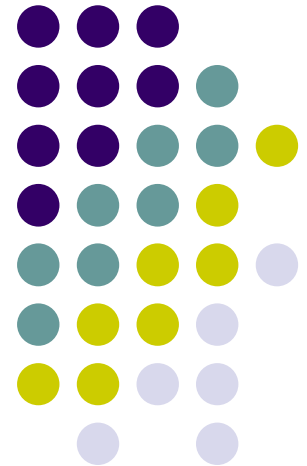
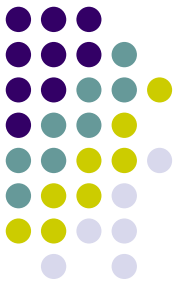


Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions

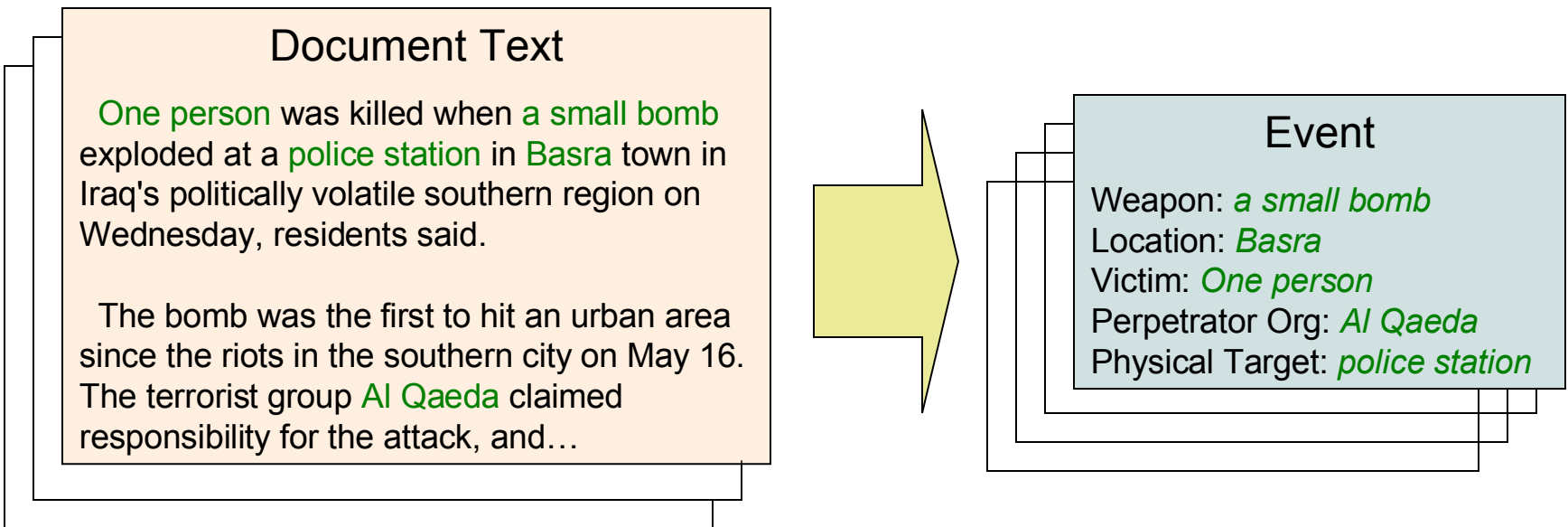
Siddharth Patwardhan & Ellen Riloff
University of Utah



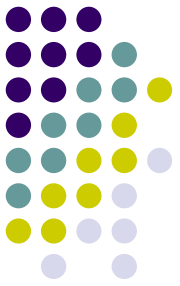


Event-oriented IE

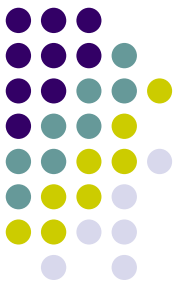
- Goal of IE system: extract facts associated with events from unstructured text



Approaches to IE

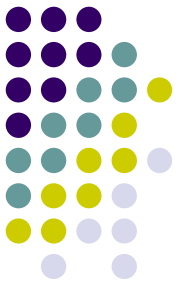


- **Pattern-based:** Kim & Moldovan 1993, Soderland et al. 1995, Riloff 1996, Soderland 1999, Yangarber et al. 2000, Agichtein & Gravano 2000, Sudo et al. 2003, Popescu et al. 2004, Yakushiji et al. 2006, among others
- **Classifier-based:** Freitag 1998, Freitag & McCallum 2000, Chieu et al. 2003, Bunescu & Mooney 2004, Ciravegna 2001, Finn & Kushmerick 2004, Li et al. 2005, Finkel et al. 2005, among others



IE Pattern Responsibilities

- IE rules or patterns have two primary responsibilities:
 - Identifying a relevant region
 - Locating a span of text to be extracted
- They rely on local context for clues to relevant information
- For example,
<subject> was assassinated

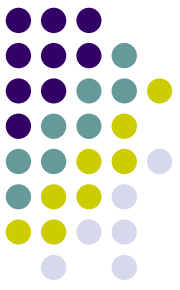


Problem #1

...the explosion ripped through the busy central neighborhood of New Delhi. **A bomb was found** under a parked car...

<subject> was found

- Lost recall due to patterns that are not event-specific

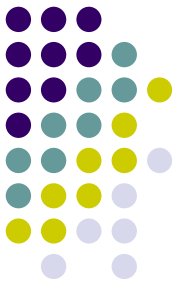


Problem #2

unleashed attacks on <np>

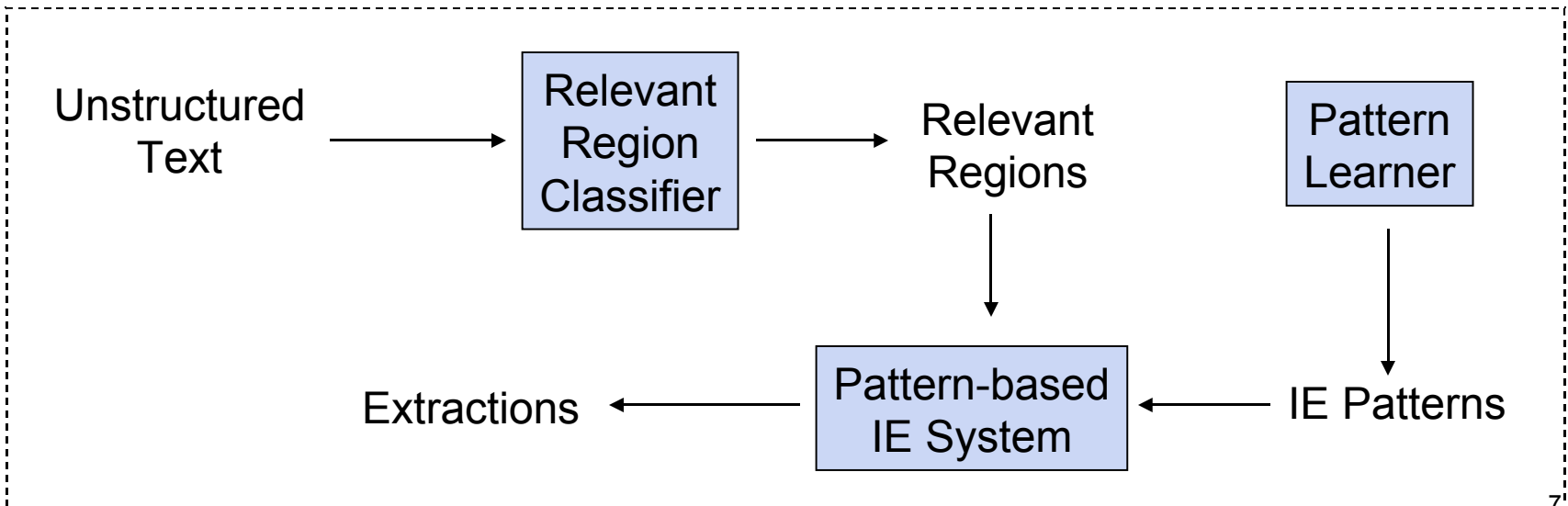
... in his speech on Sunday, Kerry **unleashed harsh attacks on George Bush** regarding his stand on immigration...

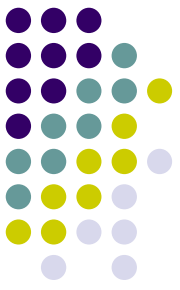
- Does not consider the “global” context
- Seemingly good patterns may result in false positives



Our Two-Pass Approach to IE

- Identify relevant regions of text using a relevant region classifier
- Selectively apply “semantically appropriate” IE patterns in these regions

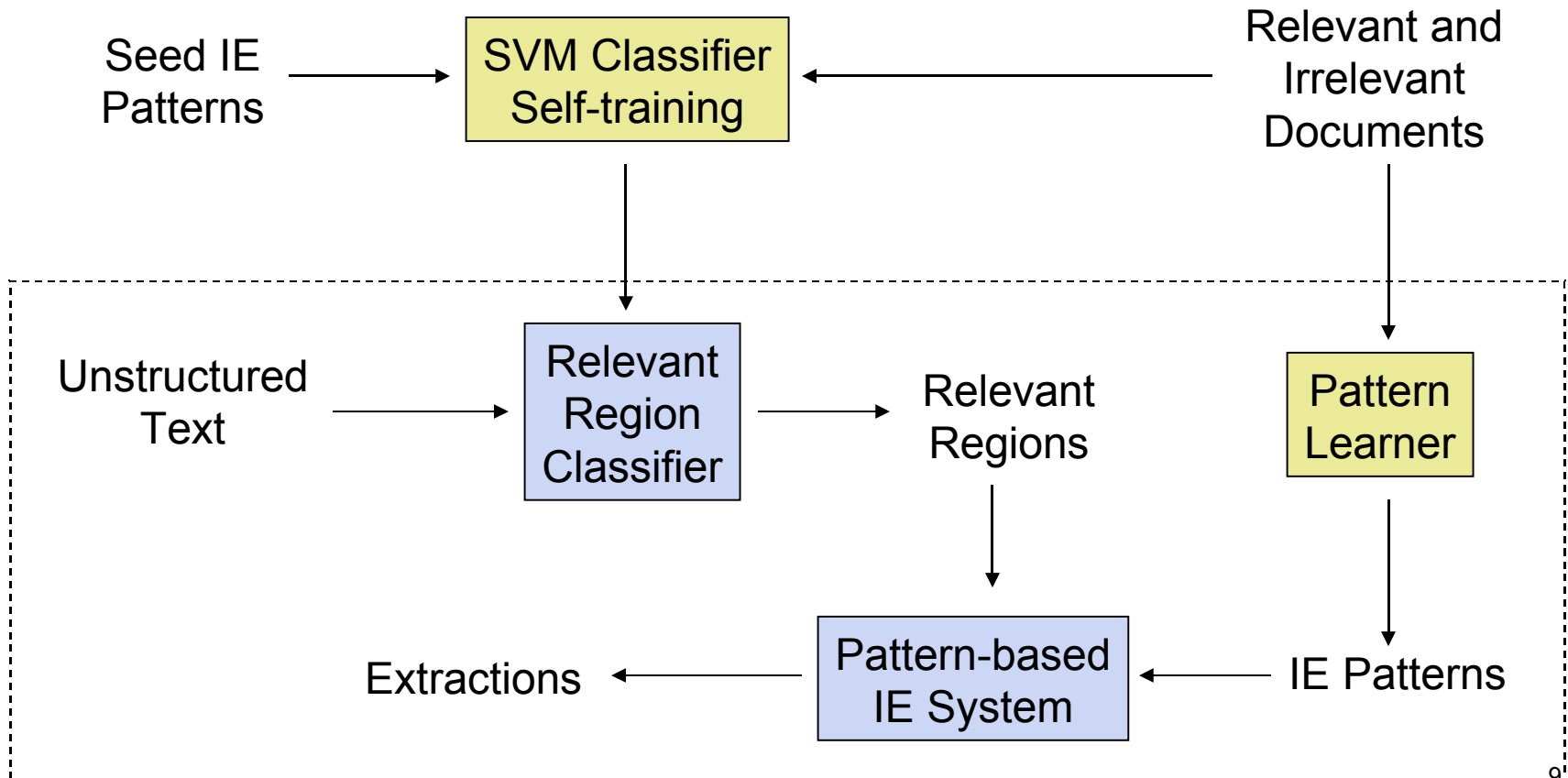
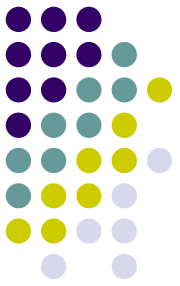




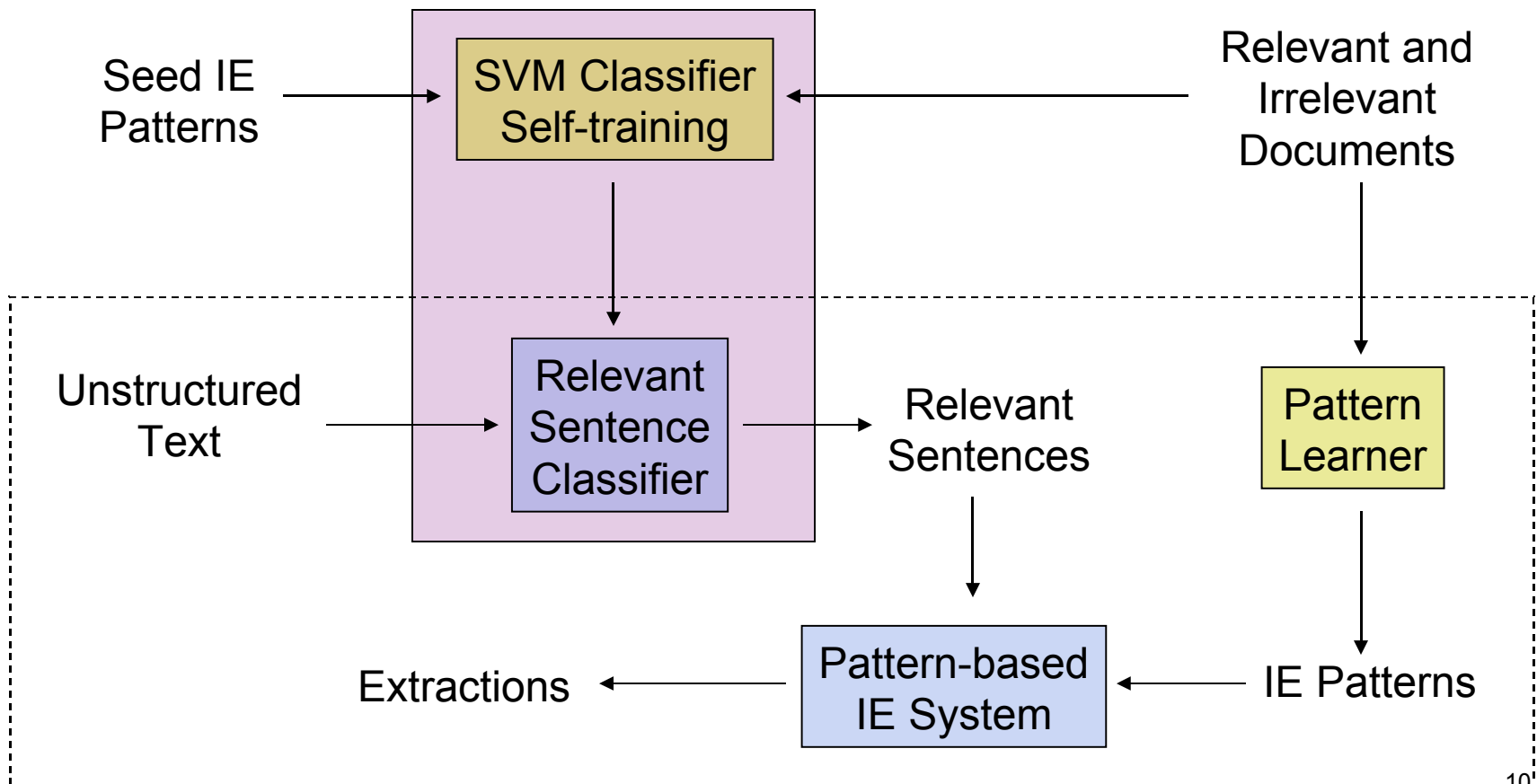
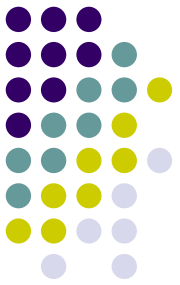
Benefits of Two Passes

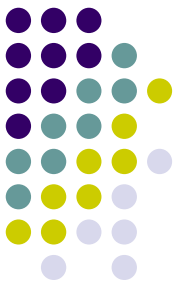
- Potentially improve precision by applying patterns only in relevant contexts
- Allows the use of non-event-specific patterns, such as *<subject> was found*
- Simplifies the learning process by having two separate subtasks

System Architecture



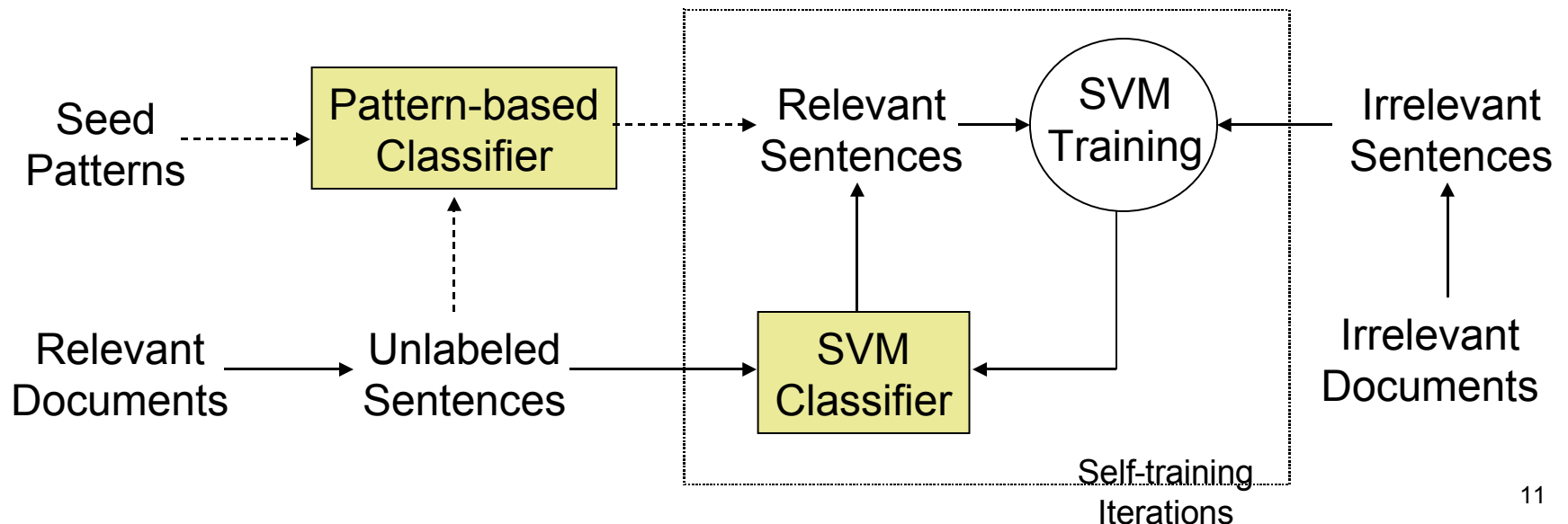
System Architecture



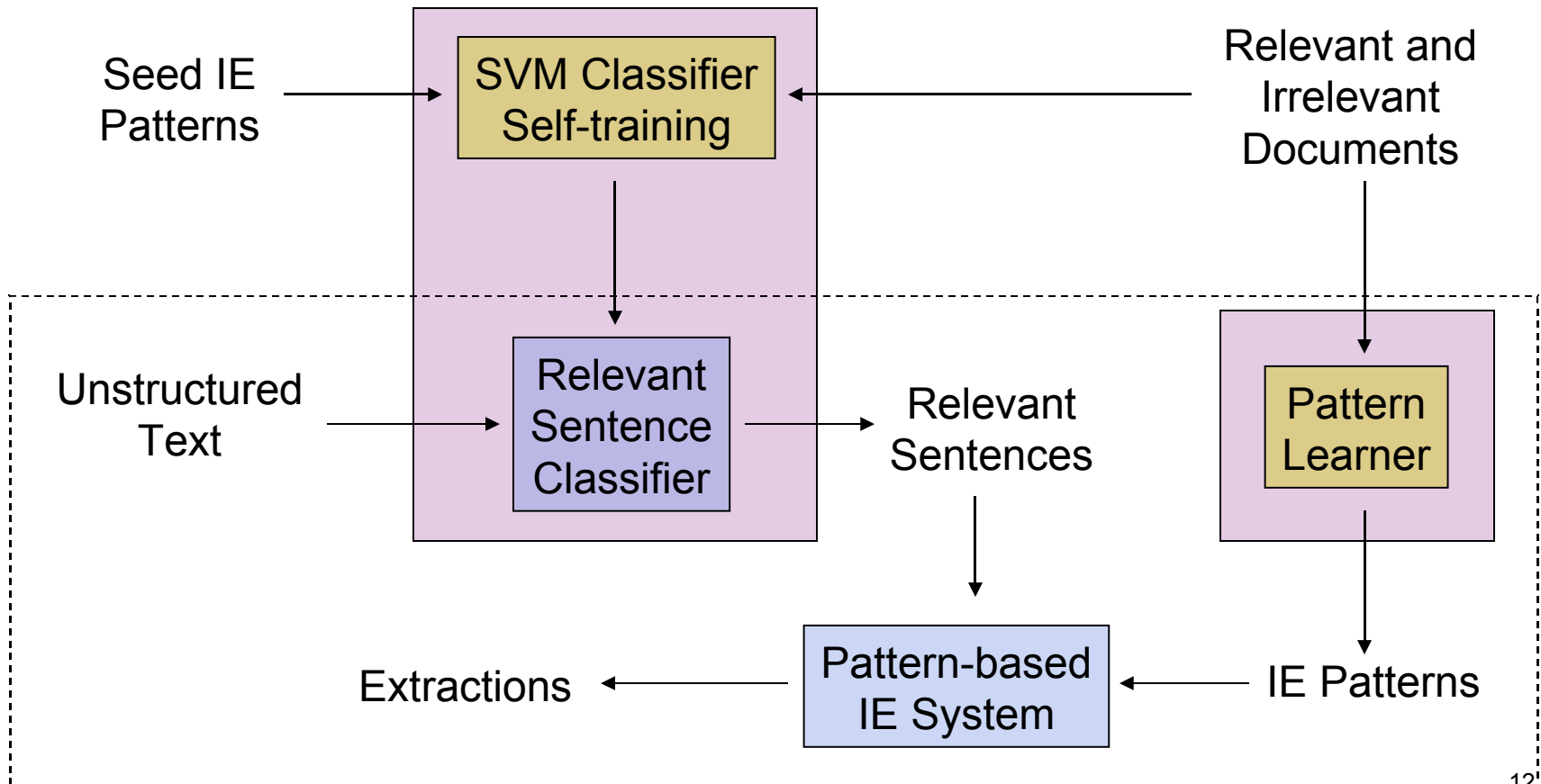
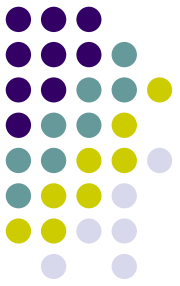


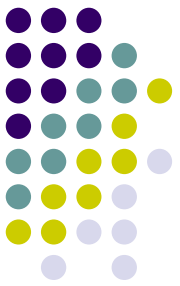
Relevant Sentence Classifier

- Iterative self-training learning process – does not require annotated data
- Input: seed IE patterns, a set of relevant documents and irrelevant documents



System Architecture

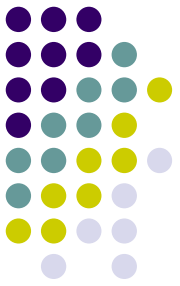




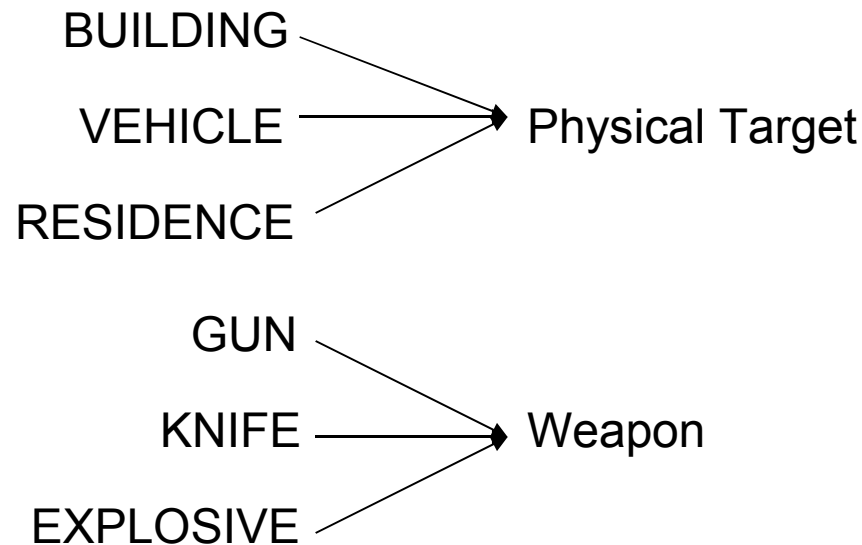
Semantic Affinity Patterns

- Given relevant regions, we only need to learn patterns that are “semantically appropriate”
- **Semantic Affinity:** *tendency of a pattern to extract phrases belonging to specific semantic categories*

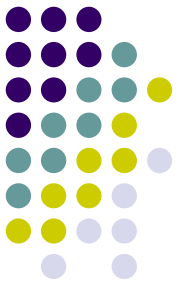
Semantic Affinity Pattern Learner



- First, create mapping between semantic categories present in semantic dictionary and the list of event roles for our task



Semantic Affinity Pattern Learner

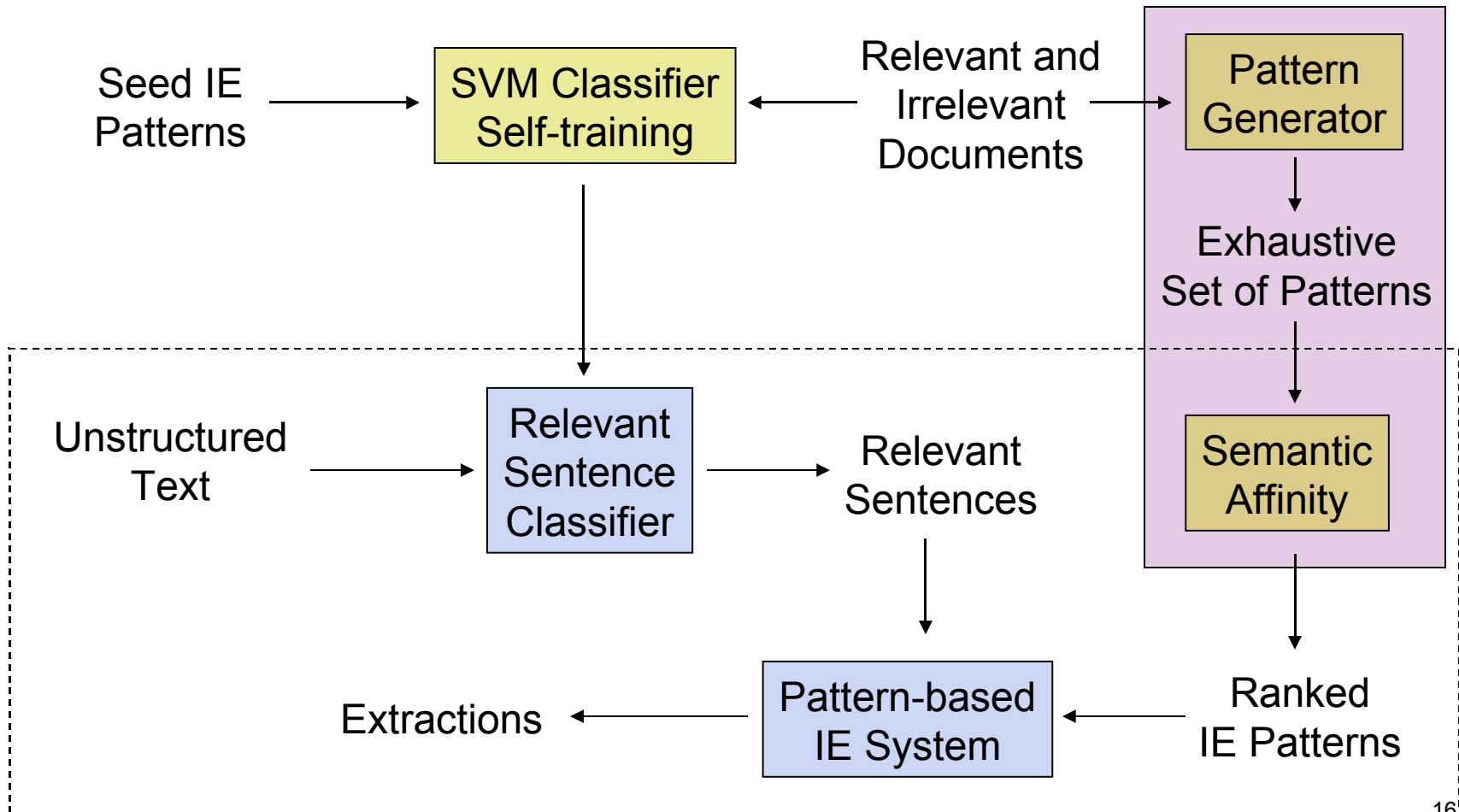
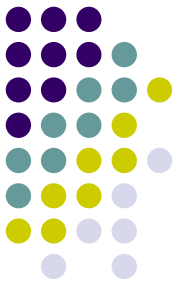


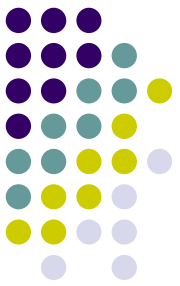
- For each pattern p and event role r_k , compute frequency $f(p, r_k)$ of the extractions whose semantic categories map to r_k
- Semantic Affinity:

$$\text{sem_aff}(p, r_k) = \frac{f(p, r_k)}{\sum_{i=1}^{|R|} f(p, r_i)} \log_2 f(p, r_k)$$

where $|R|$ is the number of event roles

System Architecture

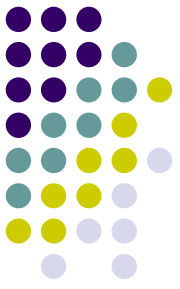




Some Top Ranked Patterns

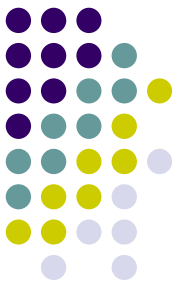
Target	Victim	Weapon	PerpOrg
destroyed <dobj> barrels of <np> shattered <dobj> <subj> was damaged blew up <np>	murder of <np> assassination of <np> killing of <np> <subj> question murdered <dobj>	<subj> exploded planted <dobj> fired <dobj> <subj> was planted explosion of <np>	<subj> claimed panama from <np> kidnapped by <np> command of <np> wing of <np>

PerpInd	Disease	Victim
<subj> blew up <subj> attacked identity of <np> bands of <np> gangs of <np>	cases of <np> spread of <np> outbreak of <np> <# th outbreak> <# outbreaks>	<# people> <# cases> <# birds> <# animals> <subj> died



Primary & Secondary Patterns

- Certain patterns are strong indicators of relevance by themselves (for e.g., *<subject> was assassinated*)
- Since the sentence classifier is not perfect, we could be missing some relevant extractions
- Potential recall gain by applying such patterns to *all* sentences

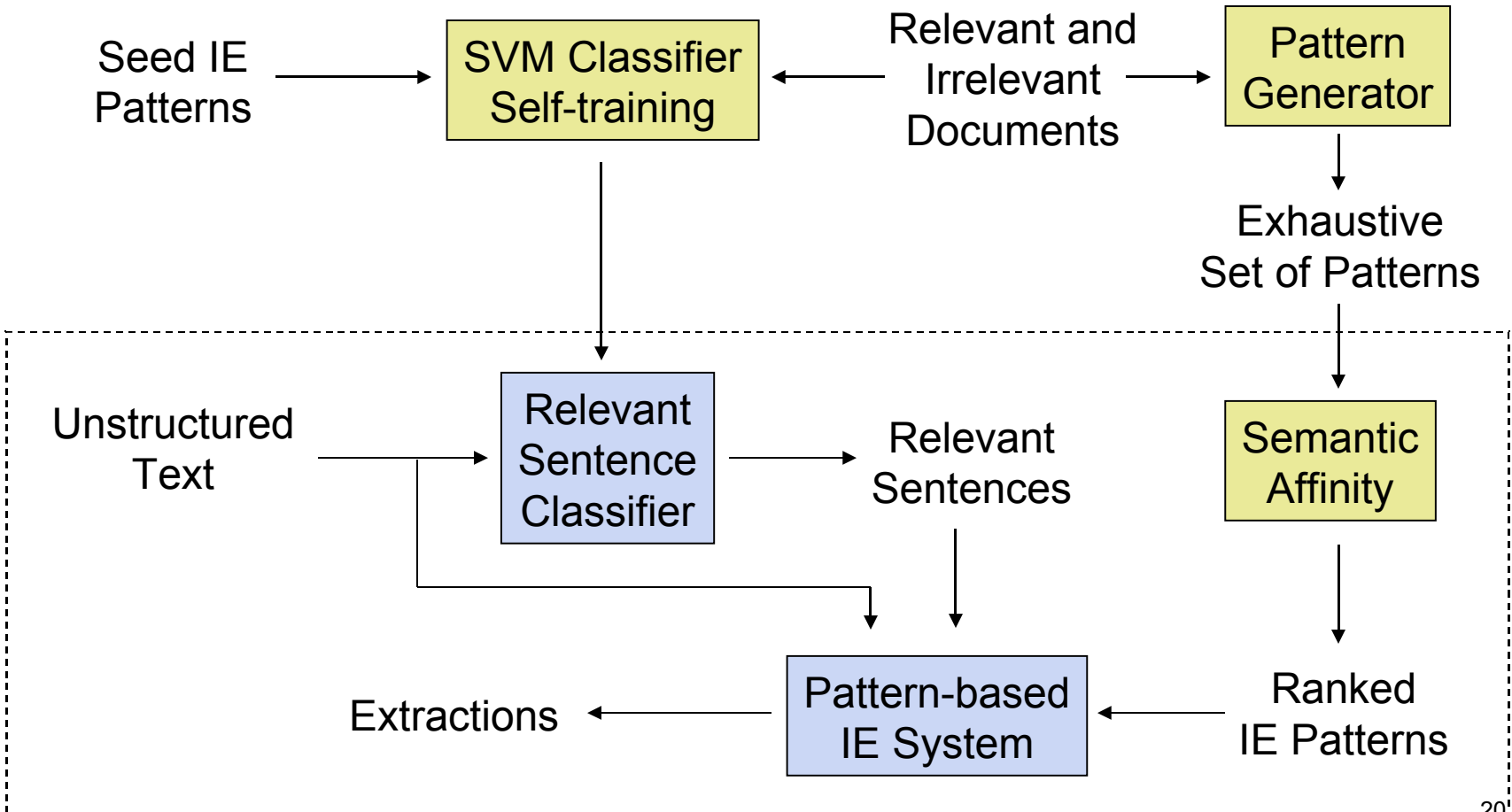
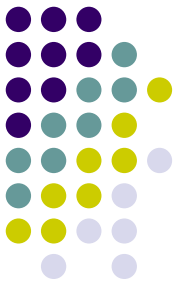


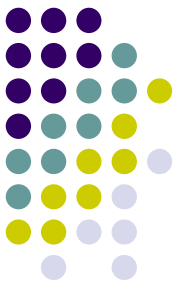
Primary & Secondary Patterns

$$P_{\text{rel}}(p_i) = \frac{\text{frequency of } p_i \text{ in relevant docs}}{\text{frequency of } p_i \text{ in all docs}}$$

- **Primary Patterns** ($P_{\text{rel}} \geq 0.8$): applied to all sentences
- **Secondary Patterns** ($P_{\text{rel}} < 0.8$) applied only in relevant sentences
- **Irrelevant Patterns** ($P_{\text{rel}} < 0.5$): patterns deleted

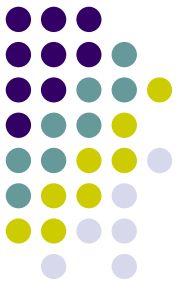
System Architecture





Evaluation

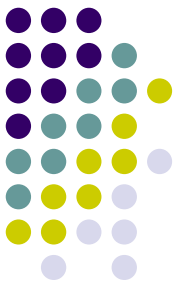
- **Terrorism: MUC-4 data**
 - 1300 training (615 relevant + 685 irrelevant)
 - 200 tuning documents
 - 200 test documents
- **Disease Outbreaks:**
 - 2000 ProMed (relevant) + 4000 PubMed (irrelevant) = 6000 training documents
 - 125 tuning documents
 - 120 test documents



Sentence Classifier

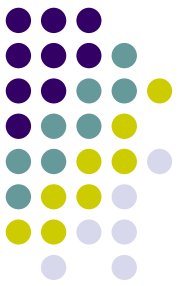
- Evaluated several iterations on tuning set, using answer key to identify relevant sentences

	Acc	Irrelevant			Relevant		
		Rec	Pr	F	Rec	Pr	F
Terrorism							
Iter #1	.84	.93	.89	.91	.41	.55	.47
Iter #2	.84	.90	.91	.90	.54	.51	.53
Iter #3	.82	.85	.92	.89	.63	.46	.53
Disease Outbreaks							
Iter #1	.75	.96	.76	.85	.21	.66	.32
Iter #2	.71	.76	.82	.79	.58	.48	.53
Iter #3	.63	.60	.85	.70	.72	.41	.52

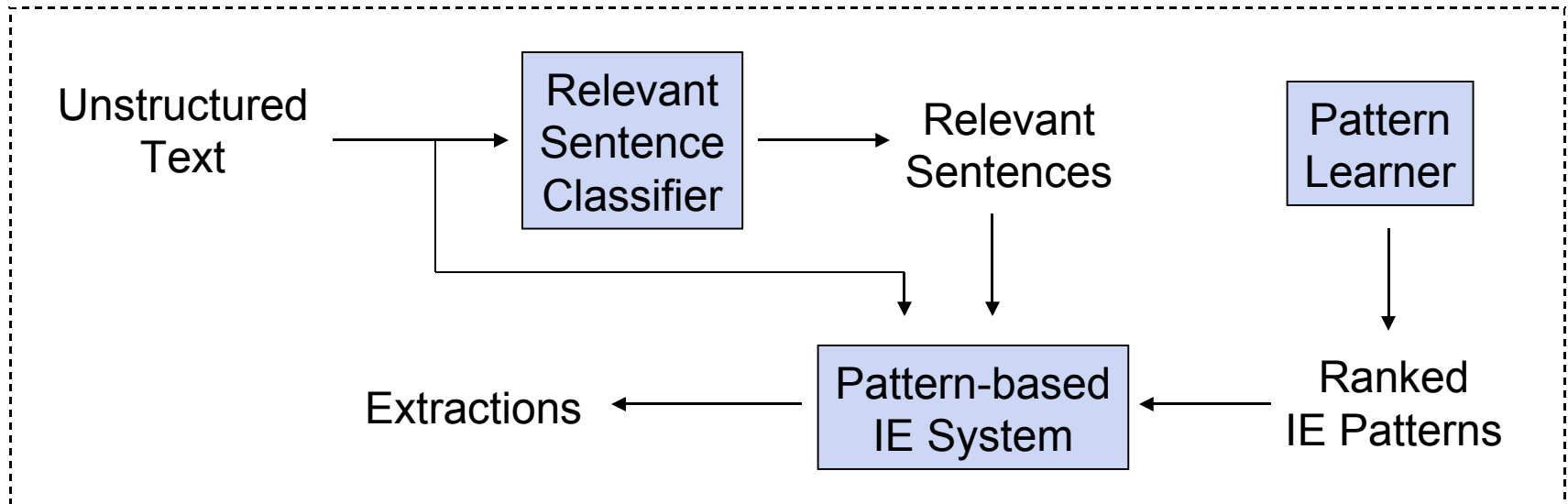


Event Roles Evaluated

- **Terrorism:** *perpetrator individual, perpetrator organization, physical target, human victims, weapon*
- **Disease Outbreaks:** *disease, victims (can be human, animals or plants)*
- AutoSlog-TS (Riloff 1996) was used as our baseline system

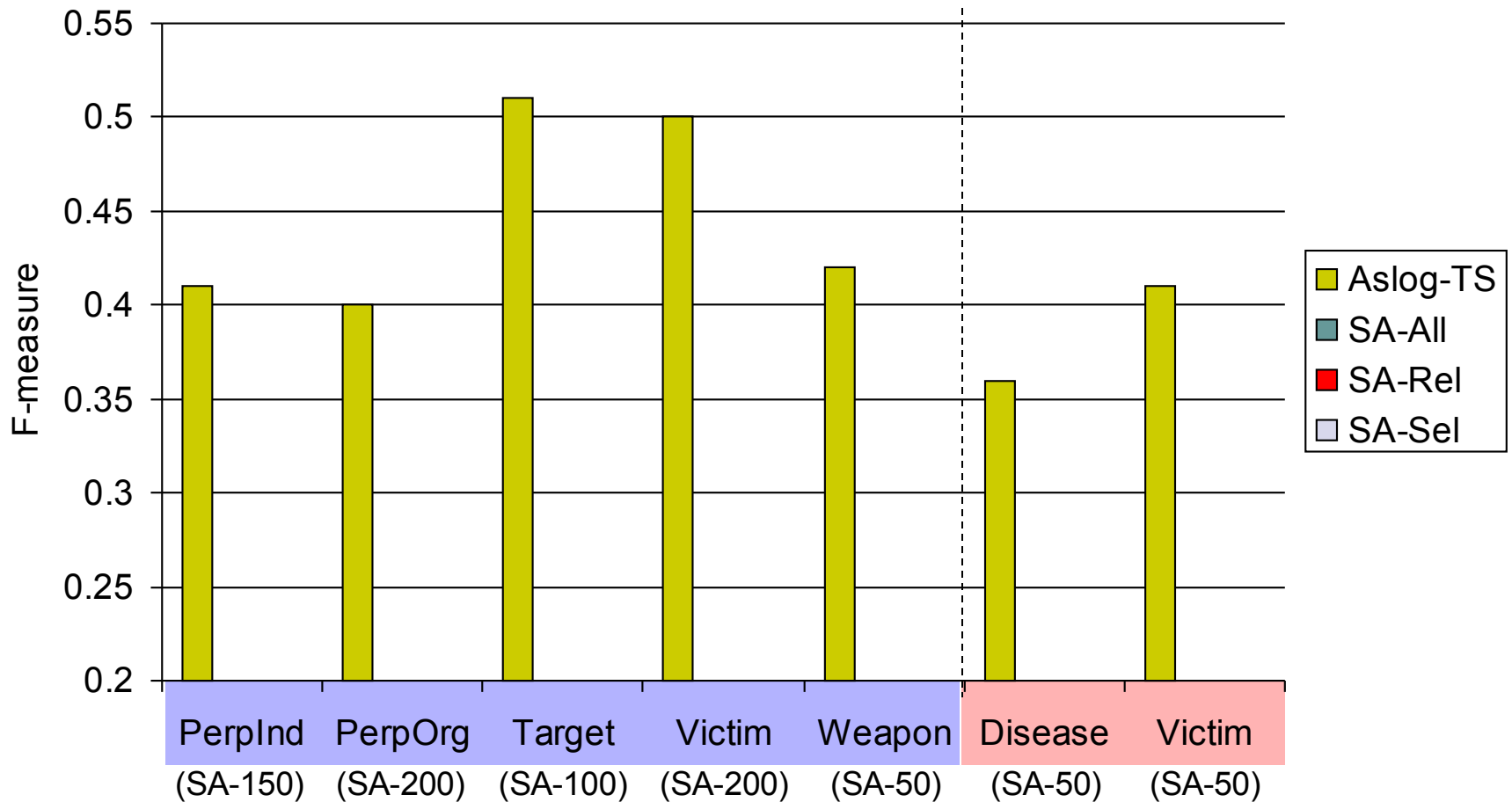
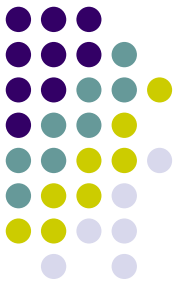


Full IE Evaluation

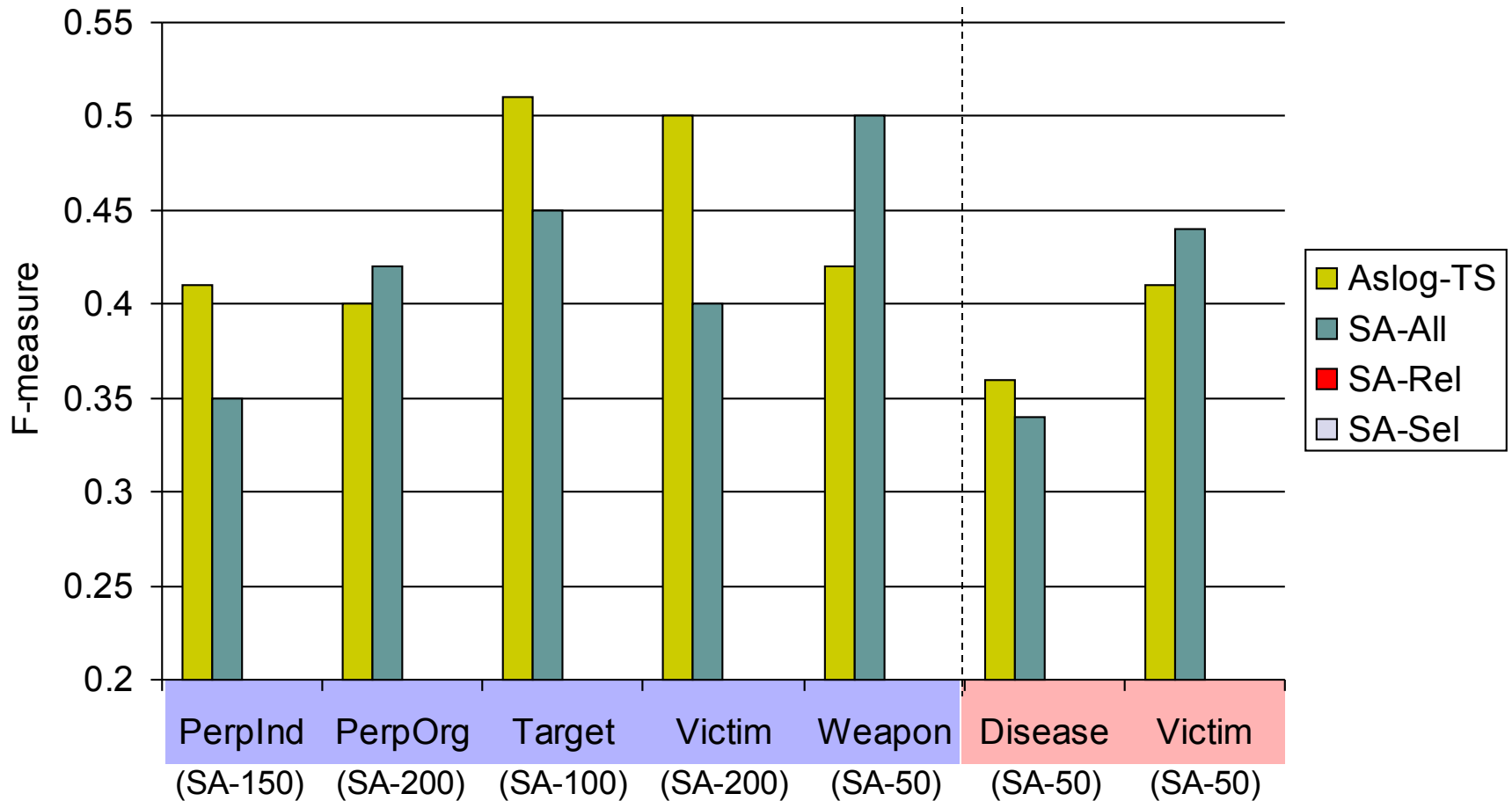
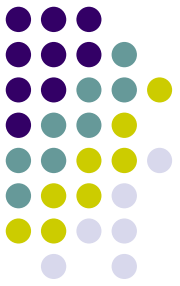


- We evaluated top N patterns ranked by *semantic affinity* for each event role

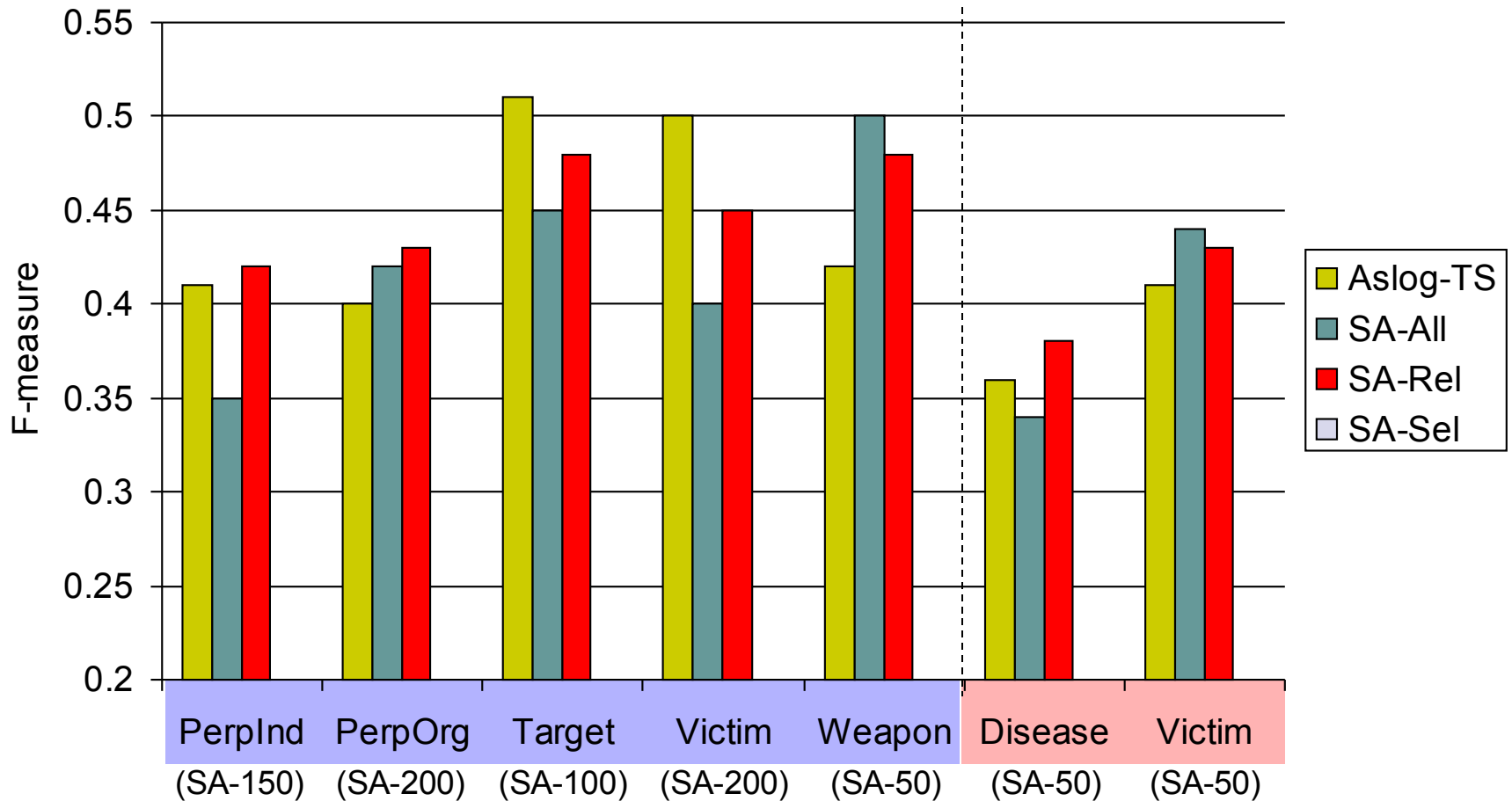
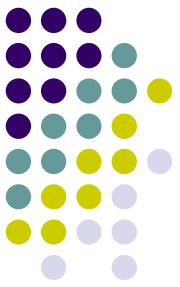
Full IE Evaluation



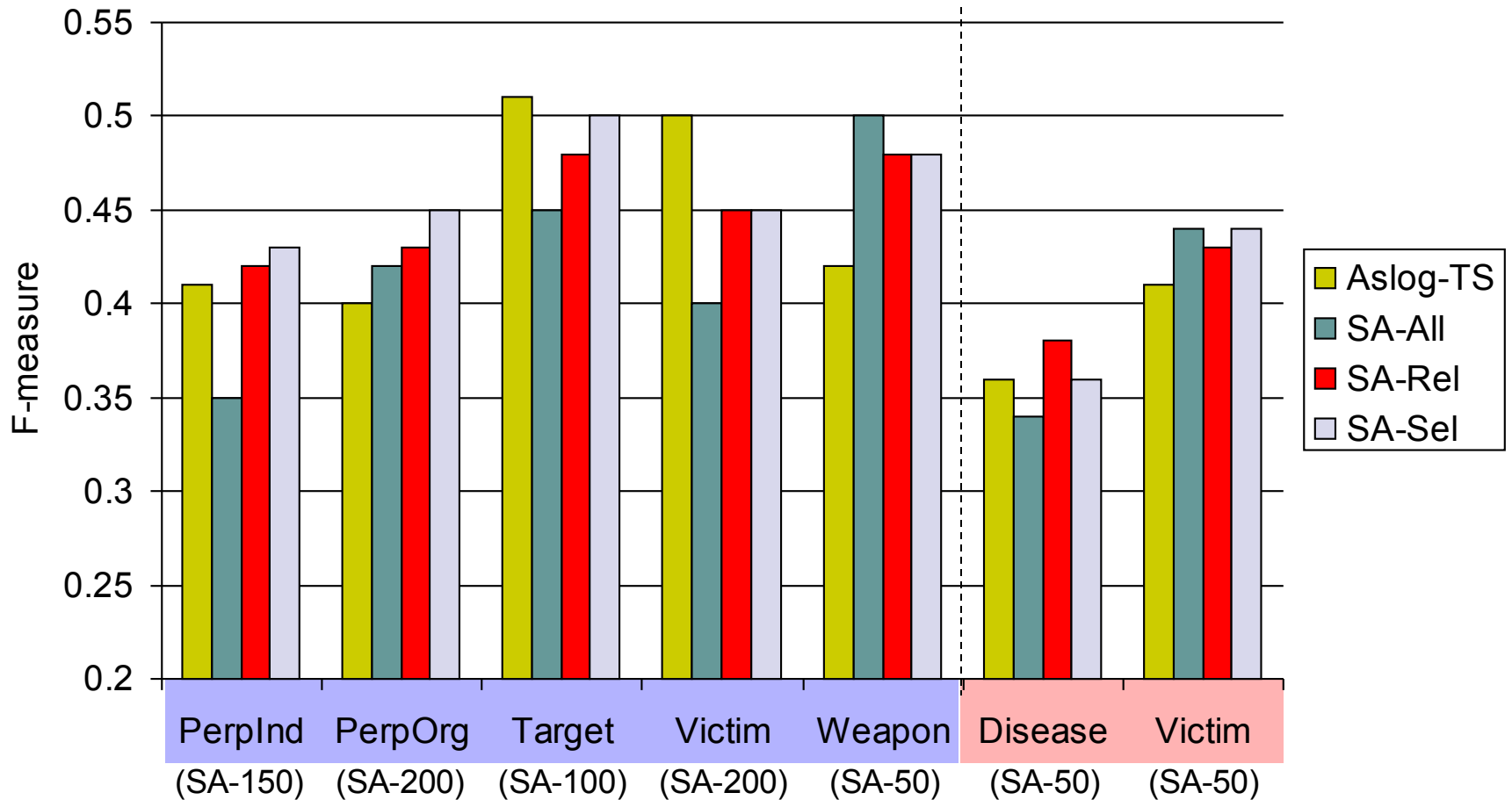
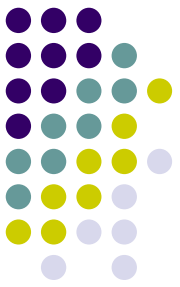
Full IE Evaluation

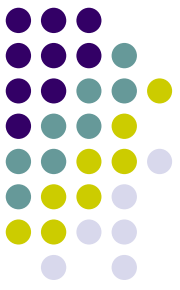


Full IE Evaluation



Full IE Evaluation

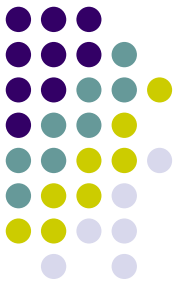




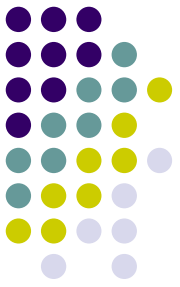
Conclusions

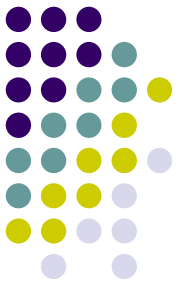
- Reducing the responsibilities of IE patterns:
 - Using a relevant sentence classifier
 - Applying semantically appropriate patterns in relevant regions
- Self-trained sentence classifier uses no sentence annotated data
- Semantic Affinity produces semantically appropriate patterns useful for IE

Questions



Questions



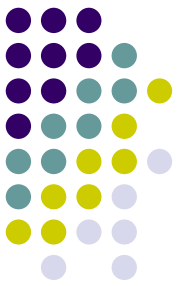


Classifier Effects

- Despite mediocre classifier precision for relevant sentences, it is useful for IE
- **Hypothesis:** It favorably alters the proportion of relevant:irrelevant sentences:

	Before	After
Terrorism Domain	17%	46%
Disease Outbreaks	28%	41%

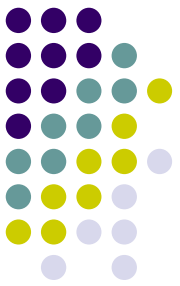
Proportion of relevant sentences in tuning set (before and after applying classifier)



Classifier Effects

- Despite mediocre precision of classifier on relevant sentences, we do not see too much drop in IE recall
- **Hypothesis:** Redundancy in answers/answer key –
 - An answer string appears multiple times in text
 - Answer key contains multiple possible strings for each answer (i.e. *JFK* or *John Kennedy*)

Terrorism domain	1.64 strings per answer
Disease outbreaks	1.77 strings per answer

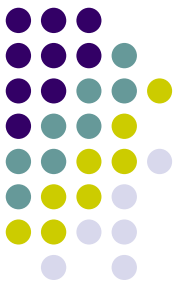


The Extraction Patterns

- Lexico-syntactic patterns based on an underlying shallow parse of the text.
- Patterns extract syntactic constituents.
- Optional selectional restrictions may be applied to the extractions.

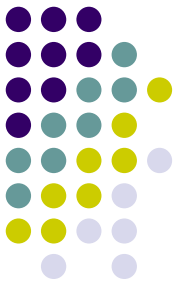
<subject> ActiveVP(murdered)

The pirates brutally murdered their leader.



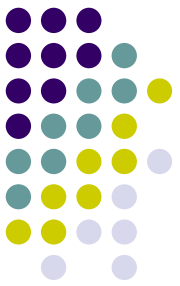
Extraction Pattern Types

ActVP <dobj>	<i>bombed <target></i>
InfVP <dobj>	<i>to kill <victim></i>
ActInfVP <dobj>	<i>planned to bomb <target></i>
PassInfVP <dobj>	<i>was planned to kill <victim></i>
Subj AuxVP <dobj>	<i>fatality is <victim></i>
NP Prep <np>	<i>attack against <target></i>
ActVP Prep <np>	<i>killed quickly with <weapon></i>
PassVP Prep <np>	<i>was killed with <weapon></i>
InfVP Prep <np>	<i>to destroy with <weapon></i>
<possessive> NP	<i><victim>'s murder</i>



Extraction Pattern Types

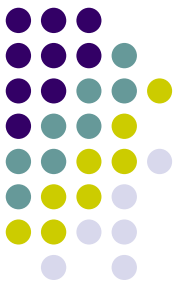
<subj> PassVP	<i><victim> was brutally murdered</i>
<subj> ActVP	<i><perp> murdered</i>
<subj> ActVP Dobj	<i><weapon> caused massive damage</i>
<subj> ActInfVP	<i><perp> tried to kill</i>
<subj> PassInfVP	<i><weapon> was intended to kill</i>
<subj> AuxVP Dobj	<i><victim> was a casualty</i>
<subj> AuxVP Adj	<i><victim> is dead</i>



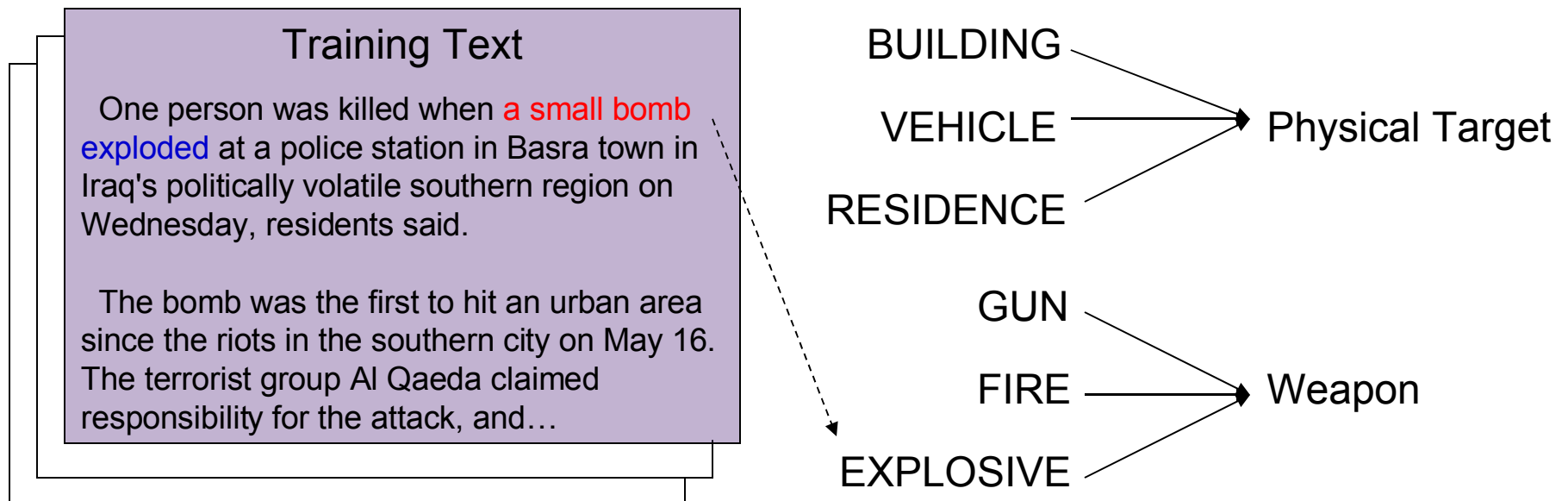
Evaluation

- Seed patterns were automatically generated from the exhaustive set of patterns:
 - Deleted patterns with $\text{freq} \leq 50$
 - Ranked patterns by probability of occurring relevant documents
 - Selected top 20 patterns as seeds (in each domain)

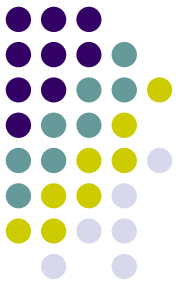
Semantic Affinity Pattern Learner



- For each pattern p and event role r_k , compute frequency $f(p, r_k)$ of the extractions whose semantic categories map to r_k



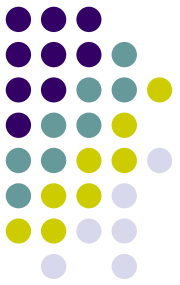
$f(\text{"<subject> EXPLODED"}, \text{Weapon}) += 1$



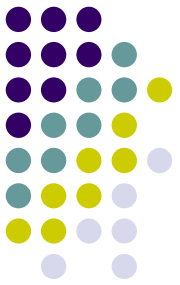
Queries Used for PubMed

- "Immunization" [MeSH] NOT "Disease Outbreaks" [MeSH] AND hasabstract [text] AND English [Lang] AND "2003/05/02 12.42" [EDAT] : "2005/05/01 12.42" [EDAT]

Evaluation

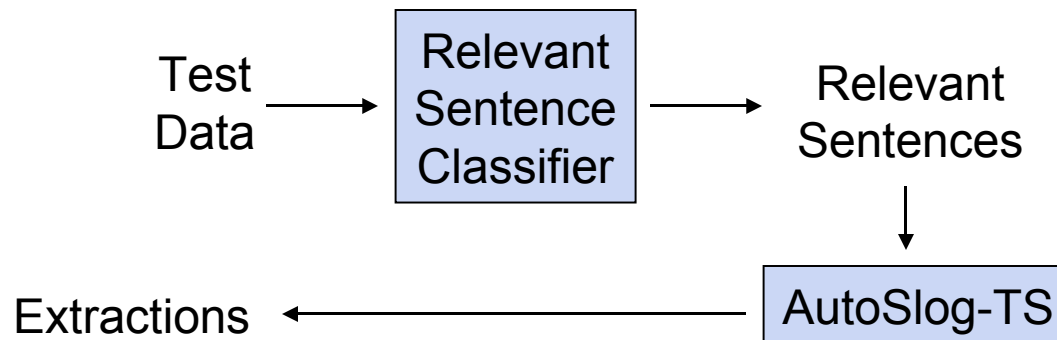


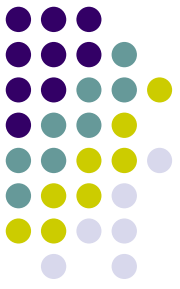
- In disease outbreaks domain, applied a higher cutoff (1.0) to SVM output



AutoSlog-TS + Classifier

- Applied the relevant sentence classifier in combination with an existing IE system





AutoSlog-TS + Classifier

- Classifier improves precision of AutoSlog-TS
- In most cases, we see improvement overall improvement (F-score)