Event Extraction

Identifying descriptions of complex events and extracting the role fillers associated with each incident.

EVENT	ROLES
Terrorist act	perpetrator, victims, target
Natural disaster	natural force, victims, damage
Plane crash	vehicle, victims, cause
Management changes	person leaving, position, successor, organization
Disease outbreaks	disease, victims, symptoms, containment measures

Event Roles vs. Named Entity Recognition

Named Entity Recognition = identifying types of entities

Event Roles = identifying entities that play a specific role with respect to an event

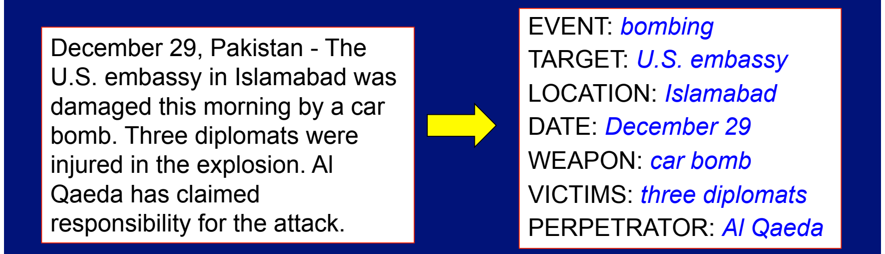
Paul Nelson killed John Smith.
Paul Nelson was killed by John Smith.

IBM purchased Microsoft.
IBM was purchased by Microsoft.
IBM was purchased on Tuesday by Microsoft.

Event Extraction

INPUT: document

OUTPUT: filled event template



December 29, Pakistan - The U.S. embassy in Islamabad was damaged this morning by a car bomb. Three diplomats were injured in the explosion. Al Qaeda has claimed responsibility for the attack.

EVENT: *bombing*
TARGET: *U.S. embassy*
LOCATION: *Islamabad*
DATE: *December 29*
WEAPON: *car bomb*
VICTIMS: *three diplomats*
PERPETRATOR: *Al Qaeda*

Event Template for Terrorist Acts

Date	<date>
Location	<location>
Event type	<set fill>
Weapon	<string list>
Perpetrator individual	<string list>
Perpetrator organization	<string list>
Physical target	<string list>
Physical target effect	<set fill>
Human target	<string list>
Human target effect	<set fill>

Filled Event Template for Terrorist Acts

Date	10 January 1990
Location	El Salvador: San Salvador (city)
Event type	BOMBING
Weapon	*“highpower bombs”*
Perpetrator individual	*“guerrilla urban commandos”*
Perpetrator organization	-
Physical target	*“car dealership”*
Physical target effect	some damage
Human target	-
Human target effect	no injury or death

Event Template for Disease Outbreaks

Story:	<document id>
ID:	<template id>
Date:	<date>
Event:	OUTBREAK
Status:	<set fill>
Containment:	<set fill>
Country:	<set fill>
Victims:	<string list>
Disease:	<string>

Filled Event Template for Disease Outbreaks

Story:	20020714.4756
ID:	1
Date:	August 14, 2002
Event:	OUTBREAK
Status:	confirmed
Containment:	none
Country:	Switzerland
Victims:	*the 27 reported cases*
Disease:	*Creutzfeldt-Jakob Disease / [sporadic] Creutzfeldt-Jakob disease (CJD) / CJD / Sporadic CJD / hereditary dominant CJD / Swiss CJD / sporadic Creutzfeldt-Jakob disease*

Unstructured vs. Semi-structured Text

Unstructured text depends 100% on language understanding. *Semi-structured text* has some visual structure (layout) that can aid in understanding.

Unstructured Text

Professor John Skvoretz, U. of South Carolina, Columbia, will present a seminar entitled “Embedded Commitment,” on Thursday, May 4th from 4-5:30 in PH 223D.

Semi-Structured Text

Laura Petitte
Department of Psychology
McGill University

Thursday, May 4, 1995
12:00 pm
Baker Hall 355

Another Semi-Structured Seminar Announcement

Name: Dr. Jeffrey D. Hermes
Affiliation: Department of AutoImmune Diseases
Research & Biophysical Chemistry Merch Research Laboratories
Title: “MHC Class II: A Target for Specific
Immunomodulation of the Immune Response”
Host/e-mail: Robert Murphy
Date: Wednesday, May 3, 1995
Time: 3:30 p.m.
Place: Mellon Institute Conference Room
Sponsor: MERCK RESEARCH LABORATORIES

IE in the Wild: ProMed Disease Outbreak Reports

```
EBOLA HEMORRHAGIC FEVER - UGANDA (09)
************************************
A ProMED-mail post
ProMED-mail, a program of ISID
<http://www.promedmail.org>

[see also:
Ebola hemorrhagic fever - Uganda20001016.1769Ebola hemorrhagic fever - Uganda (08)20001022.1826]

[1]
Date: Sun, 22 Oct 2000 22:18:31 -0200
From: ProMED-mail <promed@promedmail.org>
Source: WHO Disease Outbreaks Report, Sun 21 Oct 2000 [edited]
<http://www.who.int/disease-outbreak-news/>

[HEADLINE : 1 line]
------------------------------------------
[TEXT : 11 lines]

******
[2]
Date: Sun, 22 Oct 2000 22:18:31 -0200
From: ProMED-mail <promed@promedmail.org>
Source: WHO Disease Outbreaks Report, Sun 21 Oct 2000 [edited]
<http://www.who.int/disease-outbreak-news/>

[HEADLINE : 1 line]
------------------------------------------
[TEXT : 3 lines]

--
[PROMED DISCLAIMER  :  22 lines]
```

Headline and Text Portions:

Ebola Haemorrhagic Fever In Uganda - Update 5

As of Sat 21 Oct 2000, the Ugandan Ministry of Health has reported 139 cases including 51 deaths. The increase of 17 cases in the last 24 hours reflects the intensified active surveillance.

A team from the WHO Collaborating Centre at the US Centers for Disease Control and Prevention (CDC), United States is establishing a field diagnostic laboratory in Gulu district. The last laboratory equipment arrived Sat 20 Oct 2000 and the laboratory is expected to be operational shortly. A WHO information officer from Geneva arrived in Uganda on Wed 18 Oct 2000 and is based in Gulu district. He is working with the Ugandan Ministry of Health as media focal point.

Ebola Haemorrhagic Fever In Uganda - Update 6

As of Sun 22 Oct 2000, the Ugandan Ministry of Health has reported 149 cases, including 54 deaths. [This represents an increase of 10 cases and 3 deaths in the last 24 hours. - Mod.CP]

Headline and Text Portions:

Ebola Haemorrhagic Fever In Uganda - Update 5

As of Sat 21 Oct 2000, the Ugandan Ministry of Health has reported 139 cases including 51 deaths. The increase of 17 cases in the last 24 hours reflects the intensified active surveillance.

A team from the WHO Collaborating Centre at the US Centers for Disease Control and Prevention (CDC), United States is establishing a field diagnostic laboratory in Gulu district. The last laboratory equipment arrived Sat 20 Oct 2000 and the laboratory is expected to be operational shortly. A WHO information officer from Geneva arrived in Uganda on Wed 18 Oct 2000 and is based in Gulu district. He is working with the Ugandan Ministry of Health as media focal point.

Ebola Haemorrhagic Fever In Uganda - Update 6

As of Sun 22 Oct 2000, the Ugandan Ministry of Health has reported 149 cases, including 54 deaths. [This represents an increase of 10 cases and 3 deaths in the last 24 hours. - Mod.CP]

Text Annotations for Event Extraction

perpetrator: Alleged guerrilla urban commandos launched

weapon: highpower bombs against *target*: a car dealership in downtown

location: San Salvador *date*: this morning. A police report said that the

attack *damage*: set the building on fire, but *injury*: did not result any casualties.

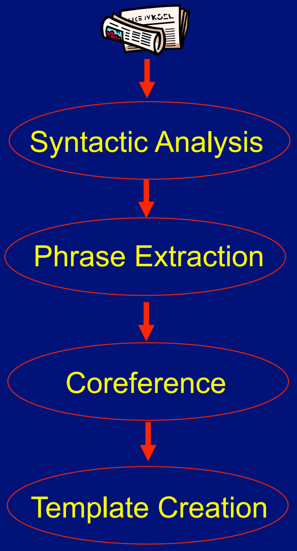
Patterns/Rules vs. Sequence Tagging

Two general approaches to event extraction:

Pattern-based systems use patterns or rules which identify phrases that should be extracted for each event role.

Machine learning classifiers label individual tokens indicating whether they should be extracted, and if so, what role they play.

Template-Filling Pipeline



IBM fired its CEO.

IBM	fired	its CEO.
Subj	*VP*	*Dobj*

Event: FIRING
Firer: IBM *Employee:* its CEO

IBM fired its CEO.
John Smith was let go on Monday.

Event:	FIRING
Firer:	IBM
Employee:	John Smith, CEO
Date:	Monday

Example of Patterns

[Alleged guerrilla urban commandos] launched

Subject = perpetrator

highpower bombs against a car dealership in downtown

San Salvador this morning .

Example of Patterns

Alleged guerrilla urban commandos launched

[highpower bombs] against a car dealership in downtown

DirectObj = instrument

San Salvador this morning .

Example of Patterns

Alleged guerrilla urban commandos launched

highpower bombs against [a car dealership] in downtown

PP(against) = target

San Salvador this morning .

IE as Sequence Tagging

- Event extraction can be modeled as a sequential tagging problem. A supervised sequential learner (e.g., MEMMs or CRFs) can be trained with manually annotated texts.
- Each document is processed sequentially and each token is labeled with respect to event roles.

  B (beginning) and I (inside) tags are needed for each role.

  For example: B_{PERP}, I_{PERP}, B_{VICTIM}, I_{VICTIM}, B_{WEAPON}, I_{WEAPON}
- Common features: words, POS tags, dependency relations, semantic types, and a small context window of preceding/ following words.

Sequence Tagging Example

Alleged	guerrilla	urban	commandos	launched	two
B_{PERP}	I_{PERP}	I_{PERP}	I_{PERP}	O	B_{WEAPON}

highpower	bombs	against	a	car	dealership	in
I_{WEAPON}	I_{WEAPON}	O	B_{TGT}	I_{TGT}	I_{TGT}	O

downtown	San	Salvador	this	morning	.
B_{LOC}	I_{LOC}	I_{LOC}	B_{DATE}	I_{DATE}	

Weakly Supervised Learning for IE

- Idea: can we train an IE system using only unannotated texts?
- Yes, if we have "preclassified" texts:
  - One pile of relevant texts
  - One pile of irrelevant texts
  - Manual review of ranked patterns
- Much easier than annotating texts!

AutoSlog-TS [Riloff 96] (Step 1)

Relevant Irrelevant

[The World Trade Center], [an icon] of [New York City], was horrifically attacked on [an otherwise beautiful day] in [September 2001] by [Al Qaeda].

Shallow Parser

Syntactic Templates

Extraction Patterns:
<subj> was attacked
icon of <np>
was attacked on <np>
was attacked in <np>
was attacked by <np>

AutoSlog-TS (Step 2)

Relevant Irrelevant

Extraction Patterns:
<subj> was attacked
icon of <np>
was attacked on <np>
was attacked in <np>
was attacked by <np>

Extraction Patterns	Freq	Prob
<subj> was attacked	100	.90
icon of <np>	5	.20
was attacked on <np>	80	.79
was attacked in <np>	85	.87
was attacked by <np>	95	.95

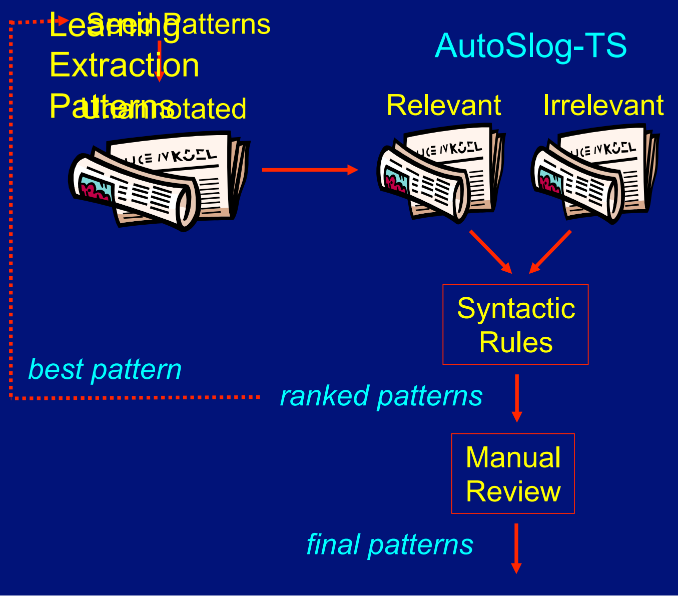
Top Terrorism Extraction Patterns

1. <subject> exploded
2. murder of <np>
3. assassination of <np>
4. <subject> was killed
5. <subject> was kidnapped
6. attack on <np>
7. <subject> was injured
8. exploded in <np>
9. death of <np>
10. <subject> took_place
11. caused <dobj>
12. claimed <dobj>
13. <subject> was wounded
14. <subject> occurred
15. <subject> was located
16. took_place on <np>
17. responsibility for <np>
18. occurred on <np>
19. was wounded in <np>
20. destroyed <dobj>
21. <subject> was murdered
22. one of <np>
23. <subject> kidnapped
24. exploded on <np>
25. <subject> died

Examples of Learned Disease Patterns

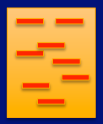
outbreak of <np>	<subj> was transmitted
<subj> spread	contracted <dobj>
cases of <np>	spread of <np>
<subj> was confirmed	<subj> infected
outbreaks of <np>	<subj> killed

ExDisco [Yangarber et al. 2000]

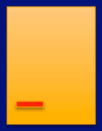


Event Extraction: Fantasy vs. Reality

NLP Fantasy: Every story is an event narrative, describing only the details of a specific event.



NLP Reality: Many stories focus on other issues but contain event information too.



News updates, investigative reports, political speeches, interviews, etc.

- *speculation or identification of suspects*
- *apprehension of perpetrators*
- *claims of responsibility by perpetrators*
- *identification of victims (hospitals, bodies, etc.)*
- *recovered weapons*

Event Narrative Example

A bomb exploded today in a Lima restaurant, and a second device that had been placed in the same establishment was deactivated by the Peruvian national police.

There were no victims, and the explosion caused very little damage to the restaurant, which is located in the commercial area of the residential district of Miraflores.

Guerrillas of the Tupac Amaru Revolutionary Movement (MRTA) have claimed credit for the terrorist act through pamphlets they left on the premises, according to the police.

Secondary Contexts

The victims were identified as David Lecky and James Donnelly.

Oqueli's body was found next to the body of Gilda Flores.

According to witnesses' reports, two 23 to 25-year-old individuals walked to Gen. Leigh's office, on the fourth floor of the building.

The destroyed UCR headquarters is in the Moreno district of Buenos Aires.

Cardenas Guerra is apparently linked to the Medellin drug cartel.

There were seven children, including four of the Vice President's children, in the home at the time.

Event Keywords

Keywords alone are not as reliable as you might think due to ambiguity, metaphor, and context.

The comedian bombed at the club …

Parliament exploded in anger about ...

Obama was attacked by House Republicans …

The Limitations of Event Keywords

Common types of civil unrest events:

strike	… or air strike, military strike, baseball strike
occupation	… or military occupation, profession
rally	… or car rally, tennis rally
riot	… or laughter ("*he's a riot*"), Riot Games
march	… or the month of March

In tweets with civil unrest event keywords, we found that only a small percentage actually described civil unrest:

English:	25%	(624/2500)
Spanish:	16%	(476/3000)

Event Phrases

Many civil unrest descriptions contain no event keywords, only multi-word event expressions. For example:

took to the streets
gathered at the square
packed the square
amassed at the plaza
thronged the capital
halted work
walked off the job
stormed parliament

Subevent expressions can also be important to recognize:

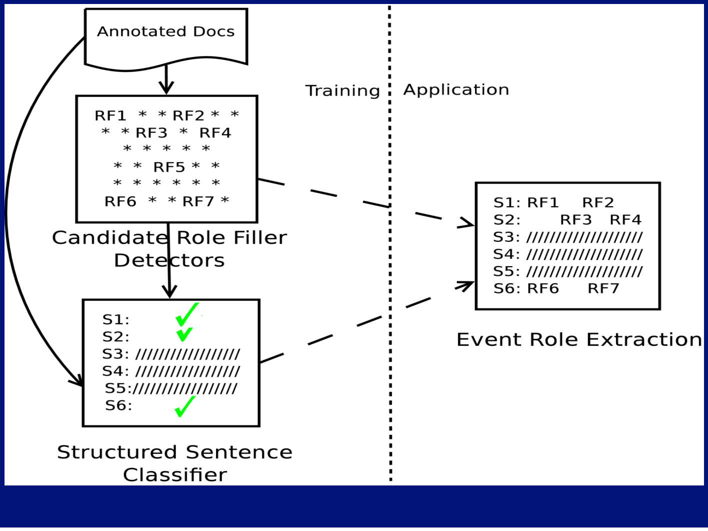
blocked a bridge, burned an effigy, the building, disrupted traffic, smashed windows

Discourse-Guided Event Extraction

[Huang & Riloff, AAAI 2012]

- Goal: use a structured sentence classifier to identify event contexts across sentences.
- The classifier can consider properties of both the current sentence and previous sentence.
- We define several types of discourse features to capture textual cohesion across sentences.

Linker [Huang & Riloff 2012]



Role Filler Extractor

- Train a machine learning classifier (SVM) to label noun phrases based on a small context window around the NP.
- The classifier contains 3 types of features:
  - lexical features (context window, head, premodifiers)
  - NER and semantic class labels
  - lexico-syntactic patterns

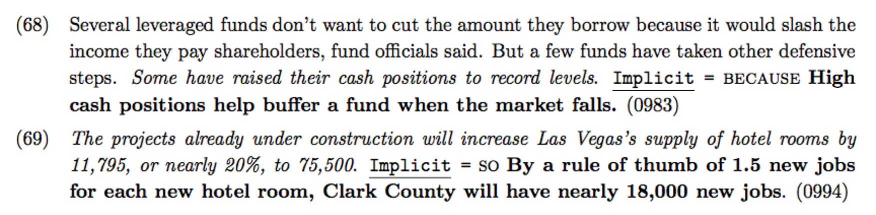
Structured Sentence Classifier

- A sequential tagging model (CRF) is trained to label each sentence as to whether it is an event context .
- Four types of discourse features:
  - Lexical bridge features
  - Discourse bridge features
  - Discourse focus features
  - Role filler distribution features

Penn Discourse Treebank

- The ***Penn Discourse Treebank (PDTB)*** contains texts that have been manually annotated with discourse relations.
- The annotations represent the argument structure, senses and attribution of discourse connectives and their arguments.
- ***Explicit discourse connectives*** require the presence of a discourse cue phrase, such as*: if, because, since, but, however, as a result*
- ***Implicit discourse connectives*** indicate that most readers would infer a discourse relation between adjacent sentences.
- Discourse parsers have been developed by training with the PDTB data.

Examples of Implicit Connectives



(68) Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls.** (0983)

(69) *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* Implicit = SO **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs.** (0994)

Lexical Bridge Features

Lexical Bridge features capture lexical associations between adjacent sentences.

Two types: <verb$_{i-1}$,verb$_{i}$>
<noun$_{i-1}$,noun$_{i}$>

Examples:
<bombed, injured>
<explosion, building>

Discourse Bridge Features

- Discourse relations between adjacent sentences, based on the PDTB discourse parser.
- Labels explicit discourse relations based on cue phrases (e.g., *if* and *because*)
- Labels implicit discourse relations such as *cause*, *condition*, *instantiation*, and *contrast*.
- We capture relations both within a sentence and between the current and previous sentence.

Discourse Focus Feature

- Hypothesis: two sentences are probably related if they have the same discourse focus.
- Create a feature for each shared NP in adjacent sentences that occurs as a Subject, Direct Object, or PP(by).

(1) A customer in the store was shot by masked men.

(2) The two men used 9mm semi-automatic pistols.

→ <*men*, PP(by), Subject>

Role Filler Distribution Features

Purpose: capture information about the presence and types of possible role fillers in the neighborhood.

Features within a sentence:

- type and head noun of each candidate role filler
- density of candidate role fillers
- pairs of different types

Features across adjacent sentences:

- head and type of pairs across sentences
- candidates that share a discourse relation
- verb and candidate pairs across sentences

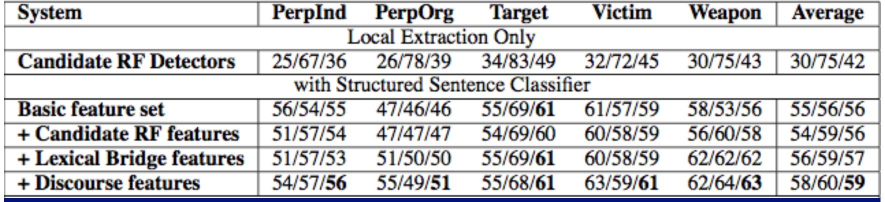
Evaluation Results

Method	**Recall**	**Precision**	**F**
Local Extractor			
Candidate RF Detectors	**75**	30	42

Evaluation Results

Method	**Recall**	**Precision**	**F**
Local Extractor			
Candidate RF Detectors	**75**	30	42
with Structured Sentence Classifier			
Basic N-gram Features	56	55	56
+Extra Features	60	**58**	**59**

Individual Event Role Result



System	**PerpInd**	**PerpOrg**	**Target**	**Victim**	**Weapon**	**Average**
			Local Extraction Only			
Candidate RF Detectors	25/67/36	26/78/39	34/83/49	32/72/45	30/75/43	30/75/42
			with Structured Sentence Classifier			
Basic feature set	56/54/55	47/46/46	55/69/**61**	61/57/59	58/53/56	55/56/56
+ Candidate RF features	51/57/54	47/47/47	54/69/60	60/58/59	56/60/58	54/59/56
+ Lexical Bridge features	51/57/53	51/50/50	55/69/**61**	60/58/59	62/62/62	56/59/57
+ Discourse features	54/57/**56**	55/49/**51**	55/68/**61**	63/59/**61**	62/64/**63**	58/60/**59**

Recent History of Event Extraction on the MUC-4 Terrorism Data Set

Event extraction performance has improved with more complex learning models:

System	Average (P/R/F)
AutoSlog-TS (1996)	45/48/46
GLACIER (2009)	48/57/52
TIER (2011)	51/62/56
Linker (2012)	58/60/59

But progress has been in small increments … there is still much room for improvement!