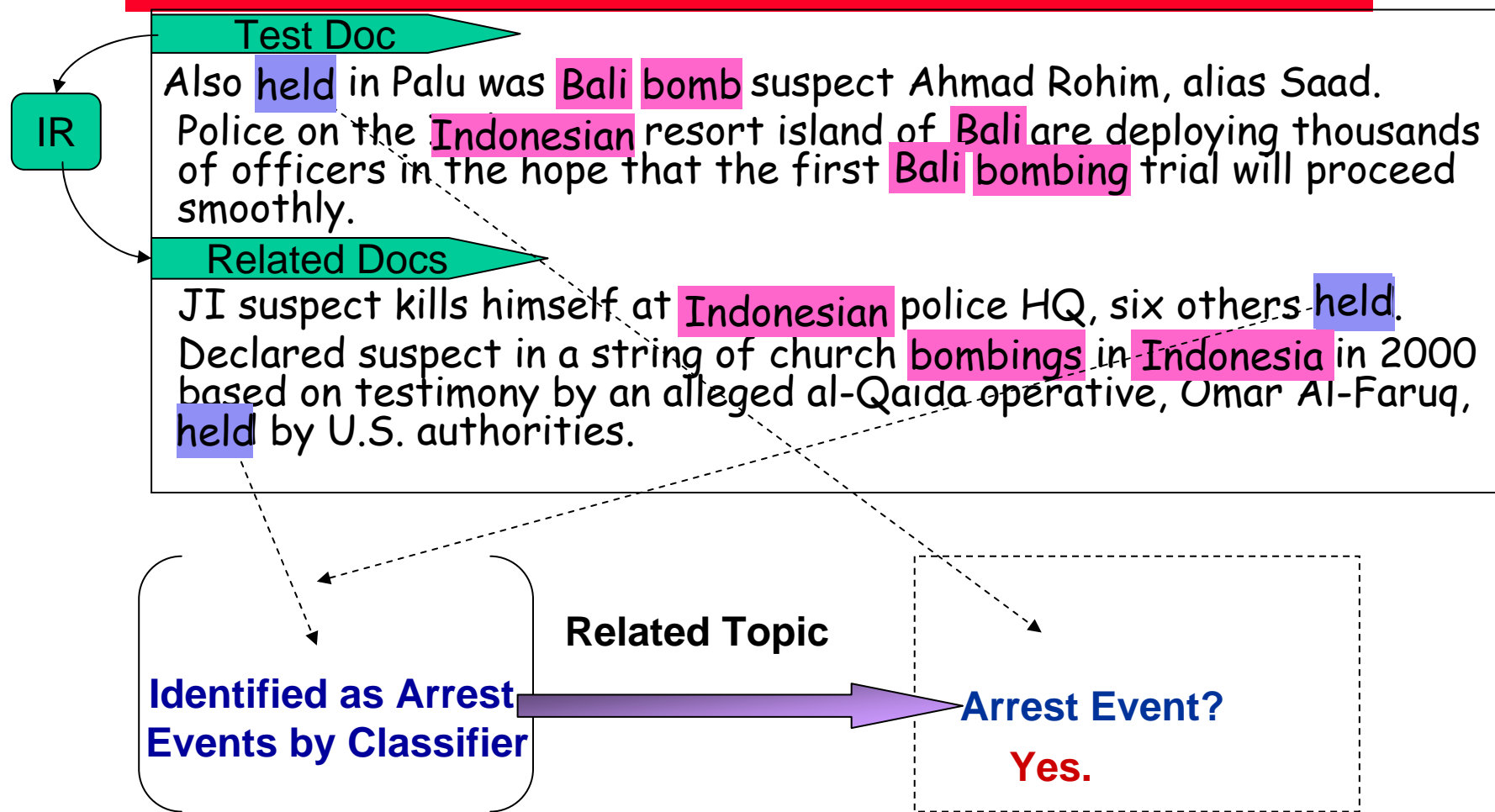


## Related Docs



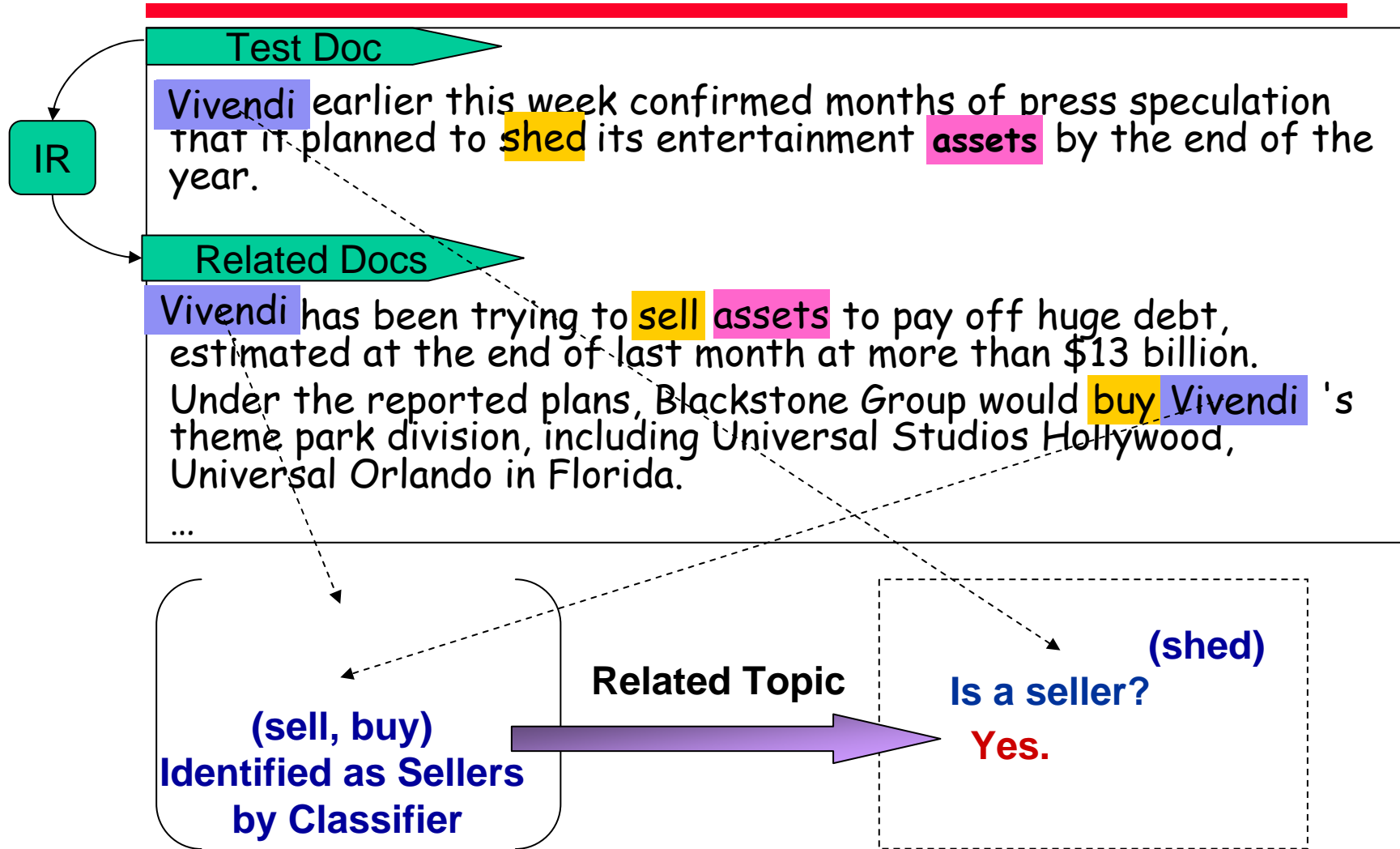


# One Trigger Sense Per Cluster (Cont')





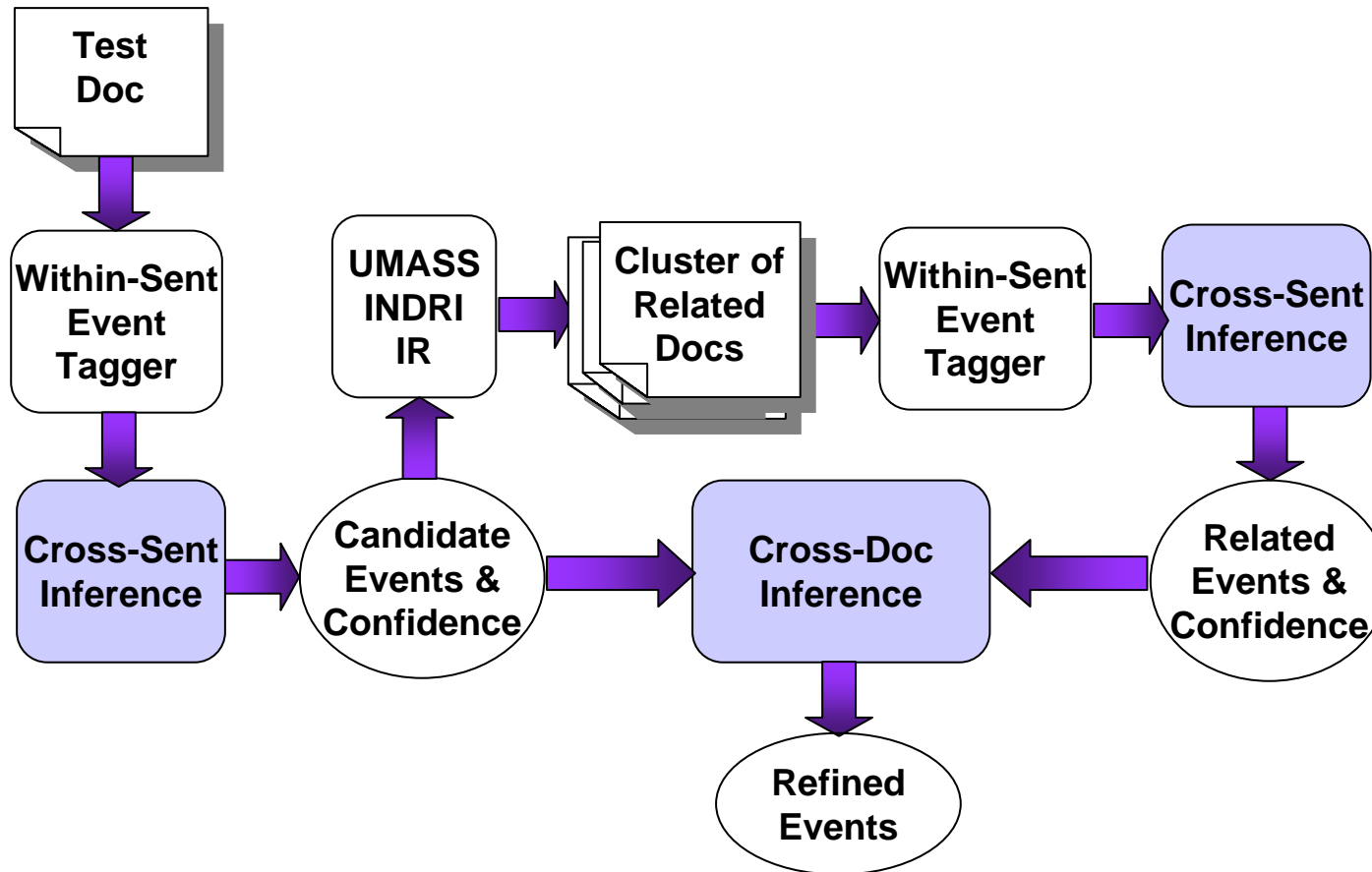
# One Argument Role Per Cluster





# Cross-Sent/Cross-Doc Event Inference Architecture

---





# Baseline Within-Sentence Event Extraction

---

## 1. Pattern matching

- Build a pattern from each ACE training example of an event
  - British and US forces reported gains in the advance on Baghdad  
→ PER report gain in advance on LOC

## 2. MaxEnt models

- ① Trigger Classifier
  - to distinguish event instances from non-events, to classify event instances by type
- ② Argument Classifier
  - to distinguish arguments from non-arguments
- ③ Role Classifier
  - to classify arguments by argument role
- ④ Reportable-Event Classifier
  - to determine whether there is a reportable event instance



# IR Query Construction and Expansion

Test Doc

**Barry Diller** on **Wednesday** quit as chief of **Vivendi Universal Entertainment**, the entertainment unit of French giant Vivendi Universal. Diller took on the "provisional" role at the top of Vivendi's US entertainment operations...

Query

```
<query>#combine( #weight(0.139503 #1(step) 0.860497 #1(quit))  
#weight( 0.076629 #1(Barry Diller) 0.197804 #1(Vivendi)  
0.043396 #1(French) 0.197804 #1(Vivendi Universal)  
0.175093 #1(Vivendi Universal Entertainment)  
0.041749 #1(chief of Vivendi Universal Entertainment)  
0.223145 #1(Diller) 0.042312 #1(Wednesday)))</query>
```

Synonym

Coreferred name

Related Name

Weighted by (local confidence \* frequency)



# Global Confidence Estimation

---

- Within-Sentence IE system produces local confidence
- IR engine returns a cluster of related docs for each test doc
- Document-wide and Cluster-wide Confidence
  - Frequency weighted by local confidence
  - *XDoc-Trigger-Freq(trigger, etype)*: The weighted frequency of string *trigger* appearing as the trigger of an event of type *etype* across all related documents
  - *XDoc-Arg-Freq(arg, etype)*: The weighted frequency of *arg* appearing as an argument of an event of type *etype* across all related documents
  - *XDoc-Role-Freq(arg, etype, role)*: The weighted frequency of *arg* appearing as an argument of an event of type *etype* with role *role* across all related documents
  - *Margin* between the most frequent value and the second most frequent value, applied to resolve classification ambiguities
  - .....



# Cross-Sent/Cross-Doc Event Inference Procedure

---

- Remove triggers and argument annotations with local or cross-doc confidence lower than thresholds
  - *Local-Remove*: Remove annotations with low local confidence
  - *XDoc-Remove*: Remove annotations with low cross-doc confidence
  
- Adjust trigger and argument identification and classification to achieve document-wide and cluster-wide consistency
  - *XSent-Iden/XDoc-Iden*: If the highest frequency is larger than a threshold, propagate the most frequent type to all unlabeled candidates with the same strings
  - *XSent-Class/XDoc-Class*: If the margin value is higher than a threshold, propagate the most frequent type and role to replace low-confidence annotations



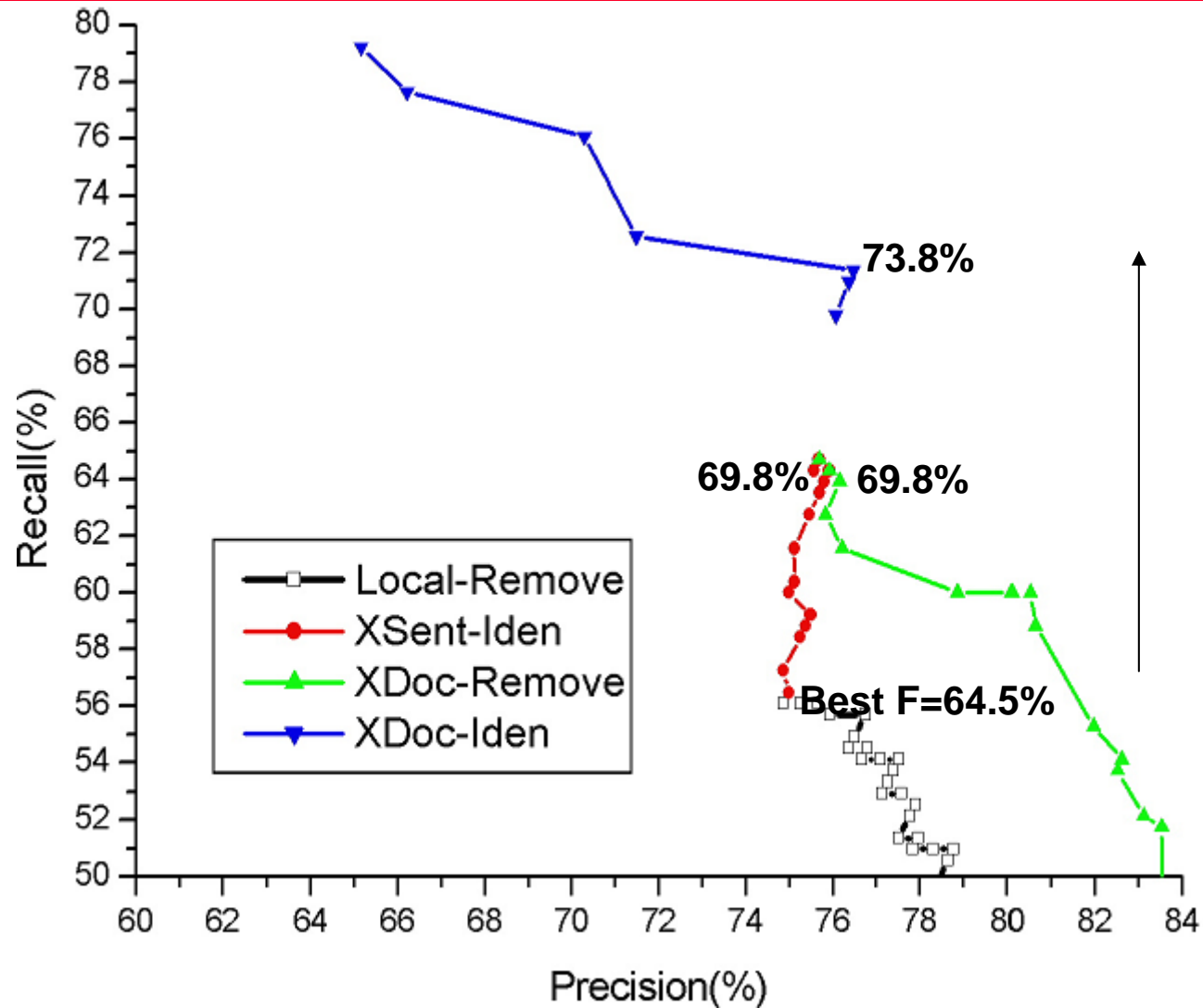
# Experiments: Data and Setting

---

- Within-Sentence baseline IE trained from 500 English ACE05 texts (from March – May of 2003)
- Use 10 ACE05 newswire texts as development set to optimize the global confidence thresholds and apply them for blind test
- Blind test on 40 ACE05 texts, for each test text, retrieved 25 related texts from TDT5 corpus (278,108 texts, from April-Sept. of 2003)



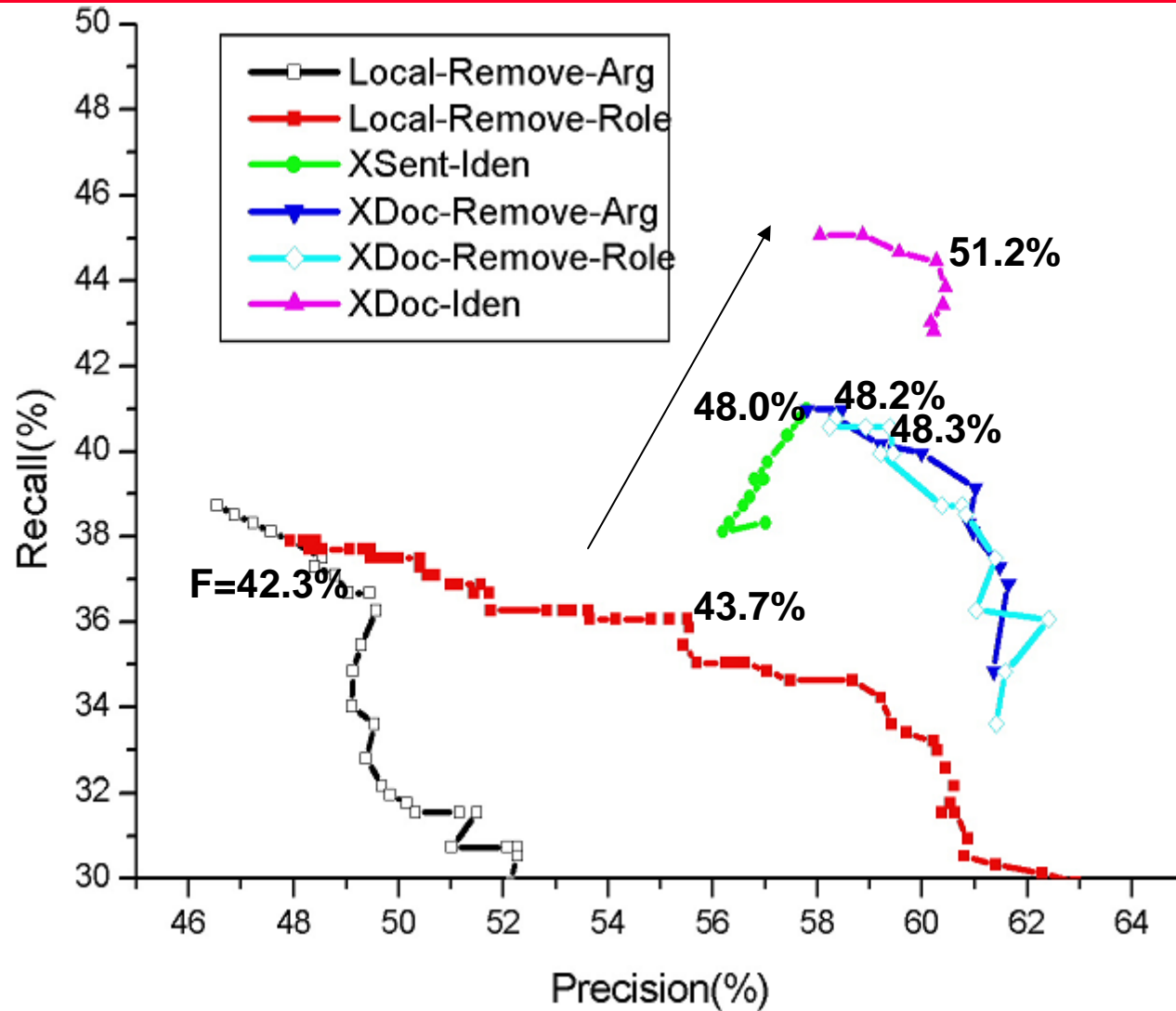
# Selecting Trigger Confidence Thresholds to optimize Event Identification F-measure on Dev Set





# Selecting Argument Confidence Thresholds

to optimize Argument Labeling F-measure on Dev Set





# Experiments: Trigger Labeling

---

Performance System/Human	Precision	Recall	F-Measure
Within-Sent IE (Baseline)	67.6	53.5	59.7
After Cross-Sent Inference	64.3	59.4	61.8
After Cross-Doc Inference	60.2	76.4	67.3
Human Annotator 1	59.2	59.4	59.3
Human Annotator 2	69.2	75.0	72.0
Inter-Adjudicator Agreement	83.2	74.8	78.8

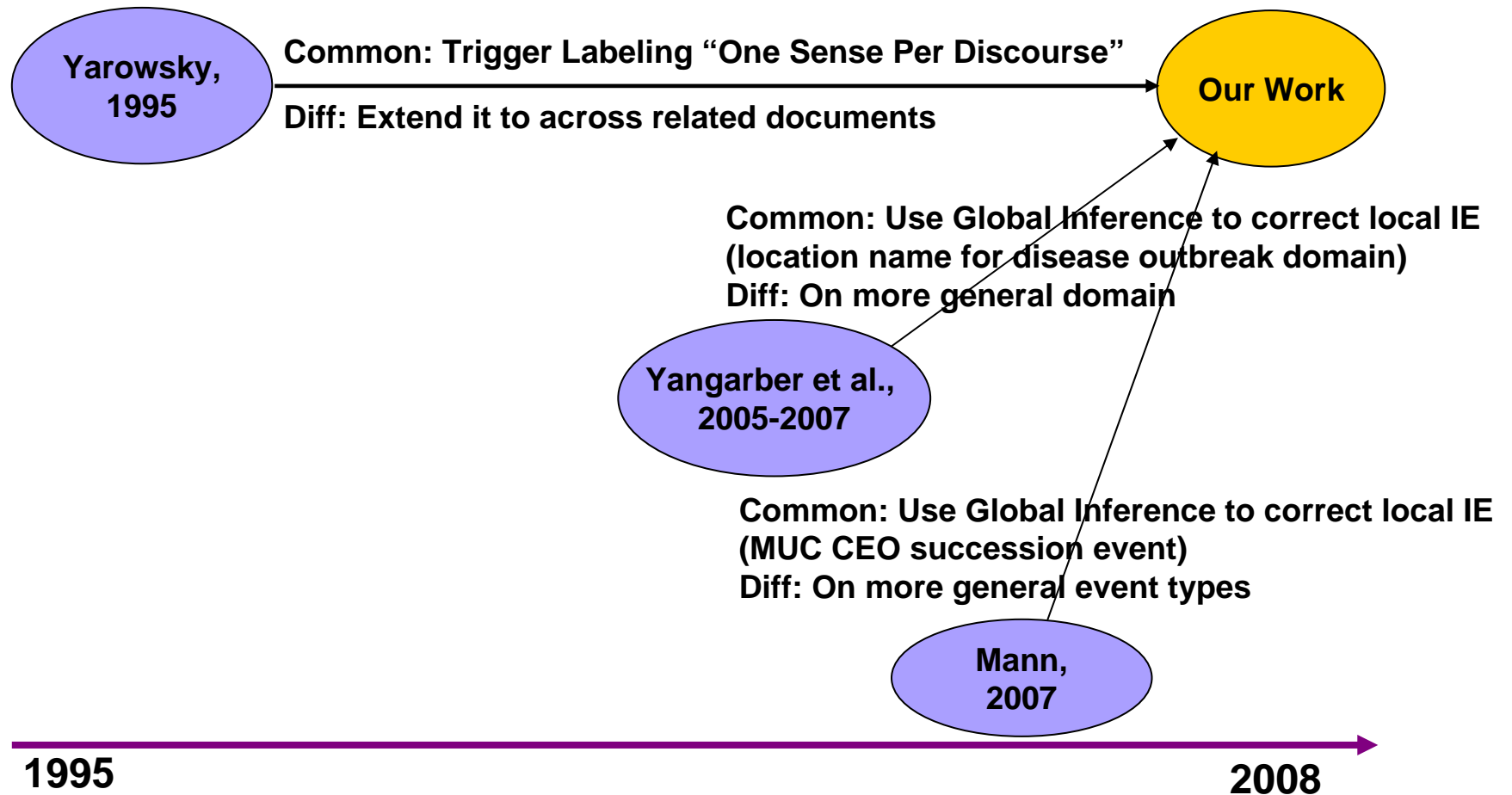


# Experiments: Argument Labeling

Performance System/Human	Argument Identification			Argument Classification Accuracy	Argument Identification + Classification		
	P	R	F		P	R	F
Within-Sent IE	47.8	38.3	42.5	86.0	41.2	32.9	36.3
After Cross-Sent Inference	54.6	38.5	45.1	90.2	49.2	34.7	40.7
After Cross-Doc Inference	55.7	39.5	46.2	92.1	51.3	36.4	42.6
Human Annotator 1	60.0	69.4	64.4	85.8	51.6	59.5	55.3
Human Annotator 2	62.7	85.4	72.3	86.3	54.1	73.7	62.4
Inter-Adjudicator Agreement	72.2	71.4	71.8	91.8	66.3	65.6	65.9



# Related Work





# Conclusion and Future Work

---

- Proposed a new approach to break the document boundaries for IE
- Gather together background knowledge from a set of related documents, and then apply inference to enhance traditional IE performance
- Recent work proved the same approach can improve relation extraction and social network extraction
- Future Work
  - Use results as seeds for unsupervised/open IE
  - Develop Event-driven multi-document summarization
  - Derive entailment rules from related events in different timeframes
  - Long term: perform essential information reasoning and event prediction



Thank you

---





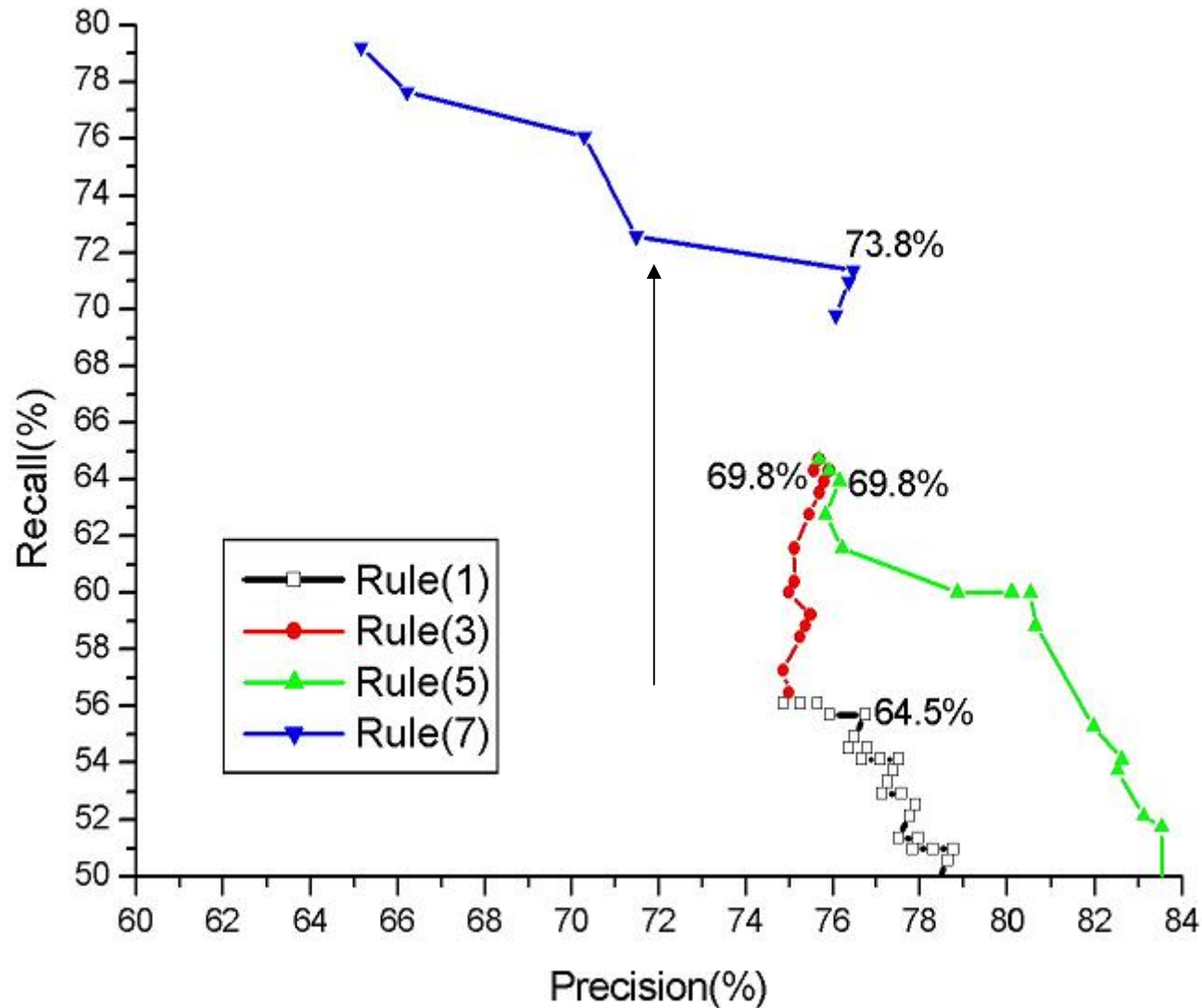
# One Trigger Sense Per Cluster

---

Event Trigger	Event Type	In Training Corpora	In Test Doc	In Related Docs
<i>advance</i>	Movement_Transport	31% of 16	50% of 12	88.9% of 27
<i>fire</i>	Personnel_End-Position	7% of 81	100% of 2	100% of 10
	Conflict_Attack	54% of 81	100% of 3	100% of 19
<i>replace</i>	Personnel_End-Position	5% of 20	100% of 1	83.3% of 6
<i>form</i>	Business_Start-Org	12% of 8	100% of 2	100% of 23
<i>talk</i>	Contact_Meet	59% of 74	100% of 4	100% of 26



# Selecting Trigger Confidence Thresholds to optimize Event Identification F-measure on Dev Set





# Selecting Argument Confidence Thresholds

to optimize Argument Labeling F-measure on Dev Set

