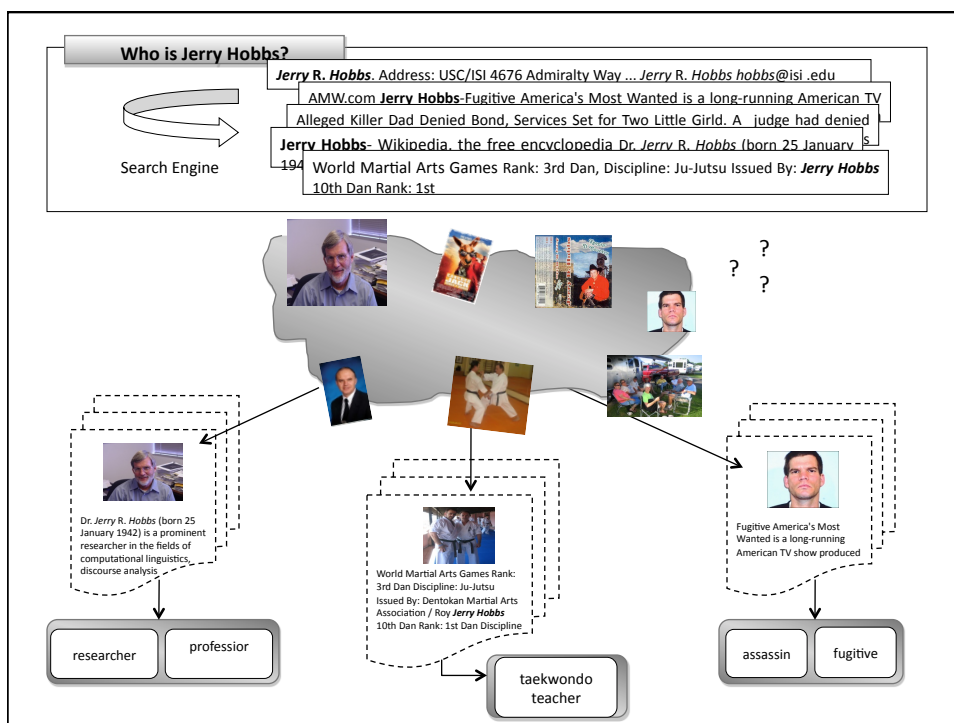


# CS544: Information Extraction, Named Entity Recognition and Classification

March 11, 2010

Zornitsa Kozareva  
USC/ISI  
Marina del Rey, CA  
kozareva@isi.edu  
www.isi.edu/~kozareva



## Named Entity Recognition and Classification

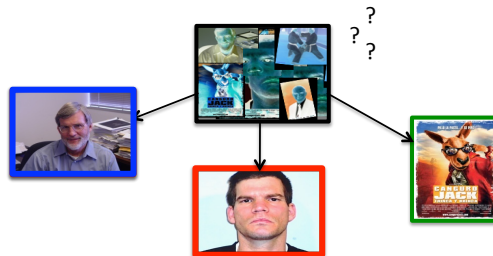
<PER>Prof. Jerry Hobbs</PER> taught CS544 during <DATE>February 2010</DATE>.  
 <PER>Jerry Hobbs</PER> killed his daughter in <LOC>Ohio</LOC>.  
 <ORG>Hobbs corporation</ORG> bought <ORG>FbK</ORG>.

- Identify mentions in text and classify them into a predefined set of categories of interest:
  - Person Names: Prof. Jerry Hobbs, Jerry Hobbs
  - Organizations: Hobbs corporation, FbK
  - Locations: Ohio
  - Date and time expressions: February 2010
  - E-mail: mkg@gmail.com
  - Web address: www.usc.edu
  - Names of drugs: paracetamol
  - Names of ships: Queen Marry
  - Bibliographic references:
  - ...

## Named Entity Discrimination

Prof. Jerry Hobbs taught CS544 during February 2010.  
 Jerry Hobbs killed his daughter in Ohio.  
 Hobbs corporation bought FbK.

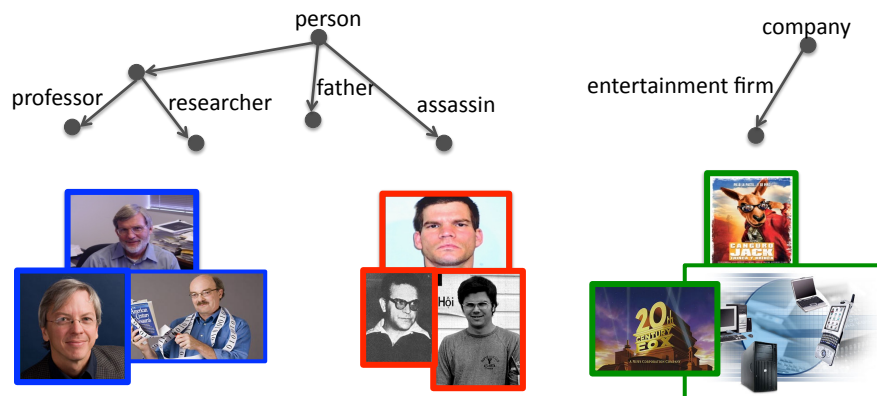
- Discover the underlying meaning of a proper name



- The number of clusters and their meaning is unknown

## Hypernym and Hyponym Learning

- Given an instance (e.g. Jerry Hobbs), learn automatically from the Web corresponding hypernyms and hyponyms



## Information Extraction

## What is “Information Extraction”?

- Goal: identify specific pieces of information from the content of unstructured or semi-structured textual documents.
- Input:
  - scenario of extraction (template schema to be filled)
  - document collection
- Output:
  - a set of instantiated templates

## MUC

- Message Understanding Conference (MUC) was an annual competition for IE systems funded by DARPA
  - MUC-1 (1987) and MUC-2 (1989)
    - Messages about naval operations
  - MUC-3 (1991) and MUC-4 (1992)
    - News articles about terrorist attacks
  - MUC-5 (1993)
    - News articles about joint ventures and microelectronics
  - MUC-6 (1995)
    - News articles about management changes
  - MUC-7 (1997)
    - News articles about space vehicle and missile launches

A **bomb** went off this morning near a **power tower** in **San Salvador** leaving a large part of the city without energy, but **no casualties** have been reported.











According to unofficial sources, the bomb-allegedly detonated by **urban guerrilla commandos** **blew up** a power tower in the northwestern part of San Salvador at 0650.

Incident type:	bombing
Date:	March 11, 2010
Location:	San Salvador (city)
Perpetrator:	urban guerrilla commandos
Physical target:	power tower
Human target:	-
Effect on physical target:	destroyed
Effect on human target:	no injury or death
Instrument:	bomb


## Today

Google **squared** labs    <http://www.google.com/squared>

Jerry Hobbs

Item Name	Image	Description	Date Of Birth	Nationality	Date Of Death	Died
<input checked="" type="checkbox"/> Franz Boas		<b>Franz Boas</b> (July 9, 1858 – December 21, 1942) was a German American anthropologist and a pioneer of modern anthropology.	July 9, 1858	American	December 21, 1942	22-Dec-1942
<input checked="" type="checkbox"/> Edward Sapir		<b>Edward Sapir</b> (pronounced /sə'pɪrɪ/), (January 26, 1884 – February 4, 1939) was a German-born American linguist.	January 26, 1884	American	February 4, 1939	4-Feb-1939
<input checked="" type="checkbox"/> Leonard Bloomfield		<b>Leonard Bloomfield</b> (April 1, 1887 – April 18, 1949) was an American linguist who led the development of structural linguistics in the United States.	April 1, 1887	American	April 18, 1949	April 18, 1949
<input checked="" type="checkbox"/> Jonathan Edwards		<b>Jonathan Edwards</b> – <b>Jonathan Edwards</b> William Tyndale: A biography - David Daniell.	October 5, 1703	American	March 22, 1758	March 22, 1758
<input checked="" type="checkbox"/> Benjamin Lee Whorf		<b>Benjamin Lee Whorf</b> (April 24, 1897 in Winthrop, Massachusetts – July 26, 1941) was an American linguist. Whorf is widely known for his theory of linguistic relativity.	April 24, 1897	American	July 26, 1941	July 26, 1941
<input checked="" type="checkbox"/> Donna Jo Napoli		<b>Donna Jo Napoli</b> writes for all ages, from picture books through young adult books. Her awards include the Lewis Foundation Grant.	02/28/1948	American		
<input checked="" type="checkbox"/> Joseph Greenberg		<b>Joseph Harold Greenberg</b> (May 28, 1915 – May 7, 2001) was a prominent and controversial American linguist, originally known as <b>Mark Baker</b> .	May 28, 1915	American	May 7, 2001	May 7, 2001
<input checked="" type="checkbox"/> Mark Baker		<b>Mark Baker</b> , OMI, Inc. City of Rio Rancho, 2003, Kenny Lewis Fort Lupton Project, OMI, Inc. Mexico, 2004.	Jun 17, 1954	United States		1 possible value
<input checked="" type="checkbox"/> Samuel Martin		<b>Martin, Samuel</b> (1976). A Reference Grammar of Japanese. Yale University Press, 47. Nadathur, Gopal, and Ischi Amund K.	1714	2 possible values	4 possible values	4 possible values
<input checked="" type="checkbox"/> Hans Kurath		<b>Hans Kurath</b> (13 December 1891–2 January 1992) was an American linguist of Austrian origin.	1891-12-13			1 possible value

# Today

 **Video Games & Cheats**  
 Submitted by [factual...](#) on October 11, 2009, [more](#) <http://www.factual.com/>

Data Visualizations Statistics Developer

[Add Row](#) [File](#) [Display](#) [Filters](#) [Fields](#) [Improve](#)

Video Game	Platform	Publisher	Developer	Genre	Theme
3 on 3 NHL Arcade	Xbox 360	EA Sports	EA Canada	Sports	Hockey
4x4 Evo 2	Xbox	2K Games	Terminal Reality, Inc.	Sports; Driving/Racing	
24: The Game	PlayStation 2	2K Games	SCE Studio Cambridge	Action	Espionage
25 to Life	PlayStation 2	Eidos Interactive	Avalanche Software LLC; Ritt	Action; Shooter	Crime
50 Cent: Bulletproof	Xbox	Vivendi Universal	Genuine Games Ltd.	Action; Shooter	Crime
1942: Joint Strike	Xbox 360	Backbone Entertainment		Coin-Op Classics	
Abuse	PC	Alive Mediasoft; Bungie Stud	Crack dot Com	Action; Shooter	Sci-Fi; Horror
AC/DC Live: Rock Band Track Pac	Xbox 360	Electronic Arts			
Ace Combat 5: The Unsung War	PlayStation 2	Namco	Namco Bandai Games Inc.	Action; Simulation; Flight Sim	Modern Military
Ace Combat X: Skies of Deception	PlayStation Portabl	Namco Games	Namco Bandai Games Inc.	Action; Simulation; Flight Sim	Modern Military
Ace Combat Zero: The Belkan Wa	PlayStation 2	Namco	Namco Bandai Games Inc.	Action; Simulation; Flight Sim	Modern Military
Activision Anthology	PlayStation 2	Activision	Activision	Action; Strategy; Sports; Drivi	
Act of War: Direct Action	PC	GFI Russia; Atari, Inc.	Eugen Systems	Strategy; Real-Time Strategy	Modern Military
Advance Guardian Heroes	Game Boy Advance	UBI Soft	Treasure	Action; Fighting	
Advance Wars: Days of Ruin	Nintendo DS	Nintendo	Intelligent Systems Co., Ltd.	Strategy	Post-Apocalyp
Advance Wars: Dual Strike	Nintendo DS	Nintendo	Intelligent Systems Co., Ltd.	Strategy	Modern Military
Advent Rising	PC	Majesco Entertainment	GlyphX Games, LLC	Action; Shooter; Adventure	Sci-Fi
Advent Rising	Xbox	Majesco Entertainment	GlyphX Games, LLC	Action; Shooter; Adventure	Sci-Fi
Aegis Wing	Xbox 360	Microsoft Game Studios			
Aeon Flux	Xbox	Majesco Sales Inc.			
Afro Samurai	Xbox 360	Namco	Namco Bandai Games Inc.	Platformer; Action	Martial Arts; Ani
Age of Booty	Xbox 360	Certain Affinity			
Age of Empires III	PC	Microsoft Game Studios; Mac	Ensemble Studios	Strategy; Real-Time Strategy;	Alternate Histor
Age of Empires: The Age of Kings	Nintendo DS	Majesco Sales Inc.	Backbone Entertainment	Strategy	Fantasy

## Other Applications

- Job postings
- Seminar announcements
- Conference call for papers
- Company information
- Apartment rental adds
- Social event announcements
- ...

Example is adapted from Raymond Mooney

Subject: US-TN-SOFTWARE PROGRAMMER  
 Date: 17 Nov 2009 17:37:29 GMT  
 Organization: Reference.Com Posting Service  
 Message-ID: <56nigp\$mrs@bilbo.reference.com>

#### SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay.

Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:  
 Kim Anderson  
 AdNET  
 (901) 458-2888 fax  
 kimander@memphisonline.com

#### computer\_science\_job\_template

id:  
 title:  
 salary:  
 company:  
 recruiter:  
 state:  
 city:  
 country:  
 language:  
 platform:  
 application:  
 area:  
 req\_years\_experience:  
 desired\_years\_experience:  
 req\_degree:  
 desired\_degree:  
 post\_date:

Example is adapted from Raymond Mooney

Subject: **US-TN-SOFTWARE PROGRAMMER**  
 Date: **17 Nov 2009** 17:37:29 GMT  
 Organization: Reference.Com Posting Service  
 Message-ID: <**56nigp\$mrs@bilbo.reference.com**>

#### SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based **Voice Mail** systems. Experienced in **C** Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay.

Prefer **5** years or more experience with **PC** Based **Voice Mail**, but will consider as little as **2** years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is **DOS**. May go to **OS-2** or **UNIX** in future.

Please reply to:  
 Kim Anderson  
 AdNET  
 (901) 458-2888 fax  
 kimander@memphisonline.com

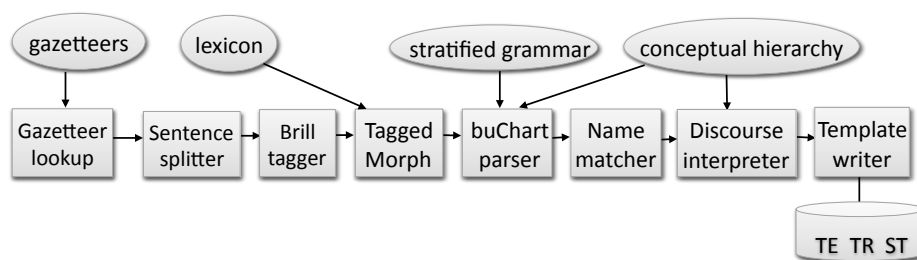
#### computer\_science\_job\_template

id: **56nigp\$mrs@bilbo.reference.com**  
 title: **SOFTWARE PROGRAMMER**  
 salary:  
 company:  
 recruiter:  
 state: **TN**  
 city:  
 country: **US**  
 language: **C**  
 platform: **PC \ DOS \ OS-2 \ UNIX**  
 application:  
 area: **Voice Mail**  
 req\_years\_experience: **2**  
 desired\_years\_experience: **5**  
 req\_degree:  
 desired\_degree:  
 post\_date: **17 Nov 2009**

## Two general approaches to IE

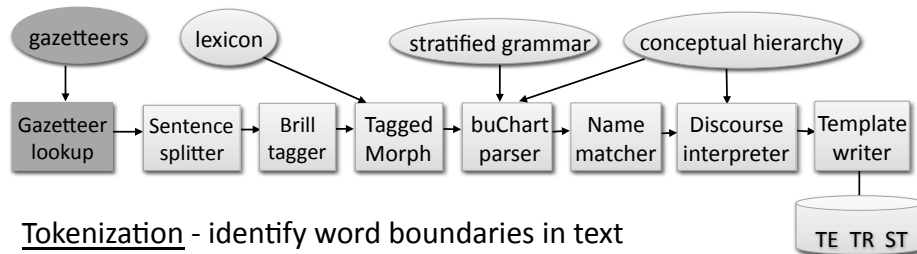
- Pattern-based systems use patterns or rules that are applied to text.
- Sequence tagging models classify individual tokens as to whether or not they should be extracted.

### LaSIE Information Extraction System





## LaSIE Information Extraction System



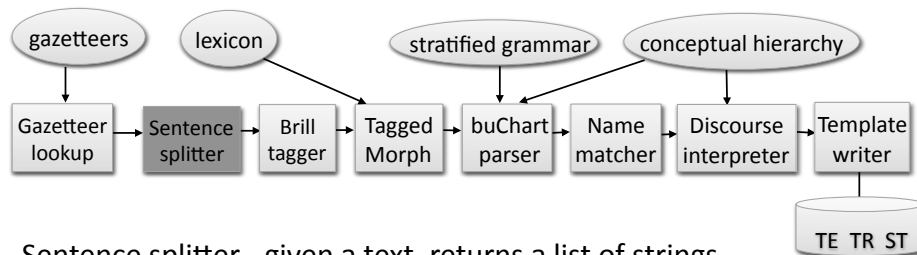
Tokenization - identify word boundaries in text

- white spaces indicate token boundaries
- full stops indicate sentences boundaries  
(not always true e.g. 1. September)

Gazetteer Lookup – recognize phrases and keywords related to named entities which were previously stored in its lists (gazetteers)

- advantage – simple, fast, language independent as one just has to create the lists
- disadvantage – collection and maintenance is time consuming, cannot deal with name variants, cannot resolve ambiguity

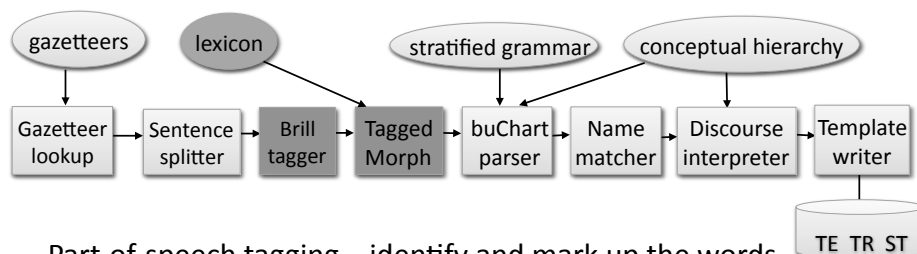
## LaSIE Information Extraction System



Sentence splitter - given a text, returns a list of strings where each element is a sentence.

- uses a set of rules like the occurrences of “.”, “?” and “!” are indicators of sentence delimiters, but the occurrence of “.” in “B. Clinton” or “U.S.” does not have this role

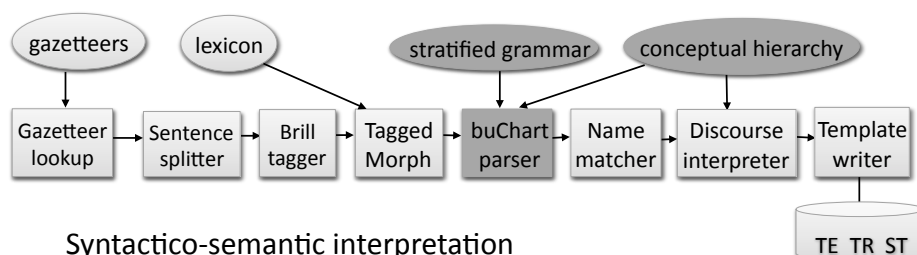
## LaSIE Information Extraction System



Part-of-speech tagging – identify and mark up the words in a text with the corresponding part of speech such as noun, verb, adjective, adverb etc.

According\_to-adv unofficial-adj source[s]-n , the-det bomb-n allegedly-adv  
 detonate[ed]-v by-prep urban-adj guerrilla-n commando[s]-n blow\_up-v a-  
 det power\_tower-n in-prep the-det northwestern-adj part-n of-prep  
 San Salvador-loc at-prep 0650-time

## LaSIE Information Extraction System



### Syntactico-semantic interpretation

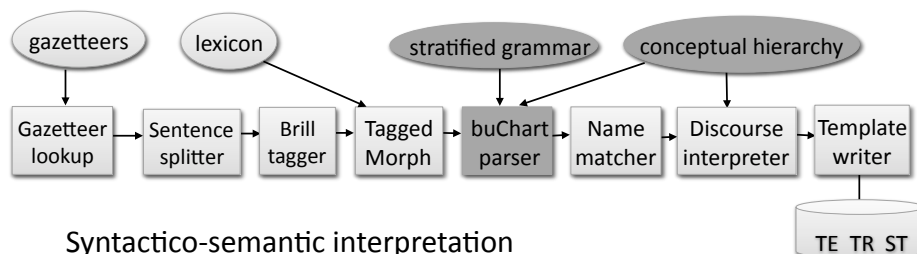
- bottom-up chart parser
- cascade of NERC grammars (eg. aircraft, person, money, time)

According\_to-adv unofficial-adj source[s]-n , the-det bomb-n allegedly-adv  
 detonate[ed]-v by-prep urban-adj guerrilla-n commando[s]-n blow\_up-v a-  
 det power\_tower-n in-prep the-det northwestern part of San Salvador-loc  
 at-prep 0650-time

NE2

NE1

## LaSIE Information Extraction System

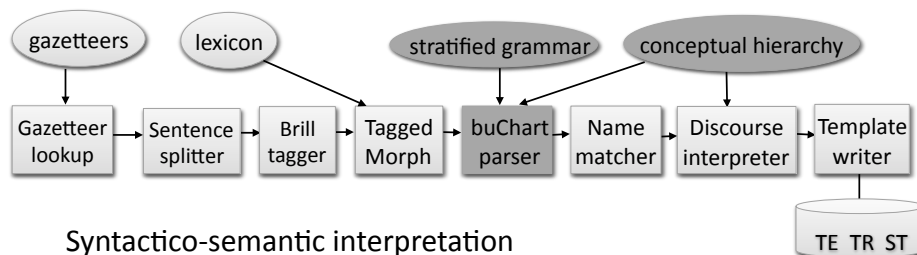


### Syntactico-semantic interpretation

- bottom-up chart parser
- cascade of NERC grammars (eg. aircraft, person, money, time)
- cascade of partial grammars (NPs, PPs, complex NP, VPs, complex VPs, RelClauses, Sentence)

S(According\_to-adv NP(unofficial-adj source[s]-n) , NP(the-det bomb-n) allegedly-adv VP(detonate[ed]-v) PP(by-prep NP(urban-adj guerrilla-n commando[s]-n)) - VP(blow\_up-v) NP(a-det power\_tower-n) PP(in-prep NP (the-det NE1-loc)) PP(at-prep NP(NE2-time)))

## LaSIE Information Extraction System

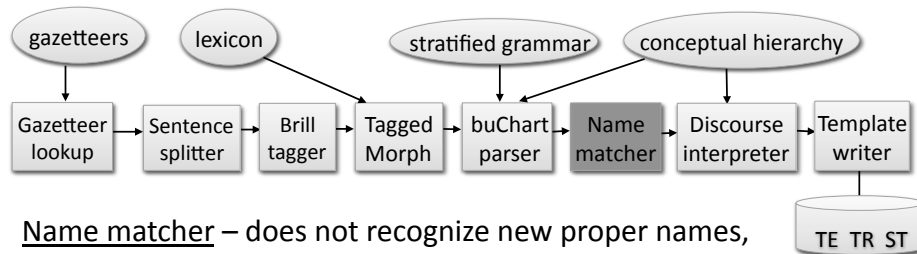


### Syntactico-semantic interpretation

- bottom-up chart parser
- cascade of NERC grammars (eg. aircraft, person, money, time)
- cascade of partial grammars (NPs, PPs, complex NP, VPs, complex VPs, RelClauses, Sentence)
- logic form

Event(E1), detonate(E1,Y,X), urban\_guerrilla\_commando(X), bomb(Y),  
 Event(E2), blow\_up(E2,Y,Z), power\_tower(Z), location\_of(Z,NE1), time\_of(E2,NE2)

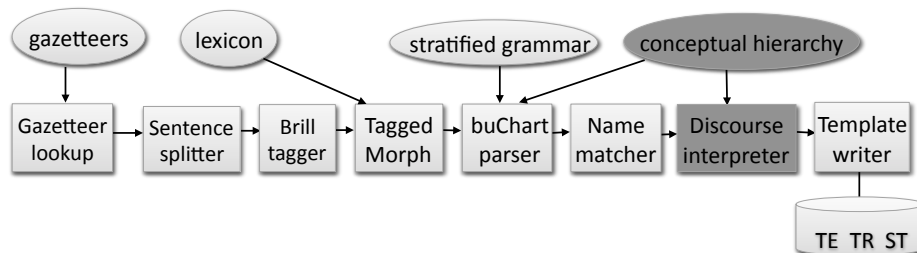
## LaSIE Information Extraction System



Name matcher – does not recognize new proper names, just adds identity relations between those found by the parser

- first token of the name matches the second name  
*"Pepsi Cola"* equals *"Pepsi"*
- one of the names is an acronym of the other  
*"ISI"* is equivalent to *"Information Sciences Institute"*
- one name is a reversal of the other  
*"Defense Department"* equals *"Department of Defense"*
- one name consists of concatenated contractions of the other  
*"Pan America"* equals *"Pan Am"*

## LaSIE Information Extraction System

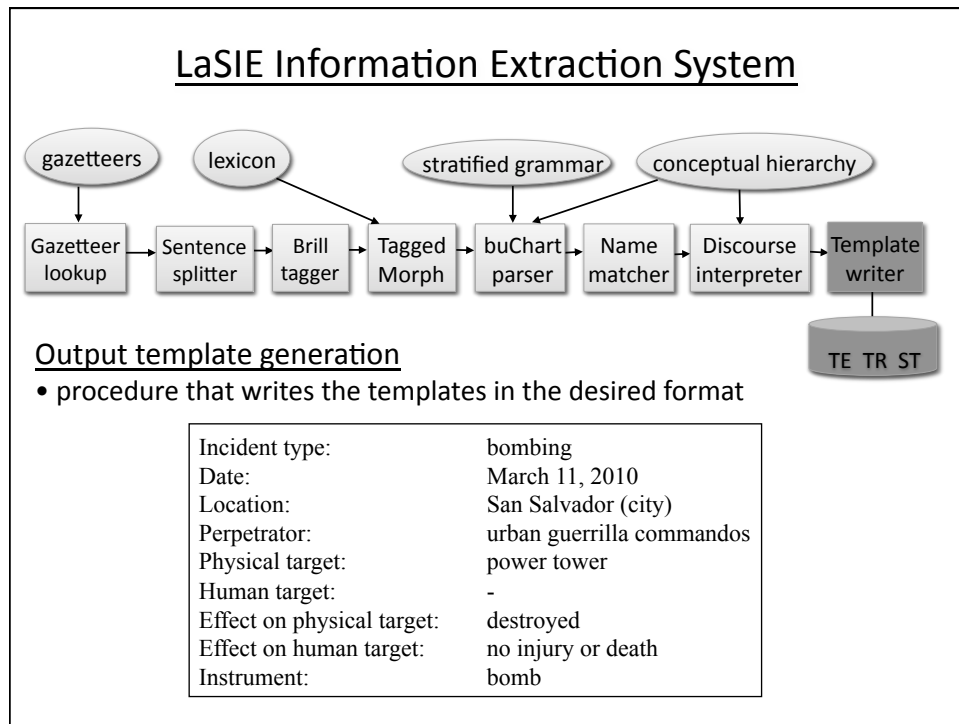


Discourse interpreter – translates the semantic representations produced by the parser into

- representation of instances, their ontological classes and attributes
- coreference resolution

Event(E1), detonate(E1,Y,X), urban\_guerrilla\_comando(X), bomb(Y),  
 Event(E2), blow\_up(E2,Y,Z), power\_tower(Z), location\_of(Z,NE1), time\_of(E2,NE2)

Diagram illustrating semantic relationships:  
 - bomb(Y) implies bombing event  
 - blow\_up(E2,Y,Z) isa destroy  
 - location\_of(Z,NE1) implies location of event



## How well does this work<sup>1</sup>?

- Evaluate system's performance on independent manually-annotated test gold data which was not used during system development
- IE systems are typically evaluated in terms of Precision (P) and Recall (R)

$$P = \frac{\text{correctly extracted facts}}{\text{extracted facts}}$$

$$R = \frac{\text{correctly extracted facts}}{\text{correct facts}}$$

$$F_1 = \frac{2PR}{P + R}$$

<sup>1</sup>[http://www-nlpir.nist.gov/related\\_projects/muc/proceedings/st\\_score\\_report.html](http://www-nlpir.nist.gov/related_projects/muc/proceedings/st_score_report.html)

## LaSIE in MUC-6

Task	Precision	Recall
Named Entity	.94	.84
Co-reference resolution	.71	.51
Template Elements	.74	.66
Scenario Templates	.73	.37

## LaSIE Named Entity

- Results for the Named Entity task over 30 texts
- Each setting indicates the contribution of LaSIE's components

No.	Setting	Precision	Recall
1	Gazetteer Look Up	.74	.37
2	1 + Parsing	.93	.80
3	2 + Name matching	.93	.88
4	3 + Discourse interpretation	.93	.89



GATE — General Architecture for Text Engineering

**ANNIE Demo**<http://services.gate.ac.uk/annie/index.jsp>

ANNIE is one of many Information Extraction systems that have been developed using GATE. It uses finite state algorithms and the JAPE language. This demo shows ANNIE recognising entities in texts.

Note: this demo uses a default set of components and IE resources; your mileage may vary! Also, complex HTML structures may prevent the system from being able to analyse the text they contain. The system does name recognition; see the [IE User Guide](#) for details of other forms of IE, and issues of domain-specificity and porting. [Contact us](#) about our [cross-domain, multi-genre](#) systems.

To use ANNIE, enter a URL in the box below. Select the types of entities that you would like to mark. GATE will then retrieve the document and extract the required information. This process may take a few seconds.

Enter a URL: 

- ☒ Person
- ☒ Location
- ☒ Organization
- ☒ Date
- ☒ Address
- ☒ Money
- ☒ Percent

Run ANNIE

**California Prius incident probed; GM offers criticized**BY JUSTIN HYDE  
FREE PRESS WASHINGTON STAFF

[Comments \(3\)](#)
[Recommend \(2\)](#)
[Print](#)
[E-mail](#)
[Letter to the editor](#)
[Share](#)

WASHINGTON -- As Toyota sought to contain the fallout from a California sudden-acceleration case caught on camera, its dealers accused General Motors of offering predatory incentives using federal money.

The Japanese automaker and the National Highway Traffic Safety Administration dispatched investigators to San Diego to analyze the 2008 Toyota Prius belonging to James Sikes, 61. Sikes called the California Highway Patrol on Monday evening reporting that his Prius was accelerating out of his control, hitting speeds of up to 94 m.p.h.

"I pushed the gas pedal to pass a car and it did something kind of funny," Sikes said at a news conference. "It jumped and it just stuck there."

An officer pulled alongside the Prius, and over a loudspeaker told Sikes to pull the emergency brake and press the regular brake hard. Sikes was able to slow the car and shut it off after 20 minutes.

The incident happened a few miles from the site of the Santee crash last August that spurred Toyota to recall 5.6 million vehicles for mechanical defects that could lead to sudden acceleration -- including Sikes' Prius, which was covered by the floor-mat recall.

But Sikes said he had taken his Prius to his Toyota dealer and was told it wasn't covered. Toyota said in a statement that the repairs for the Prius had not yet been sent to dealers. The automaker had told owners to remove the driver's-side floor mats until the repairs could be made, but Sikes had left the floor mat in his vehicle.

Toyota reiterated Tuesday that the Prius was under recall. It had said last year that it could take months for all of the recalled vehicles to be fixed.

In another sign of the pressure facing Toyota, its national dealer panel accused GM of using "taxpayer dollars to fund ... a nationwide predatory advertising campaign." Shortly after Toyota began its recalls last month, GM began incentives for Toyota owners that now include zero-percent financing and up to \$1,000 cash back.

"It is outrageous that GM is using our taxpayer dollars against us, making me and other Toyota dealers pay to undermine our own businesses," said Paul Atkinson, president of Toyota's U.S. dealer council.

GM's move was matched by other automakers. Last week, Toyota launched an incentive campaign of its own after a 9% drop in February sales.

WASHINGTON -- As Toyota sought to contain the fallout from a California sudden-acceleration case caught on camera, its dealers accused General Motors of offering predatory incentives using federal money.

&gt;

The Japanese automaker and the National Highway Traffic Safety Administration dispatched investigators to San Diego to analyze the 2008 Toyota Prius belonging to James Sikes, 61. Sikes called the California Highway Patrol on Monday evening reporting that his Prius was accelerating out of his control, hitting speeds of up to 94 m.p.h.

"I pushed the gas pedal to pass a car and it did something kind of funny," Sikes said at a news conference. "It jumped and it just stuck there."

An officer pulled alongside the Prius, and over a loudspeaker told Sikes to pull the emergency brake and press the regular brake hard. Sikes was able to slow the car and shut it off after 20 minutes.

The incident happened a few miles from the site of the Santee crash last August that spurred Toyota to recall 5.6 million vehicles for mechanical defects that could lead to sudden acceleration -- including Sikes' Prius, which was covered by the floor-mat recall.

But Sikes said he had taken his Prius to his Toyota dealer and was told it wasn't covered. Toyota said in a statement that the repairs for the Prius had not yet been sent to dealers. The automaker had told owners to remove the driver's-side floor mats until the repairs could be made, but Sikes had left the floor mat in his vehicle.

Toyota reiterated Tuesday that the Prius was under recall. It had said last year that it could take months for all of the recalled vehicles to be fixed.

In another sign of the pressure facing Toyota, its national dealer panel accused GM of using "taxpayer dollars to fund ... a nationwide predatory advertising campaign." Shortly after Toyota began its recalls last month, GM began incentives for Toyota owners that now include zero-percent financing and up to \$1,000 cash back.

"It is outrageous that GM is using our taxpayer dollars against us, making me and other Toyota dealers pay to undermine our own businesses," said Paul Atkinson, president of Toyota's U.S. dealer council.

GM's move was matched by other automakers. Last week, Toyota launched an incentive campaign of its own after a 9% drop in February sales.

"We understand why Toyota dealers would be frustrated, but they know better," said GM spokesman Kerry

## Rule-based IE: Pros and Cons

- PROs:
  - clearly understood technology
  - hand-written rules are relatively precise
  - people can write rules with a reasonable amount of training
- CONs:
  - rules need to be written by hand
  - requires experienced grammar developers
  - difficult to port to a different domain

## Can we automatically learn IE?

- In the mid-1990's supervised IE systems were created.
- Supervised learning requires annotated training data.
- Trade-off: annotating texts vs. manual knowledge engineering
  - weeks vs. months
  - domain experts vs. computational linguists



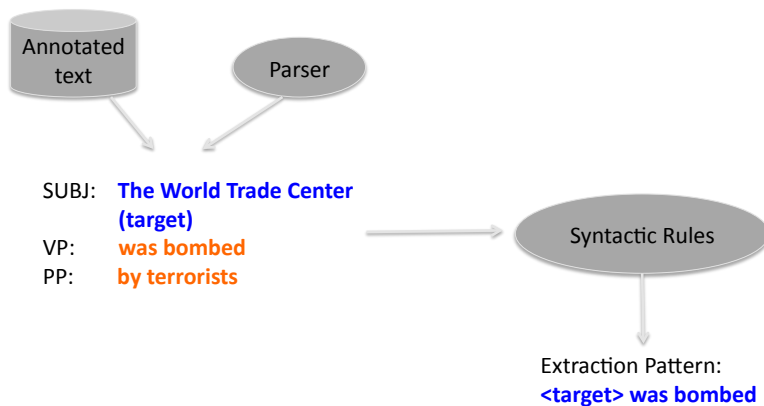
## Annotating Texts for IE

Alleged guerilla urban commandos <sup>perpetrator</sup> launched  
 two highpower bombs <sup>weapon</sup> against a car dealership <sup>target</sup>  
 in downtown San Salvador <sup>location</sup> this morning <sup>date</sup> . A  
 police report said that the attack set the building <sup>damage</sup>  
on fire <sup>injury</sup> but did not result any casualties .

## Pattern Learning Algorithms

- A variety of different pattern/rule representations have been developed, but very commonly:
  - IE systems learn patterns by beginning with highly specialized patterns and iteratively generalizing them.
  - rule-learning stops when a set of patterns has been generated to sufficiently “cover” the training examples.

## AutoSlog [Riloff 1993]



Examples of learned extraction patterns by AutoSlog:

<subject> active-vp	<perpetrator> bombed
<subject> active-vp dobj	<perpetrator> threw dynamite
<subject> active-vp infinitive	<perpetrator> tried to kill
passive-vp infinitive <dobj>	was hired to kill <victim>
subject auxiliary <dobj>	fatality was <victim>
passive-vp prep <np>	was killed by <perpetrator>