Information Extraction Overview

Bonan Min

bonanmin@gmail.com

Some slides are based on class materials from Thien Huu Nguyen, Ralph Grishman, Dan Jurasky, James Martin

Information Extraction (IE)

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	

.....

Event Knowledge Base

Trigger	Type	Person1	Person2	Time
divorce	Divorce	Giuliani	Donna Hanover	July

.....

IE = automatically extracting structured information from unstructured and/or semi-structured machine-readable documents

Data Mining

Reasoning

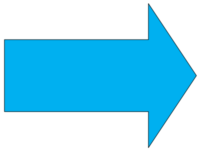
Monitoring

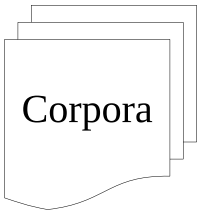
...

Information Extraction Pipeline

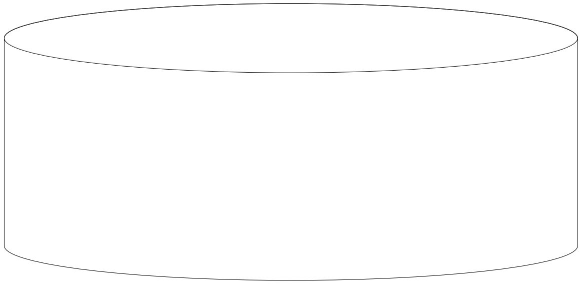
Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

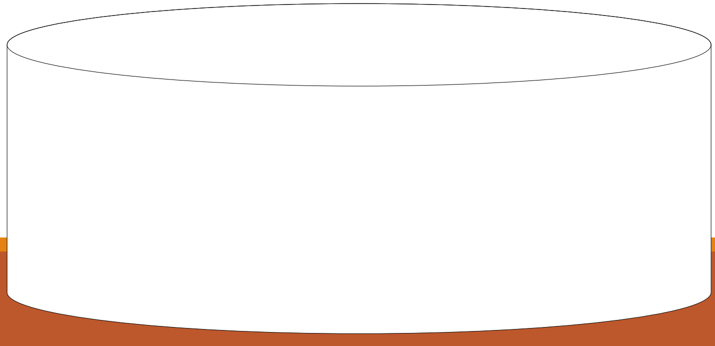




Relation Knowledge Base



Event Knowledge Base



Information Extraction Pipeline

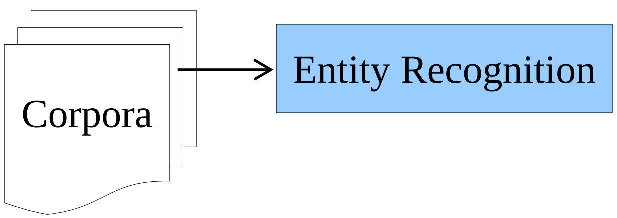
Person

Location

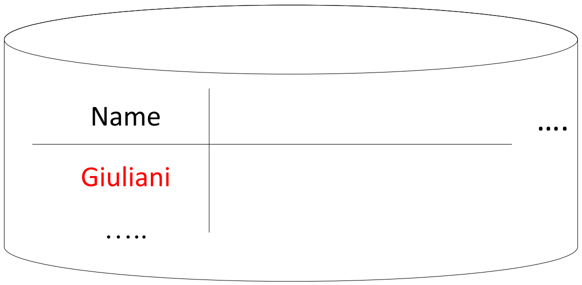
Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

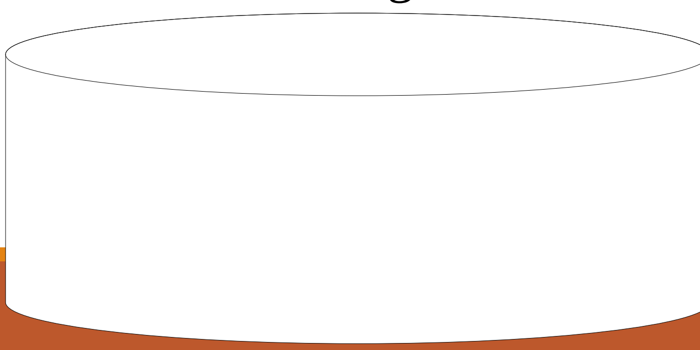
In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



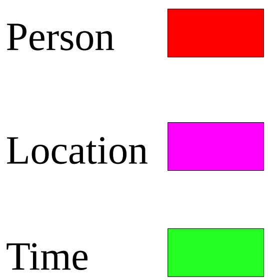
Relation Knowledge Base



Event Knowledge Base

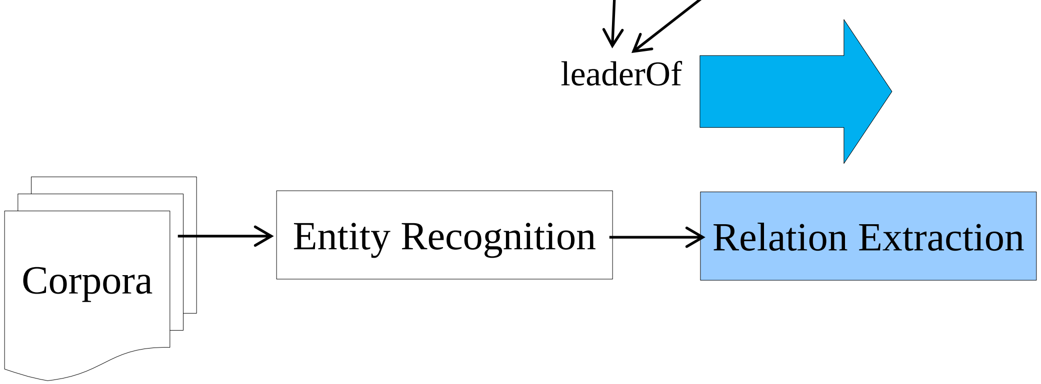


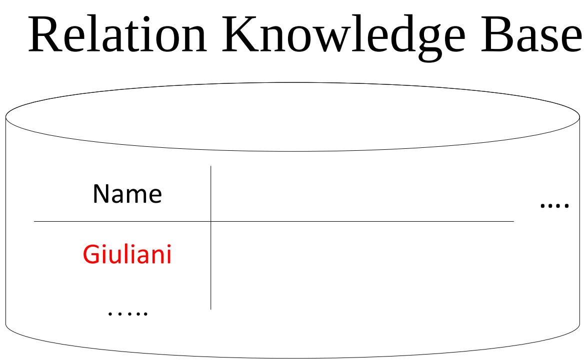
Information Extraction Pipeline

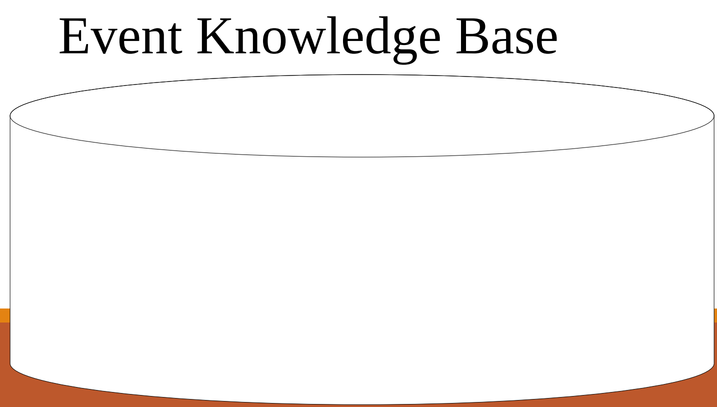


Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

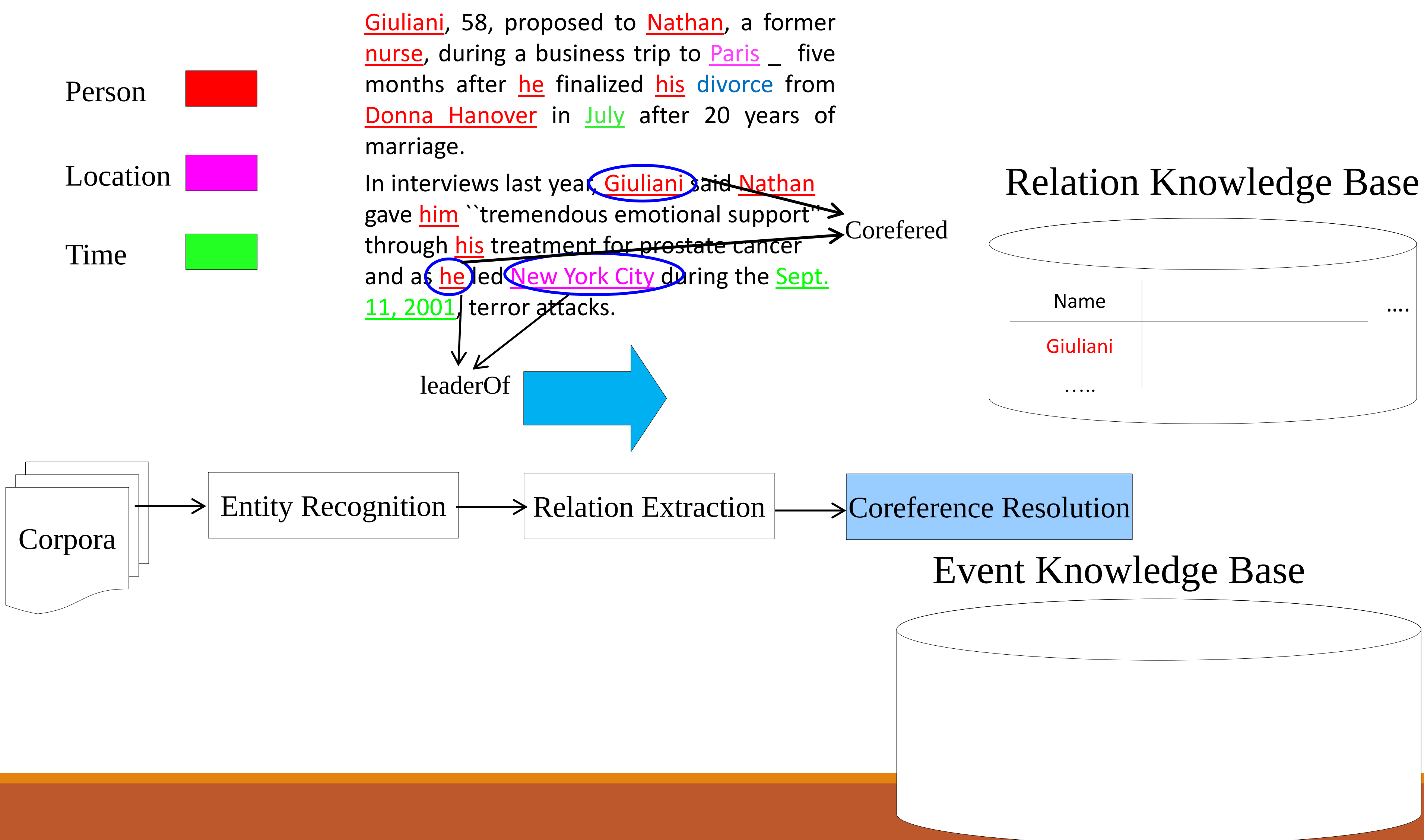
In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.







Information Extraction Pipeline



Information Extraction Pipeline

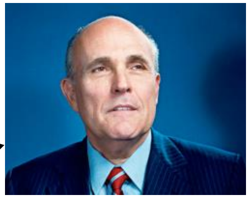
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Corefered

leaderOf

Relation Knowledge Base

Name		....
Giuliani		
…..		

Corpora → Entity Recognition → Relation Extraction → Coreference Resolution → Entity Linking

Event Knowledge Base

Information Extraction Pipeline

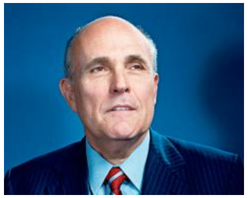
Person

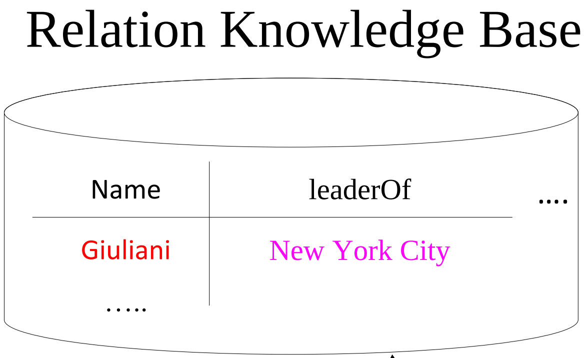
Location

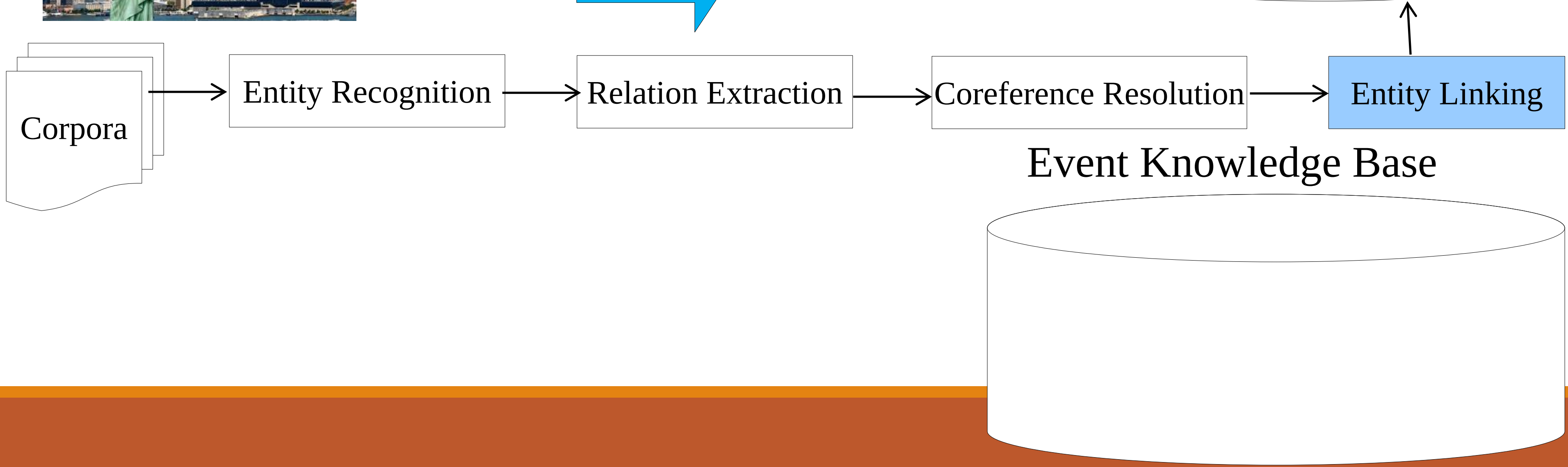
Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.





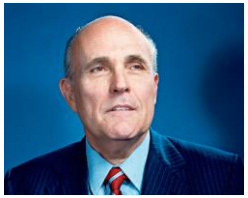


Information Extraction Pipeline

Person

Location

Time



Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
…..		

Corpora → Entity Recognition → Relation Extraction → Coreference Resolution → Entity Linking

Entity Recognition → Trigger Prediction

Event Knowledge Base

Trigger	Type	Person1	Person2	Time
divorce	Divorce			
.....				

Information Extraction Pipeline

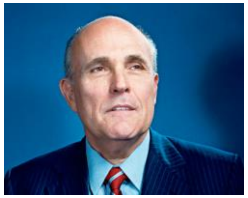
Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.



Corpora → Entity Recognition → Relation Extraction → Coreference Resolution → Entity Linking → Relation Knowledge Base

Entity Recognition → Trigger Prediction → Argument Prediction

Relation Knowledge Base

Name	leaderOf	....
Giuliani	New York City	
…..		

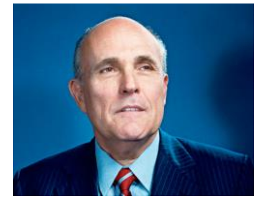
Event Knowledge Base

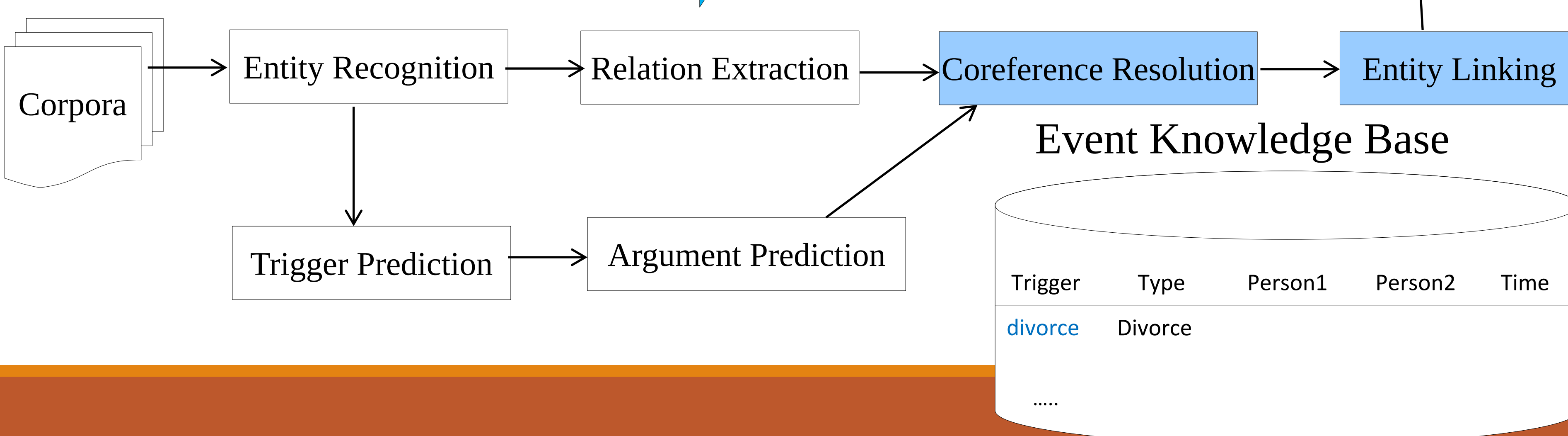
Trigger	Type	Person1	Person2	Time
divorce	Divorce			
…..				

Information Extraction Pipeline

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

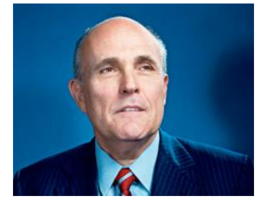


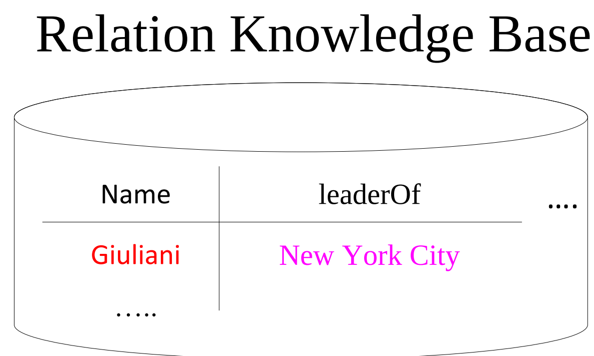


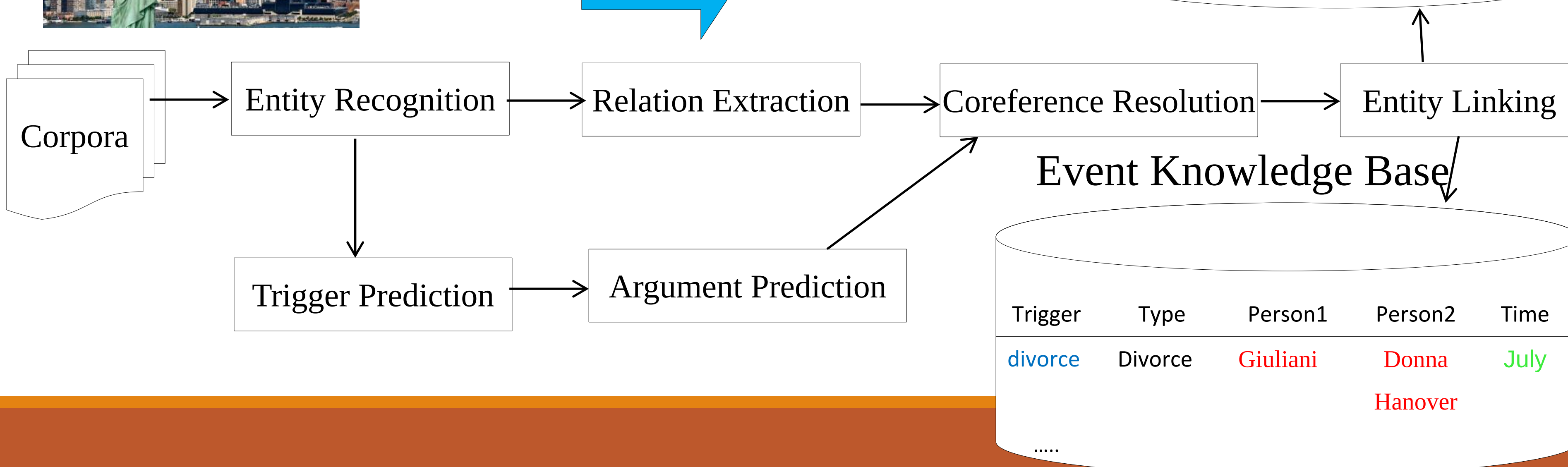
Information Extraction Pipeline

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

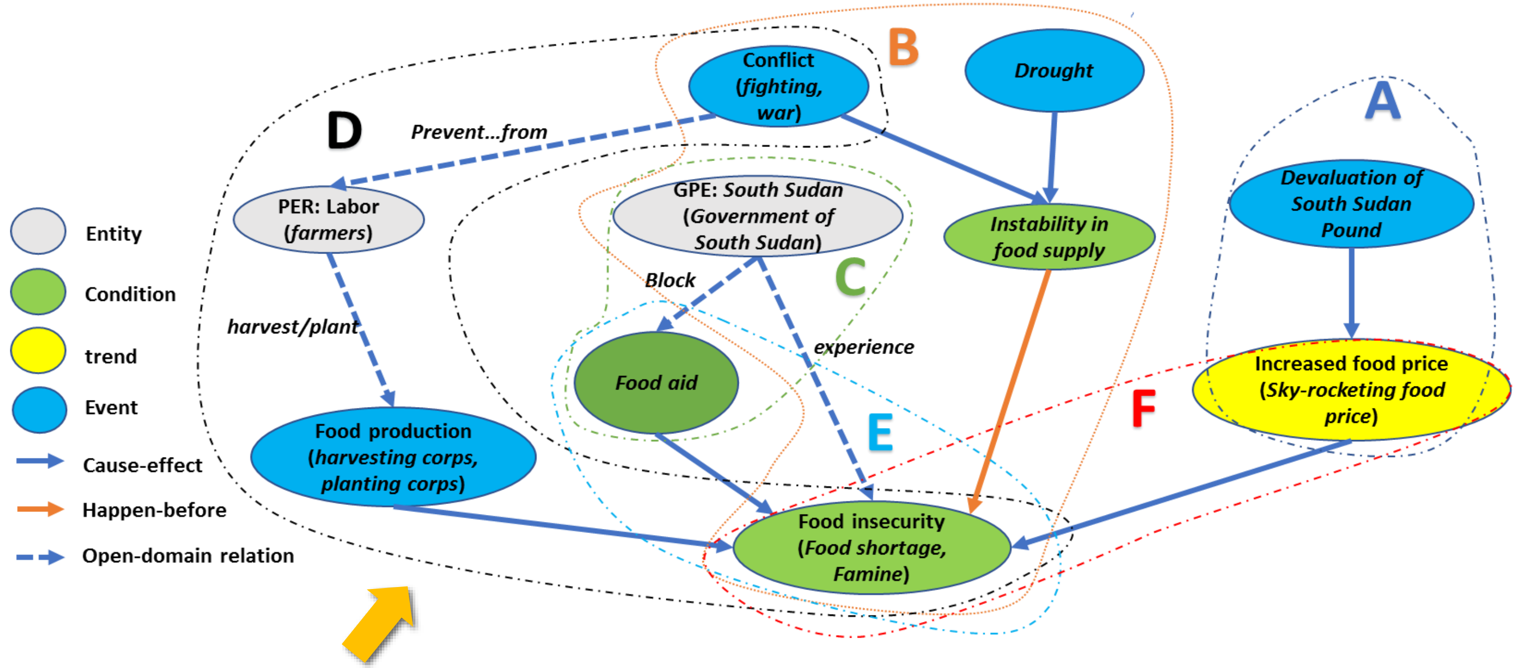






Events, Conditions, Trends, and Causal and Temporal Relations

Raytheon BBN Technologies



A: ***Devaluation of South Sudan Pound*** *essentially leads to* ***increased prices of food****, ...*

B: *...**South Sudan** has been experiencing a **famine** following several years of **instability in the country's food supply** caused by **war** and **drought**...*

C: ***South Sudan's government*** *is* ***blocking food aid****...*

D: *...**fighting** has prevented **farmers** from **planting or harvesting crops**, causing **food shortages** nationwide.*

E: ***Food aid*** *is ... the most efficient means of addressing* ***food insecurity****.*

F: ***Sky-rocketing food prices*** *in South Sudan are deepening* ***food insecurity***

This work was supported by DARPA/I2O and U.S. Army Research Office under the World Modelers program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the DoD or the U.S. Government. This document does not contain technology or technical data controlled under either the U.S. ITAR or the U.S. EAR.

Tasks

Entity Recognition

Relation Extraction

Coreference resolution

Entity Linking

Event extraction (triggers & arguments)

Event-event relation extraction

Entity Recognition

Identifying the entities mentioned in a text (the "entity mentions")

Three types:

- Named mentions: *Giuliani*
- Nominal mentions: *a former nurse*
- Pronominal mentions: *he, his*

The pronominal mentions (and some of the nominal mentions) are references to previously mentioned entities

- Resolving these is the subject of *reference resolution*

Entity Recognition con'd

For named and nominal mentions, we want to be able to define (semantic) classes of entities and identify instances of these classes.

- We have already defined broad classes of names: *Named Entity Recognition (NER)*
- For nominal mentions the semantic class is generally determined by the head of the phrase.

Person

Location

Time

Giuliani, 58, proposed to Nathan, a former nurse, during a business trip to Paris _ five months after he finalized his divorce from Donna Hanover in July after 20 years of marriage.

In interviews last year, Giuliani said Nathan gave him ``tremendous emotional support'' through his treatment for prostate cancer and as he led New York City during the Sept. 11, 2001, terror attacks.

Tasks and Evaluations: ACE and CoNLL 2003 NER

Major tasks and evaluations :

- **Automatic Content Extraction (ACE)** tasks identified seven types of entities: Person, Organization, Location, Facility, Weapon, Vehicle and Geo-Political Entity (GPEs)
- **The CoNLL (Conference on Natural Language Learning) 2003 NER task** consists of newswire text from the Reuters RCV1 corpus tagged with four different entity types (PER, LOC, ORG, MISC)

Large exhaustively annotated corpora was provided for training/development/testing

- Laborious to annotate!

Type	Subtypes
FAC (Facility)	Airport, Building-Grounds, Path, Plant, Subarea-Facility
GPE (Geo-Political Entity[3])	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual
VEH (Vehicle)	Air, Land, Subarea-Vehicle, Underspecified, Water
WEA (Weapon)	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified

Automatic Content Extraction (ACE) entity types

Sekine's Extended Named Entity Hierarchy (https://nlp.cs.nyu.edu/ene/)

Appendix: Extended NE hierarchy

```
TOP
 NAME
  PERSON                    # Bill Clinton, George W. Bush, Satoshi Sekine,
   LASTNAME                 # Clinton, Bush, Sekine,
   MALE_FIRSTNAME           # Bill, George, Satoshi,
   FEMALE_FIRSTNAME         # Mary, Catherine, Ilene, Yoko

  ORGANIZATION              # United Nations, NATO
   COMPANY                  # IBM, Microsoft
   COMPANY_GROUP            # Star Alliance, Tokyo-Mitsubishi Group
   MILITARY                 # The U.S Navy
   INSTITUTE                # the National Football League, ACL
   MARKET                   # New York Stock Exchange, NASDAQ
   POLITICAL_ORGANIZATION #
    GOVERNMENT              # Department of Education, Ministry of Finance
    POLITICAL_PARTY         # Republican Party, Democratic Party, GOP
    PUBLIC_INSTITUTION      # New York Post Office,
   GROUP                    # The Beatles, Boston Symphony Orchestra
    SPORTS_TEAM             # the Chicago Bulls, New York Mets
   ETHNIC_GROUP             # Han race, Hispanic
   NATIONALITY              # American, Japanese, Spanish

  LOCATION                  # Times Square, Ground Zero
   GPE                      # Asia, Middle East, Palestine
    CITY                    # New York City, Los Angeles
    COUNTY                  # Westchester
    PROVINCE                # State (US), Province (Canada), Prefecture (Japan)
    COUNTRY                 # the United States of America, Japan, England
   REGION                   # Scandinavia, North America, Asia, East coast
   GEOLOGICAL_REGION        # Altamira
    LANDFORM                # Rocky Mountains, Manzano Peak, Matterhorn
    WATER_FORM              # Hudson River, Fletcher Pond
    SEA                     # Pacific Ocean, Gulf of Mexico, Florida Bay
   ASTRAL_BODY              # Halley's comet, the Moon
    STAR                    # Sirius, Sun, Cassiopeia, Centaurus
    PLANET                  # the Earth, Mars, Venus
   ADDRESS                  #
    POSTAL_ADDRESS          # 715 Broadway, New York, NY 10003
    PHONE_NUMBER            # 212-123-4567
    EMAIL                   # sekine@cs.nyu.edu
    URL                     # http://www.cs.nyu/cs/projects/proteus
  FACILITY                  # Empire State Building, Hunter Montain Ski Resort
   GOE                      # Pentagon, White House, NYU Hospital
    SCHOOL                  # New York University, Edgewood Elementary School
    MUSEUM                  # MOMA, the Metropolitan Musium of Art
    AMUSEMENT_PARK          # Walt Disney World, Oakland Zoo
    WORSHIP_PLACE           # Canterbury Cathedral, Westminster Abbey
    STATION_TOP             #
     AIRPORT                # JFK Airport, Narita Airport, Changi Airport
     STATION                # Grand Central Station, London Victoria Station
     PORT                   # Port of New York, Sydney Harbour
     CAR_STOP               # Port Authority Bus Terminal, Sydney Bus Depot
```

```
NATURAL_OBJECT         # mitochondria, shiitake mushroom
 ANIMAL                # elephant, whale, pig, horse
 VEGETABLE             # spinach, rice, daffodil
 MINERAL               # Hydrogen, carbon monoxide,

COLOR                  # black, white, red, blue

TIME_TOP
 TIMEX
  TIME                 # 10 p.m.,  afternoon
  DATE                 # August 10, 2001,  10 Aug. 2001,
  ERA                  # Glacial period, Victorian age

 PERIODX               # 2 semesters, summer vacation period
  TIME_PERIOD          # 10 minutes, 15 hours, 50 hours
  DATE_PERIOD          # 10 days, 50 days
  WEEK_PERIOD          # 10 weeks, 50 weeks
  MONTH_PERIOD         # 10 months, 50 months
  YEAR_PERIOD          # 10 years, 50 years

NUMEX                  # 100 pikel, 10 bits
 MONEY                 # $10, 100 yen, 20 marks
 STOCK_INDEX           # 26 5/8,
 POINT                 # 10 points
 PERCENT               # 10%, 10 1/2%
 MULTIPLICATION        # 10 times
 FREQUENCY             # 10 times a day
 RANK                  # 1st prize, booby prize
 AGE                   # 36, 77 years old

 MEASUREMENT           # 10 bytes, 10 Pa, 10 millibar
  PHYSICAL_EXTENT      # 10 meters, 10 inches, 10 yards, 10 miles
  SPACE                # 10 acres, 10 square feet,
  VOLUME               # 10 cubic feet, 10 cubic yards
  WEIGHT               # 10 milligrams, 10 ounces, 10 tons
  SPEED                # 10 miles per hour, Mach 10
  INTENSITY            # 10 lumina, 10 decibel
  TEMPERATURE          # 60 degrees
  CALORIE              # 10 calories
  SEISMIC_INTENSITY    # 6.8 (on Richter scale)

 COUNTX
  N_PERSON             # 10 biologists, 10 workers, 10 terrorists
  N_ORGANIZATION       # 10 industry groups, 10 credit unions
  N_LOCATION           # 10 cities, 10 areas, 10 regions, 10 states
   N_COUNTRY           # 10 countries
  N_FACILITY           # 10 buildings, 10 schools, 10 airports
  N_PRODUCT            # 10 systems, 20 paintings, 10 supercomputers
  N_EVENT              # 5 accidents, 5 interviews, 5 bankruptcies
  N_ANIMAL             # 10 animals, 10 horses, 10 pigs
  N_VEGETABLE          # 10 flowers, 10 daffodils
  N_MINERAL            # 10 diamonds
```

Fine-Grained NER (Ling and Weld, 2012)

Most NER systems are restricted to produce labels from a small set of classes

- E.g., PER, ORG, location (in CoNLL, or ACE)

In order to intelligently understand text, it is useful to *more precisely* determine the semantic classes of entities mentioned in unstructured text

Three challenges impeding the development of a fine-grained NER

- Selection of the tag set: use 112 frequent types from Freebase
- Creation of training data
  - Too large to rely on traditional, manual labeling
  - Exploit the anchor links in Wikipedia text to automatically label entity segments with appropriate tags
- Development of a fast and accurate multi-class labeling algorithm
  - MEMM, CRF, BiLSTM-CRF

person	doctor	**organization**	terrorist_organization
actor	engineer	airline	government_agency
architect	monarch	company	government
artist	musician	educational_institution	political_party
athlete	politician	fraternity_sorority	educational_department
author	religious_leader	sports_league	military
coach	soldier	sports_team	news_agency
director	terrorist		

location	body_of_water	**product**	camera	**art**	written_work
city	island	engine	mobile_phone	film	newspaper
country	mountain	airplane	computer	play	music
county	glacier	car	software	**event**	military_conflict
province	astral_body	ship	game	attack	natural_disaster
railway	cemetery	spacecraft	instrument	election	sports_event
road	park	train	weapon	protest	terrorist_attack
bridge					

building	time	chemical_thing	website
airport	color	biological_thing	broadcast_network
dam	award	medical_treatment	broadcast_program
hospital	educational_degree	disease	tv_channel
hotel	title	symptom	currency
library	law	drug	stock_exchange
power_station	ethnicity	body_part	algorithm
restaurant	language	living_thing	programming_language
sports_facility	religion	animal	transit_system
theater	god	food	transit_line

Xiao Ling and Daniel S. Weld. Fine-Grained Entity Recognition. AAAI 2012.

Relation and Event Extraction

We also would like to extract predications asserted about these entities

- The predications range from simple *relations* to complex *events* which may have multiple arguments (agent, patient, time, location, ...)

We will focus on simple binary relationships with two arguments

- Mainly because these have been most intensively studied, particularly from a machine learning point of view

Binary relationships

- Relations: ***Bill Gates**, co-founder of **Microsoft***.
- Event-argument relations: *I* ***ate*** *a burger* *this morning*

Relation Extraction

A *relation* is a predication about a pair of entities:

- *Rodrigo works for UNED.*
- *Alfonso lives in Tarragona.*
- *Otto's father is Ferdinand.*

Typically they represent information which is permanent or of extended duration.

Rudolph William Louis Giuliani *(/ˌdʒuːliˈɑːni/, Italian: [dʒuˈljaːni]; born May 28, 1944) is an American politician, attorney, and public speaker who served as the 107th Mayor of New York City from 1994 to 2001. He currently acts as an attorney to President Donald Trump.[1] Politically first a Democrat, then an Independent in the 1970s, and a Republican since the 1980s, Giuliani served as United States Associate Attorney General from 1981 to 1983. That year he became the United States Attorney for the Southern District of New York, holding the position until 1989.[2]*

Rudy Giuliani	
107th Mayor of New York City	
In office	
January 1, 1994 – December 31, 2001	
Preceded by	David Dinkins
Succeeded by	Michael Bloomberg
United States Attorney for the Southern District of New York	
In office	
June 3, 1983 – January 1, 1989	
President	Ronald Reagan
Preceded by	John S. Martin Jr.
Succeeded by	Benito Romano (Acting)
United States Associate Attorney General	
In office	
February 20, 1981 – June 3, 1983	
President	Ronald Reagan
Preceded by	John Shenefield
Succeeded by	D. Lowell Jensen
Personal details	
Born	Rudolph William Louis Giuliani May 28, 1944 (age 75) New York City, New York, U.S.
Political party	Republican (1980–present)
Other political affiliations	Independent (1975–1980) Democratic (before 1975)

ACE (2005-2008)

Pre-defined types between ACE entities

Large exhaustively annotated corpora (599 files for ACE 2005) was provided for training/development/testing

- Laborious to annotate!

Focus on mention-mention relation extraction

Influential in relation extraction research

Type	**Subtype**
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (Gen-affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*none*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

Relation types in ACE

KBP Slot Filling (2009-2017)

Slot Filling (SF): The **slot filling** task is to search the document collection to **fill** in values for specific attributes ("**slots**") for specific entities

Question-answering style evaluation:
- *What's the age of Barack Obama?*
- *Who is the spouse of Barack Obama?*

Focus on getting the answer
- System needs to deduplicate answers

Do not provide annotated corpus for training
- System needs to come up with clever ways to find heuristically labeled data to train an extractor

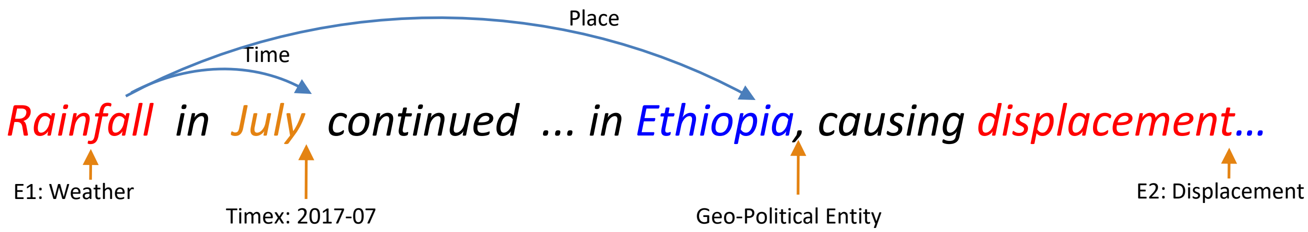
Train from Wikipedia
with distant supervision!

Person Slots			Organization Slots		
Name	**Type**	**List?**	**Name**	**Type**	**List?**
per:alternate_names	Name	Yes	org:alternate_names	Name	Yes
per:date_of_birth	Value		org:political_religious_affiliation	Name	Yes
per:age	Value		org:top_members_employees	Name	Yes
per:country_of_birth	Name		org:number_of_employees_members	Value	
per:stateorprovince_of_birth	Name		org:members	Name	Yes
per:city_of_birth	Name		org:member_of	Name	Yes
per:origin	Name	Yes	org:subsidiaries	Name	Yes
per:date_of_death	Value		org:parents	Name	Yes
per:country_of_death	Name		org:founded_by	Name	Yes
per:stateorprovince_of_death	Name		org:date_founded	Value	
per:city_of_death	Name		org:date_dissolved	Value	
per:cause_of_death	String		org:country_of_headquarters	Name	
per:countries_of_residence	Name	Yes	org:stateorprovince_of_headquarters	Name	
per:statesorprovinces_of_residence	Name	Yes	org:city_of_headquarters	Name	
per:cities_of_residence	Name	Yes	org:shareholders	Name	Yes
per:schools_attended	Name	Yes	org:website	String	
per:title	String	Yes			
per:employee_or_member_of	Name	Yes			
per:religion	String	Yes			
per:spouse	Name	Yes			
per:children	Name	Yes			
per:parents	Name	Yes			
per:siblings	Name	Yes			
per:other_family	Name	Yes			
per:charges	String	Yes			

Slots in KBP Slot Filling

Event Extraction

Event Extraction: An Example



Preprocessing

- Tagging named entities, mentions and value mentions (e.g., time)

Event Extraction

- Event detection: detect and classify event mentions
- Argument Extraction: attach event arguments *Who, When* (Time), and *Where (Place)*

Event Extraction con'd

Scenario Template (MUC: Message Understanding Conference)

- The scenario template task originally was *the* IE task for the MUC evaluations
  - Identify participants, locations, dates etc. of a class of events -- a naval engagement, a terrorist incident, a joint venture.
  - A single template included related information, such as an **attack and its effects**; this led to some **relatively complex templates**
- With later MUCs (6 and 7), the task narrowed to single events or closely related events -- executive succession, rocket launchings

For the ACE evaluations, this became the event extraction task. An event is

- a specific occurrence involving participants.
- something that happens.
- frequently described as a change of state.

ACE (2005-2008)

Pre-defined types

- Event types over trigger words
- Event argument roles are relationships between pairs of trigger words and ACE entity mentions or value mentions (e.g., time, charge)

Large exhaustively annotated corpora (599 files for ACE 2005) was provided for training/development/testing

- Laborious to annotate!

Hugely influential in event extraction

Types	Subtype
Life	Be-Born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-Ownership, Transfer-Money
Business	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
Personnel	Start-Position, End-Position, Nominate, Elect
Justice	Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

Event types in ACE 2005

Information Extraction vs. Information Retrieval

Information Retrieval returns a set of documents given a query.

Information Extraction returns facts from documents

E.g., What you search for in real estate advertisements:

- Town/suburb. You might think easy, but:
  - Real estate agents: Coldwell Banker, Mosman
  - Phrases: Only 45 minutes from Parramatta
  - Multiple property ads have different suburbs in one ad
- Money: want a range not a textual match
  - Multiple amounts: was $155K, now $145K
- Bedrooms
  - Variations: br, bdr, beds, B/R

Information Extraction Evaluations

CoNLL has annual evaluations of IE components for about 15 years

NIST has organized (annual) US government-sponsored evaluations of information extraction for about 25 years

- covering both components and integrated systems
- MUC [Message Understanding Conferences] in the 1990's
- ACE [Automatic Content Extraction] 2000-2008
- KBP [Knowledge Base Population] since 2009

Named Entity Recognition

Supervised Learning for NER

Named entities are crucial to different IE and QA tasks

For Named Entity Recognition (NER) (find and classify names in text) , we can use the sequence labeling methods discussed previously (i.e., MEMM, CRF, RNN).

Person				Organization	
Fred	Smith	works	for	Time	inc.
B_PER	I_PER	O	O	B_ORG	I_ORG

Sequence Tagging Models for NER

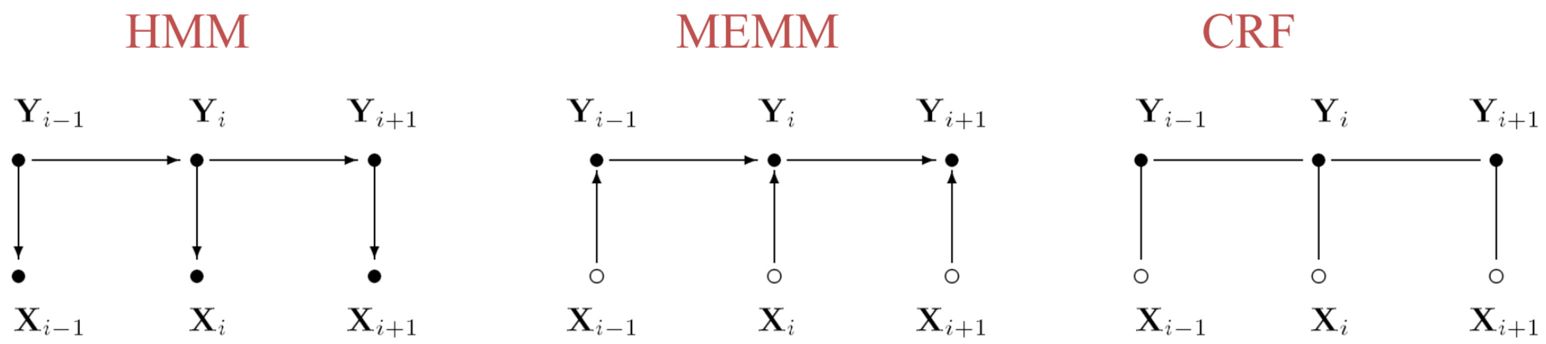


Figure 2. Graphical structures of simple HMMs (left), MEMMs (center), and the chain-structured case of CRFs (right) for sequences. An open circle indicates that the variable is not generated by the model.

From Lafferty et al. 2001

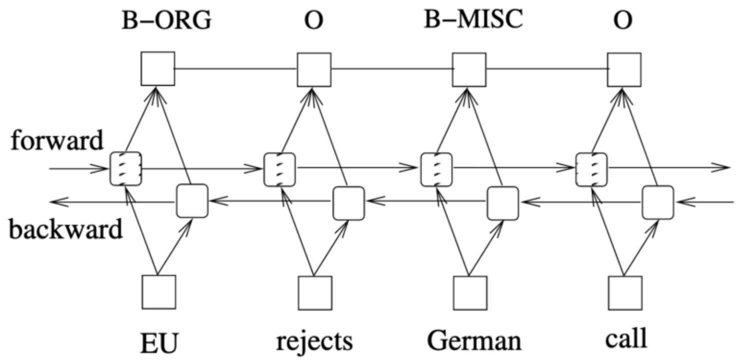


Figure 7: A BI-LSTM-CRF model.

Huang et al. 2015, “Bidirectional LSTM-CRF Models for Sequence Tagging

Features for NER

Feature-based models: the key is to design good feature sets to feed into the sequence labeling models (i.e., feature engineering with MEMM or CRF)

Mr. Gates said…

B-PER I-PER

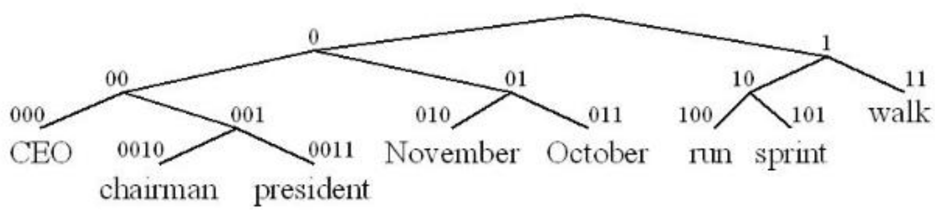
Features for each token:

- previous, current, and next tokens
- POS and phase chunk tags: NPs are more likely to be names
- The tag assigned to the previous token (generated on the fly)
- Combinations of the above
- Word clusters, e.g., Brown word clusters
- Word embeddings

Good indicators of person, organization, and location names

- A name that is followed by a comma and a state or country name is probably the name of a city
- Lists of common first/last person names (from census), or location names from WikiData.

10110111	Richard	50145
10110111	Martin	50309
10110111	Thomas	51997
10110111	Peter	54179
10110111	Mark	55466
10110111	Ali	61819
10110111	James	64596
10110111	Mike	68074
10110111	Robert	74256
10110111	Paul	80092
10110111	George	93816
10110111	Michael	99169
10110111	David	119402
10110111	General	120541
10110111	John	169172
101110000	371,000	10
101110000	1,295	10
101110000	422.5	10
101110000	237,000	10
101110000	51.02	10
101110000	97.3	10
101110000	12.24	10
101110000	314,000	10



Brown word clusters

Features for NER

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
w_i is all upper case
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
presence of hyphen

Figure 17.5 Typical features for a feature-based NER system.

prefix(w_i) = `L`	suffix(w_i) = `tane`
prefix(w_i) = `L'`	suffix(w_i) = `ane`
prefix(w_i) = `L'O`	suffix(w_i) = `ne`
prefix(w_i) = `L'Oc`	suffix(w_i) = `e`
word-shape(w_i) = **`X'Xxxxxxx`**	short-word-shape(w_i) = **`X'Xx`**

Features for NER

Word shape features: Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

Shorter word shape features: consecutive character types are removed (i.e., DC10-30 -> Xd-d, I.M.F -> X.X.X)

Gazetteers: Lists of common names for different types

- Millions of entries for locations with detailed geographical and political information (www.geonames.org)
- Lists of first names and surnames derived from its decadal census in the U.S (www.census.gov)
- Typically implemented as a binary feature for each name list
- Unfortunately, such lists can be difficult to create and maintain, and their usefulness varies considerably.

Homework Assignment #3

Design and extract features for CoNLL-style NER

Target classes: PER, ORG, LOC, and MISC
- We'll simply the task to only use IO schema (I=Inside; O=Outside)

Data: We'll provide training/development/test data
- Training & development data comes with NER labels
- We also provide POS & chunk tags
- Format is similar to homework #2 (one word per line)

We recommend using opennlp MaxEnt package (Java) and will provide code for
- Training a MaxEnt model from a feature file
- Decode a MaxEnt model over the testing data
- A sample program (GenerateFeaturesForNER.java) on how to extract features from the training data

You task is not to implement the Machine Learning model, but to implement code that **extracts feature**s from the training, development, and testing data
- In other words, implement a fancier version of *GenerateFeaturesForNER.java*
- Note: you don't have to write this in Java, as long as you extract the features and write them into the file following the format that MaxEnt package requires.

What make good features for a NER model?

Deep learning for NER

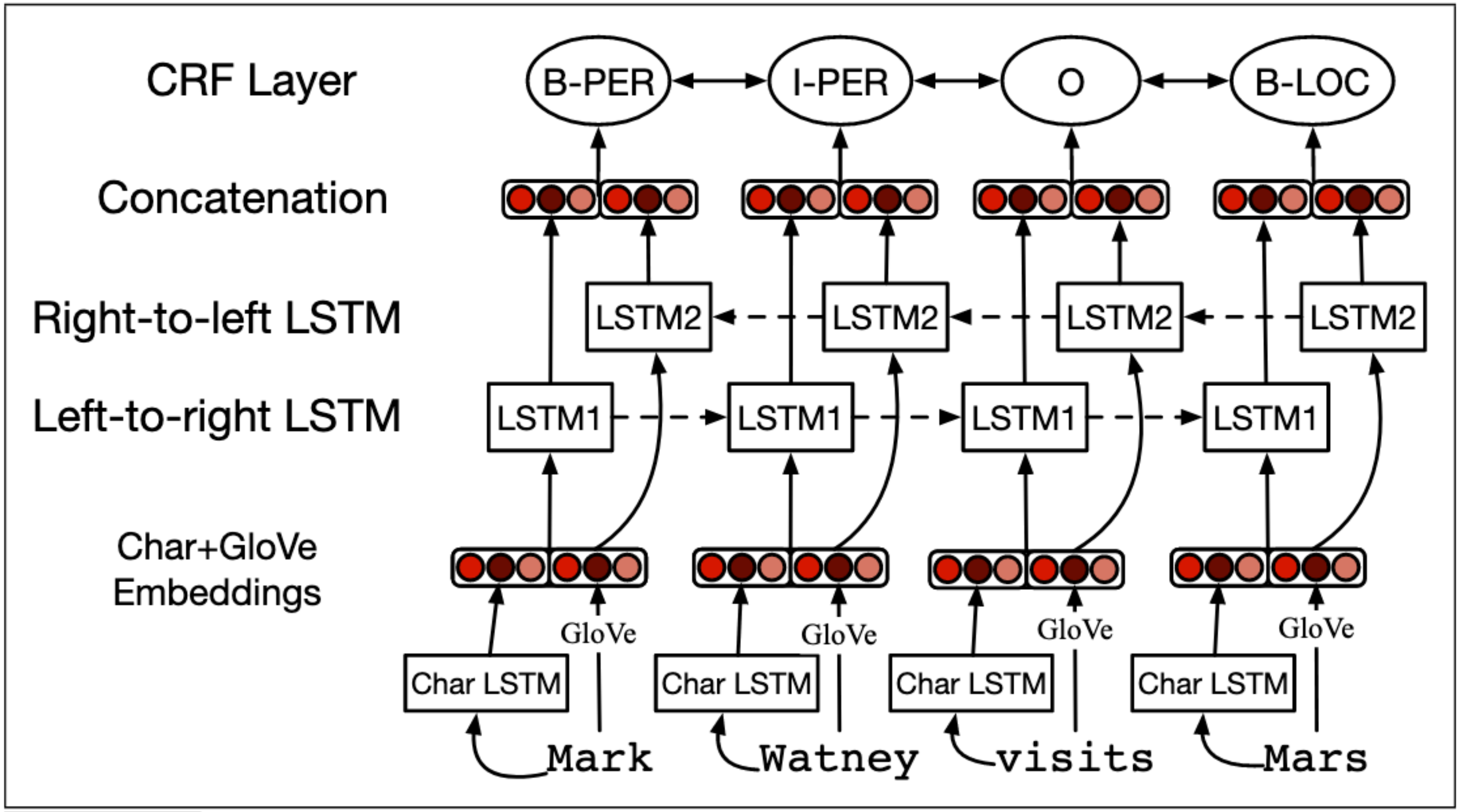


Figure 17.8 Putting it all together: character embeddings and words together a bi-LSTM sequence model. After Lample et al. (2016).

Evaluation for NER systems

	1	2	3	4	5	6	7
	tim	cook	is	the	CEO	of	Apple
gold	B-PER	I-PER	O	O	O	O	B-ORG
system	B-PER	O	O	O	B-PER	O	B-ORG

<start, end, type>

gold

<1,2,PER>
<7,7,ORG>

system

<1,1,PER>
<5,5,PER>
<7,7,ORG>

Precision	1/3
Recall	1/2

Other Types Of Learning (Not limited to NER)

We have discussed hand-coded rules and supervised models (HMM, MEMM, CRF, RNN) for NER [named entity recognition]

- A large labeled training dataset is required
- Annotating a large corpus to train a high-performance NER is fairly expensive

Semi-supervised learning

- Part of training data is labeled ('the seed') (the rest is unlabeled)
- Make use of redundancies to learn labels of additional data, then train model
- Co-training
- Reduces amount of data which must be hand-labeled to achieve a given level of performance

Active learning

- Start with partially labeled data
- System selects additional 'informative' examples for user to label

Semi-supervised Learning

L = labeled data

U = unlabeled data

1. L = seed

   repeat 2-4 until stopping condition is reached

2. C = classifier trained on L
3. Apply C to U.
   N = most confidently labeled items
4. L += N; U -= N

Confidence

How to estimate confidence?

Binary probabilistic classifier

- Confidence = | P – 0.5 | * 2

N-ary probabilistic classifier

- Confidence = $P_1 - P_2$

where

P_1 = probability of most probable label

P_2 = probability of second most probable label

SVM

- Distance from the separating hyperplane

Co-Training (Multi-View Learning)

Two 'views' of data (subsets of features)

- Producing two classifiers $C_1(x)$ and $C_2(x)$

Ideally

- Independent
- Each sufficient to classify data

Apply classifiers in alternation (or in parallel)

1. L = seed
   -- repeat 2-7 until stopping condition is reached
2. C_1= classifier trained on L
3. Apply C_1 to U.
   N = most confidently labeled items
4. L += N; U -= N
5. C_2 = classifier trained on L
6. Apply C_2 to U.
   N = most confidently labeled items
7. L += N; U -= N

When to stop?

- U is exhausted
- Reach performance goal using held-out labeled sample
- After fixed number of iterations based on similar tasks

Poor confidence estimates

- Errors from poorly-chosen data rapidly magnified

Co-Training for NER

We can split the features for NER into two sets:

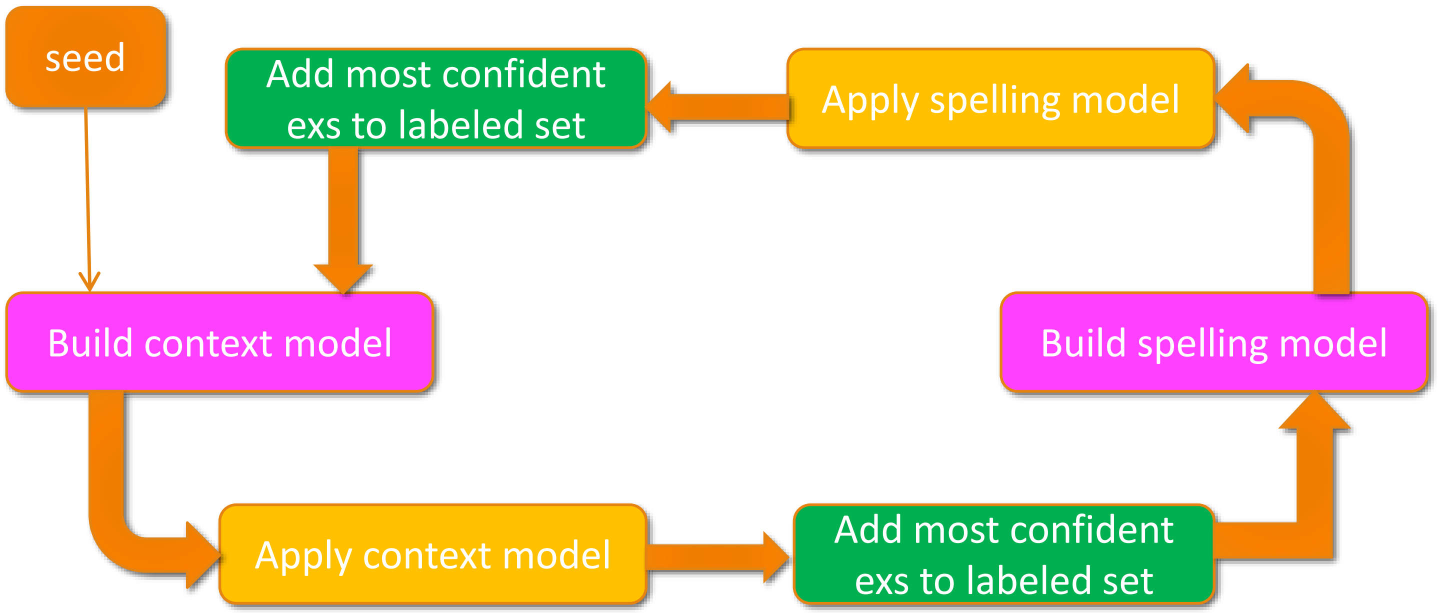
- Spelling features
  (the entire name + tokens in the name)
- Context features
  (left and right contexts + syntactic context)

Start with a seed

- E.g., some common unambiguous full names

Iteratively grow seed, alternatively applying spelling and context models and adding most -confidently-labeled instances to seed

Co-Training for NER



Name Co-training: Results from Collins and Singer (1999)

- 3 classes: person, organization, location (and 'other')
- Data: 1M sentences of news
- Seed:

  New York, California, U.S. → location
  contains(Mr.) → person
  Microsoft, IBM → organization
  contains(Incorporated) → organization

- Apply constraints, e.g., took names appearing with appositive modifier
- Accuracy: 83% or 91% (Clean accuracy: ignoring names not in one of the 3 categories)

Semi-supervised NER: When To Stop

Semi-supervised NER labels a few more examples at every iteration

- It stops when it runs out of examples to label

This is fine if

- Names are easily identified (e.g., by capitalization in English)
- Most names fall into one of the categories being trained (e.g., people, organizations, and locations for news stories)

Semi-supervised NER: Semantic Drift

Semi-supervised NER doesn’t work so well if

- The set of names is hard to identify
  - Monocase languages
  - Extended name sets including lower-case terms
- The categories being trained cover only a small portion of the set of names

The result is *semantic drift* and *semantic spread*

- The name categories gradually grow to include related terms

Fighting Semantic Drift

We can fight drift by training a larger, more inclusive set of categories

- Including ‘negative’ categories
  - Categories we don’t really care about but include to compete with the original categories
- These negative categories can be built
  - By hand (Yangarber et al. 2003)
  - Or automatically (McIntosh 2010)

Active Learning

For supervised learning, we typically annotate text data sequentially

Not necessarily the most efficient approach

- Most natural language phenomena have a Zipfian distribution … a few very common constructs and lots of infrequent constructs
- After you have annotated "Spain" 50 times as a location, the NER model is little improved by annotating it one more time

We want to select the most *informative* examples and present them to the annotator

- The data which, if labeled, is most likely to reduce NER error

How To Select Informative Examples?

Uncertainty-based sampling

- For binary classifier
  - For MaxEnt, probability near 50%
  - For SVM, data near separating hyperplane
- For n-ary classifier, data with small margin

Committee-based sampling

- Data on which committee members disagree
- (co-testing … use two classifiers based on independent views)

Representativeness

Selecting examples that are representative (centroid of clusters)

Or it’s more helpful to annotate examples involving less common features

- Weighting these features correctly will have a larger impact on error rate
- So we rank examples by frequency of features in the entire corpus

Batching and Diversity

Each iteration of active learning involves running classifier on (a large) unlabeled corpus

- This can be quite slow
- Meanwhile annotator is waiting for something to annotate

So we run active learning in batches

- Select best n examples to annotate each time
- But all items in a batch are selected using the same criteria and same system state, and so are likely to be similar

To avoid example overlap, we impose a diversity requirement with a batch: limit maximum similarity of examples within a batch

- Compute similarity based on example feature vectors

Simulated Active Learning

True active learning experiments are
- Hard to reproduce
- Very time consuming

So most experiments involve *simulated active learning*:
- “unlabeled” data has really been labeled, but the labels have been hidden
- When data is selected, labels are revealed
- Disadvantage: “unlabeled” data can’t be so bit

This leads us to ignore lots of issues of true active learning:
- An annotation unit of one sentence or even one token may not be efficient for manual annotation
- So reported speed-ups may be optimistic
  (typical reports reduce by half the amount of data to achieve a given NER accuracy)

Limitations

Cited performance is for well matched training and test

- Same domain
- Same source
- Same epoch

- Performance deteriorates rapidly if less matched
  - NER trained on Reuters (F=91),
    tested on Wall Street Journal (F=64) [Ciaramita and Altun 2003]
- Work on NER adaptation is vital

Adding rarer classes to NER is difficult

- Supervised learning inefficient
- Semi-supervised learning is subject to semantic drift