Relation Extraction

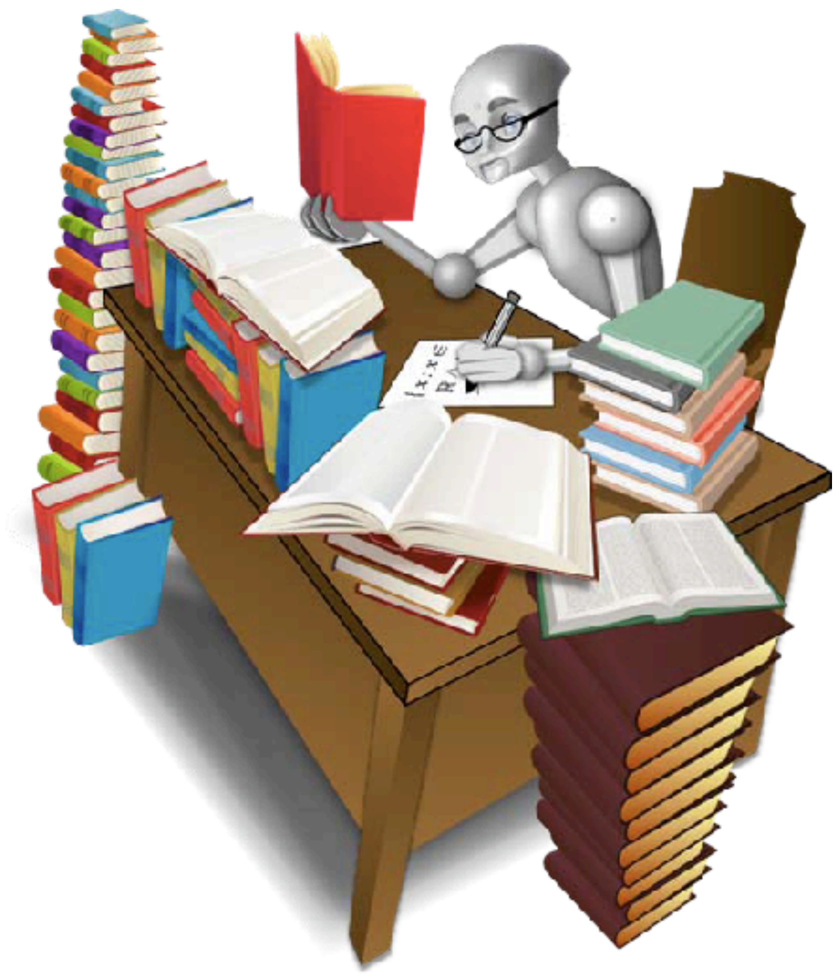
Bill MacCartney

CS224U

26 January 2012

A mish-mash of slides from many people, including Dan Jurafsky, Rion Snow, Jim Martin, Chris Manning, William Cohen, and others

Goal: “Machine Reading”



DARPA slide

Background: Information Extraction

- IE = extracting information from text
- Sometimes called *text analytics* commercially
- Extract entities
  - (the people, organizations, locations, times, dates, genes, diseases, medicines, etc. in a text)
- Extract the relations between entities
- Figure out the larger events that are taking place

What is Information Extraction?

As a task: Filling slots in a database from sub-segments of text.

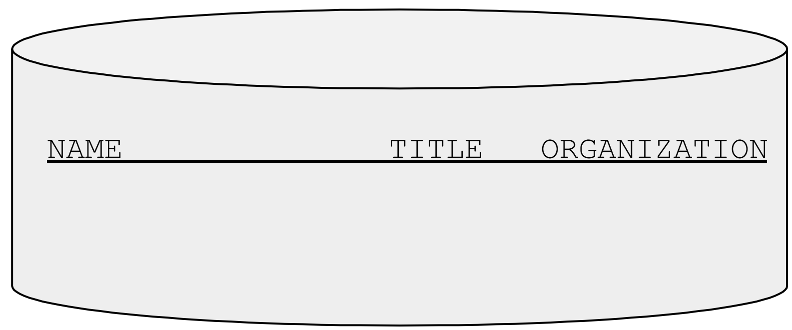
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



Slide from William Cohen

What is Information Extraction?

As a task: **Filling slots in a database from sub-segments of text.**

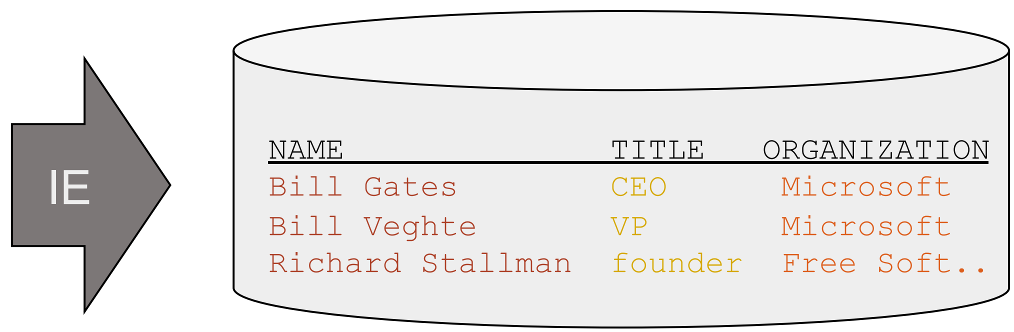
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Slide from William Cohen

What is Information Extraction?

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

“named entity extraction”

Slide from William Cohen

What is Information Extraction?

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Slide from William Cohen

What is Information Extraction?

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

Slide from William Cohen

What is Information Extraction?

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

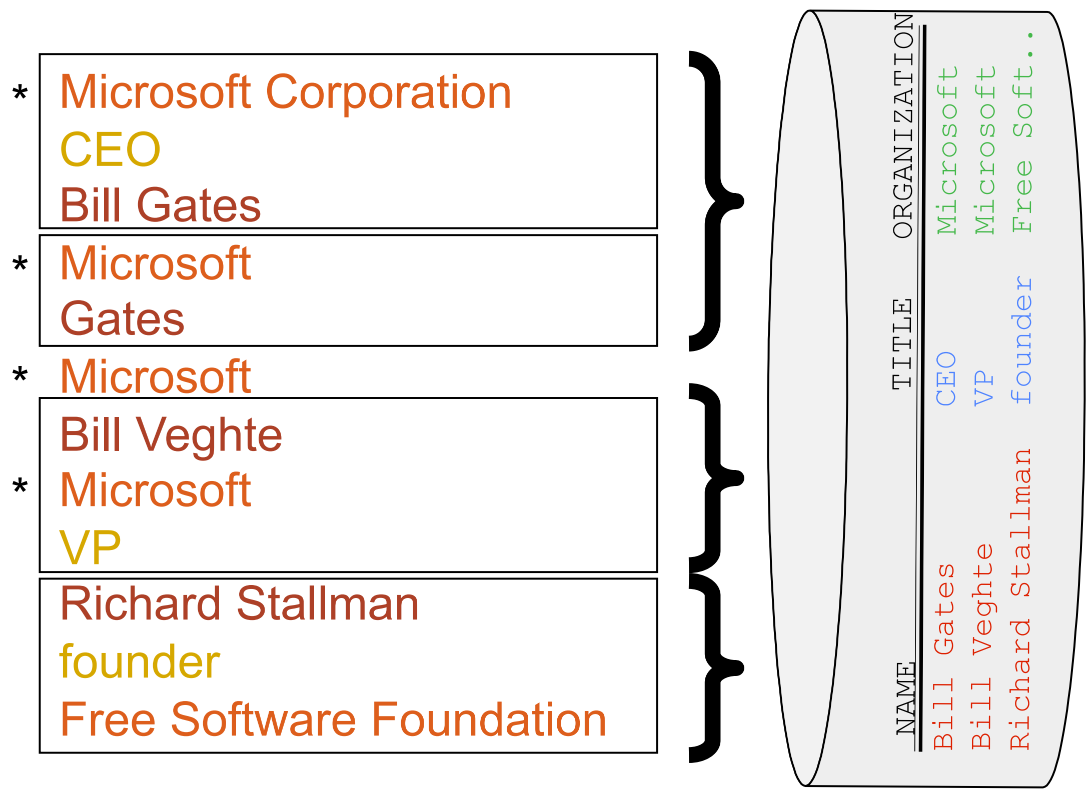
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

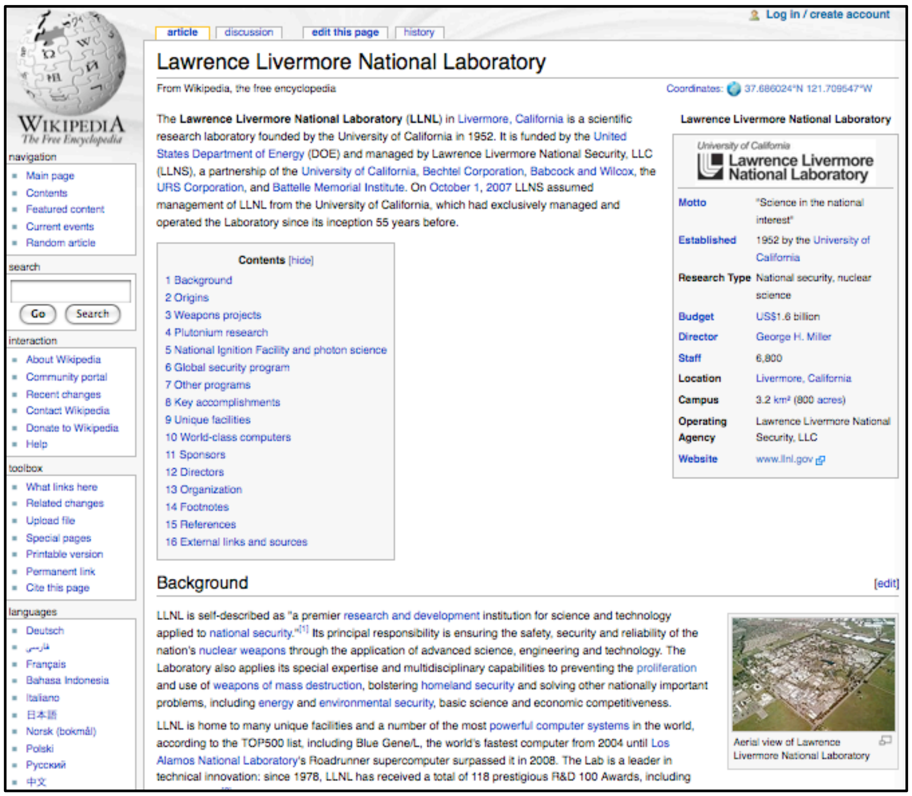
Richard Stallman, founder of the Free Software Foundation, countered saying…



Slide from William Cohen

Extracting Structured Knowledge

Each article can contain hundreds or thousands of items of knowledge...



"The Lawrence Livermore National Laboratory (LLNL) in Livermore, California is a scientific research laboratory founded by the University of California in 1952."

LLNL EQ Lawrence Livermore National Laboratory
LLNL LOC-IN *California*
Livermore LOC-IN *California*
LLNL IS-A scientific research laboratory
LLNL FOUNDED-BY University of California
LLNL FOUNDED-IN 1952

Goal: Machine-readable summaries


Involvement of Tumor Necrosis Factor Receptor-associated Protein 1 (TRAP1) in Apoptosis Induced by β-Hydroxyisovalerylshikonin*

Textual abstract:
Summary for human

Subject	Relation	Object
p53	**is_a**	protein
Bax	**is_a**	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

Structured knowledge extraction:
Summary for machine

From Unstructured Text to Structured Knowledge

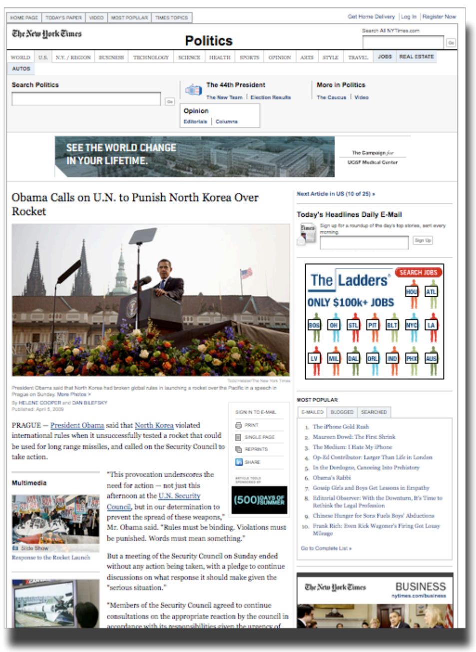
Unstructured Text

News articles...

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text


Politics

Obama Calls on U.N. to Punish North Korea Over Rocket

Blog posts....

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Involvement of Tumor Necrosis Factor Receptor-associated Protein 1 (TRAP1) in Apoptosis Induced by β-Hydroxyisovalerylshikonin*

Scientific journal articles...

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Tweets, instant messages, chat logs...

slide from Rion Snow

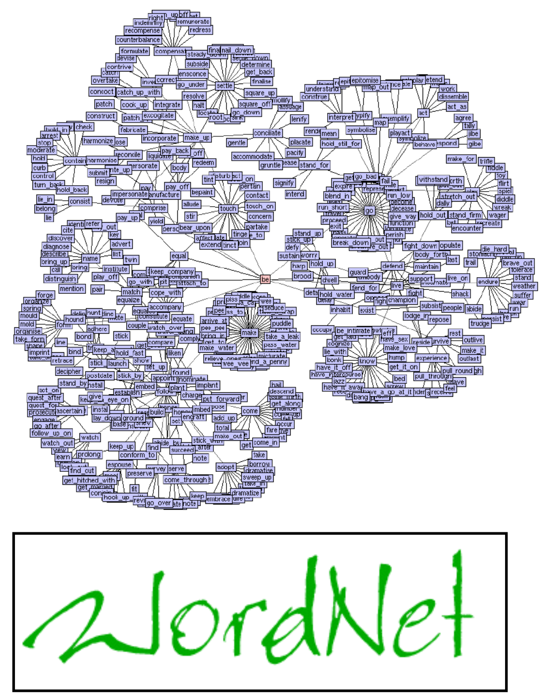
Unstructured Text

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge

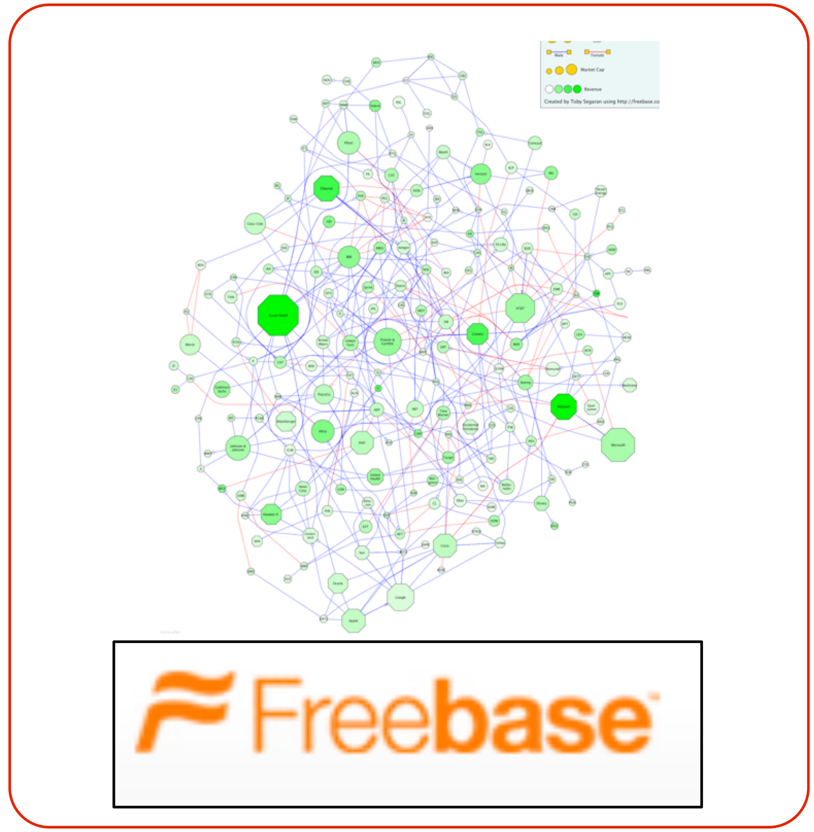


slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge

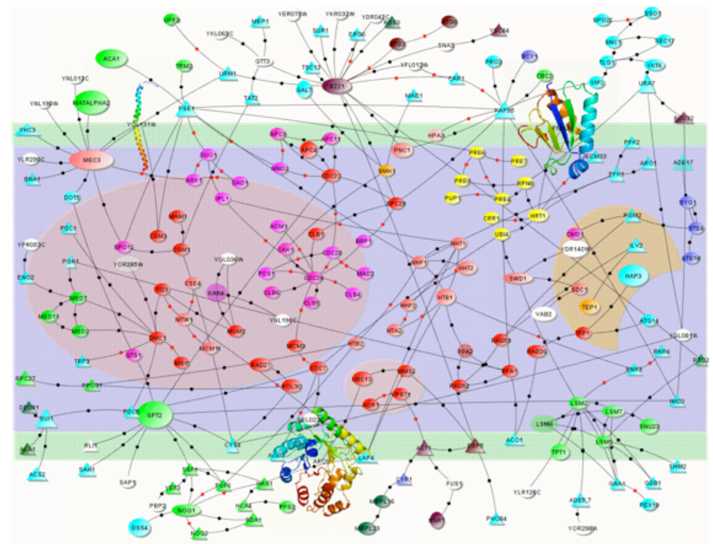


slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge



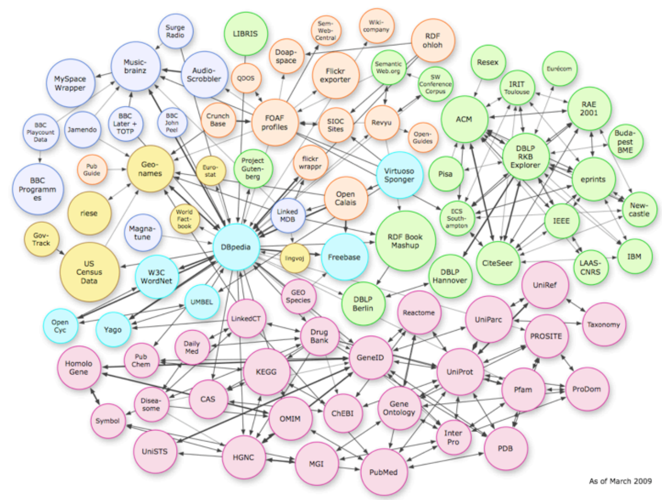
the Gene Ontology

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge



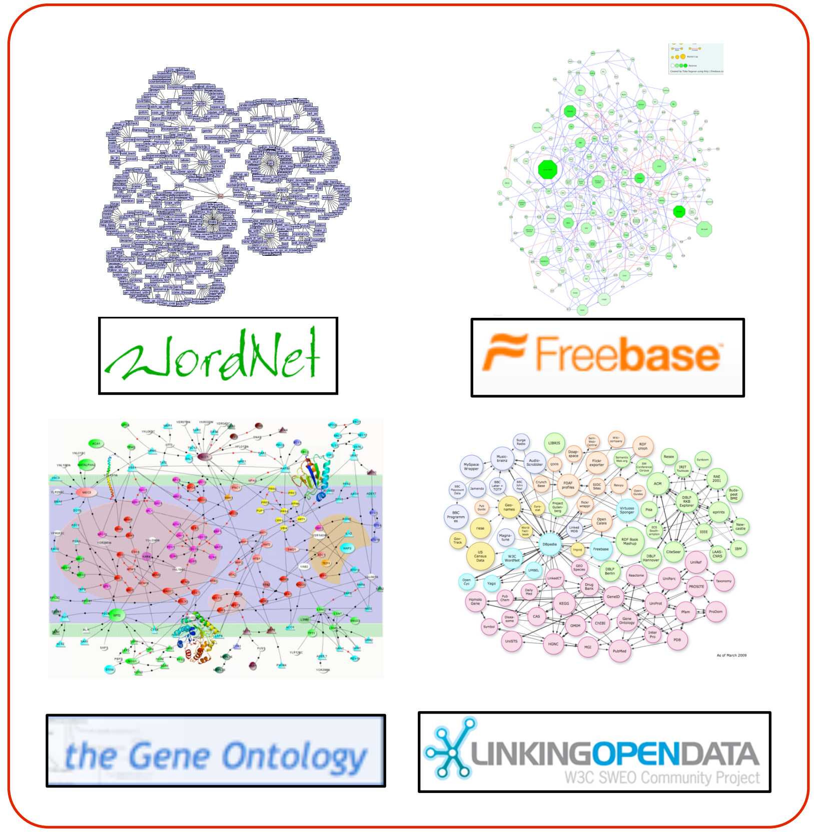
LINKINGOPENDATA
W3C SWEO Community Project

slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge

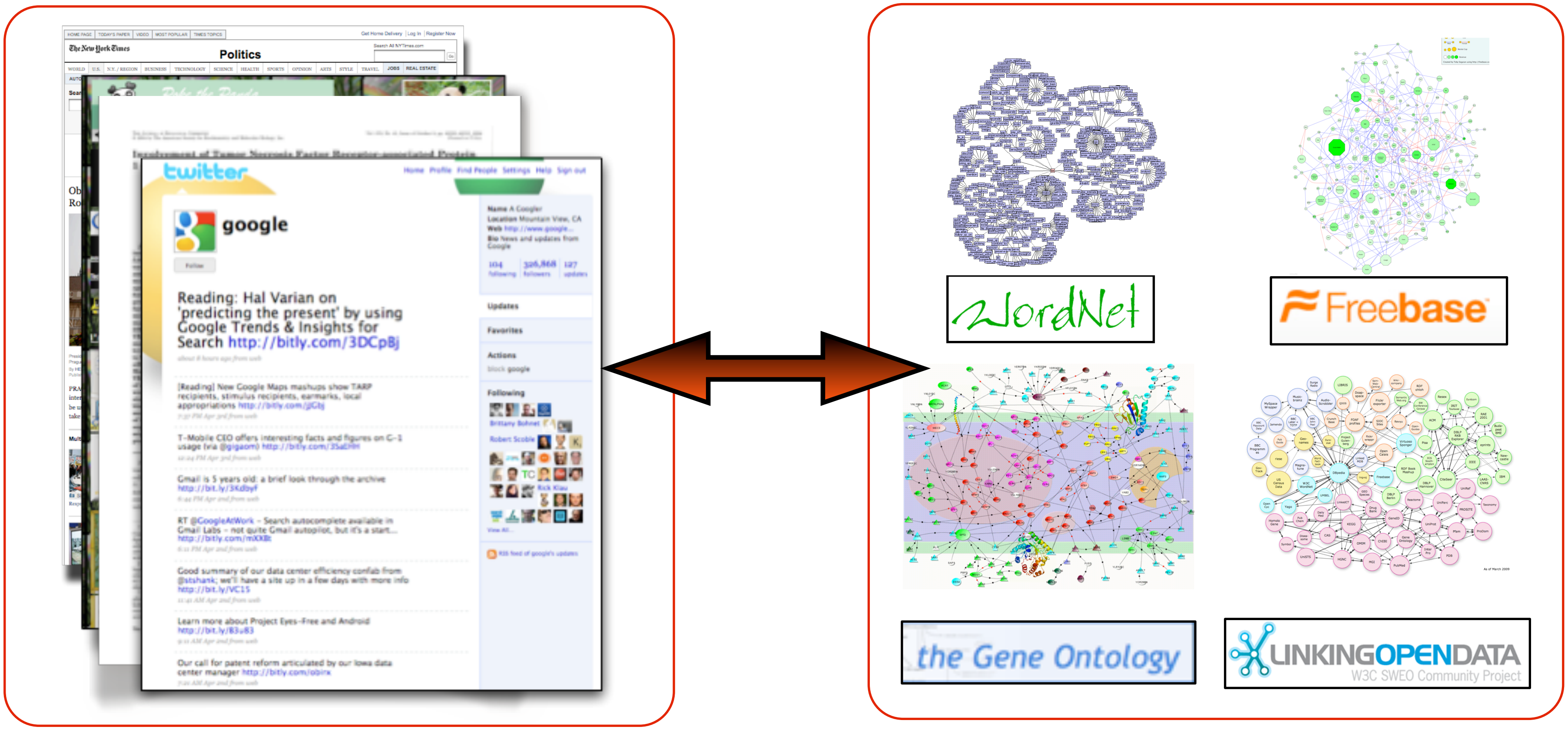


slide from Rion Snow

From Unstructured Text to Structured Knowledge

Unstructured Text

Structured Knowledge



slide from Rion Snow

More applications of IE?

More applications of IE

- Building & extending knowledge bases and ontologies
- Scholarly literature databases: Google Scholar, CiteSeerX
- People directories: Rapleaf, Spoke, Naymz
- Shopping engines & product search
- Bioinformatics: clinical outcomes, gene interactions, …
- Patent analysis
- Stock analysis: deals, acquisitions, earnings, hirings & firings
- SEC filings
- Intelligence analysis for business & government

Google Squared

Web Images Video Maps News Shopping Gmail more ▾ | Send us feedback | Help | Sign in

Google squared labs — roller coasters — Square it — Add to this Square

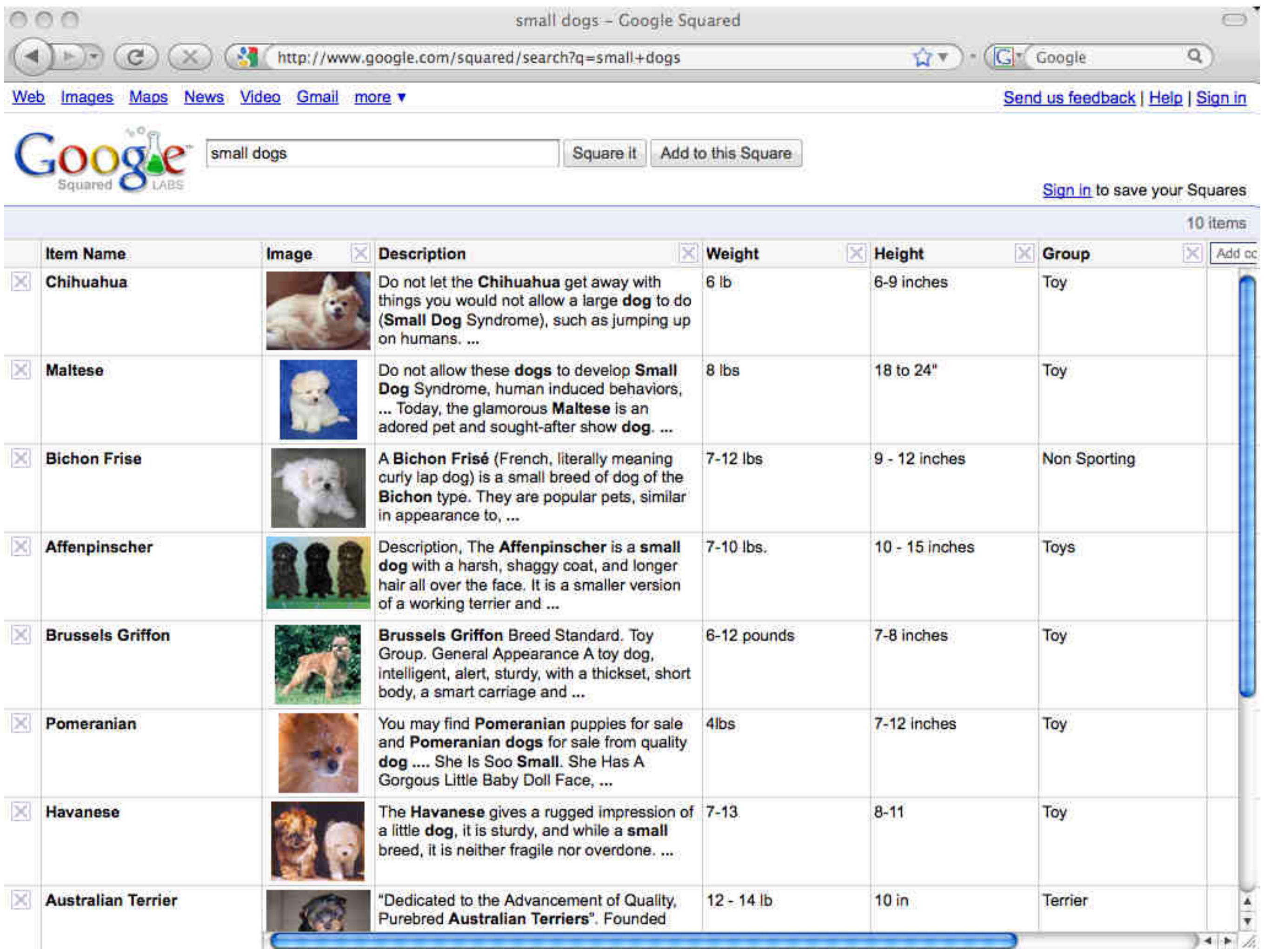
Sign in to save your Squares

roller coasters — 7 items

	Item Name	Image	Description	Capacity	Height	Speed	Add co
	Millennium Force		**Millennium Force** is a high-speed giga roller coaster. Guests must be in good health to ride this ride. Guests with disabilities are mainstreamed through the ...	1,600 riders per hour	310 ft	93 mph	
	Texas Giant		At more than 14 stories tall, the **Texas Giant** is one of the tallest, fastest wooden roller coasters to be found anywhere. The **Giant** was named the #1 roller ...	1600 riders per hour	143 ft	62-MPH	
	Cyclone		For other roller coasters named **Cyclone**, see **Cyclone** (disambiguation). .... The Coney Island **Cyclone** is an ACE Coaster Classic and Coaster Landmark; ...	1200 riders per hour	85 feet	60-MPH	
	Mean Streak		The **Mean Streak** is a high-speed wooden roller coaster. The lap bar and seatbelt must be fastened and tightened securely. Special access is via the exit ramp ...	1,600 riders per hour	161 ft	65 mph	
	SuperMan The Escape		**Superman: The Escape** is a launched shuttle roller coaster located in the Samurai Summit area of Six Flags Magic Mountain in Valencia, California that opened ...	1050 riders per hour	415 feet	100 mph	
	Riddler's Revenge		This article is about the Six Flags Magic Mountain roller coaster. For the episode of The Batman, see **Riddler's Revenge** (The Batman). ...	No value found	156'	65 mph	
	Alpengeist		**Alpengeist** is the tallest and one of the fastest full circuit inverted roller coasters in the world. **Alpengeist** opened in 1997 at Busch Gardens Williamsburg ...	1820 riders per hour	195 ft	67 mph	
	Add items — Add	or Add next 10 items					

Not finding the right items? Start with an empty Square.

Google Squared



Item Name	Image	Description	Weight	Height	Group
Chihuahua		Do not let the **Chihuahua** get away with things you would not allow a large **dog** to do (**Small Dog** Syndrome), such as jumping up on humans. ...	6 lb	6-9 inches	Toy
Maltese		Do not allow these **dogs** to develop **Small Dog** Syndrome, human induced behaviors, ... Today, the glamorous **Maltese** is an adored pet and sought-after show **dog**. ...	8 lbs	18 to 24"	Toy
Bichon Frise		A **Bichon Frisé** (French, literally meaning curly lap dog) is a small breed of dog of the **Bichon** type. They are popular pets, similar in appearance to, ...	7-12 lbs	9 - 12 inches	Non Sporting
Affenpinscher		Description, The **Affenpinscher** is a **small dog** with a harsh, shaggy coat, and longer hair all over the face. It is a smaller version of a working terrier and ...	7-10 lbs.	10 - 15 inches	Toys
Brussels Griffon		**Brussels Griffon** Breed Standard. Toy Group. General Appearance A toy dog, intelligent, alert, sturdy, with a thickset, short body, a smart carriage and ...	6-12 pounds	7-8 inches	Toy
Pomeranian		You may find **Pomeranian** puppies for sale and **Pomeranian dogs** for sale from quality **dog** .... She Is Soo **Small**. She Has A Gorgous Little Baby Doll Face, ...	4lbs	7-12 inches	Toy
Havanese		The **Havanese** gives a rugged impression of a little **dog**, it is sturdy, and while a **small** breed, it is neither fragile nor overdone. ...	7-13	8-11	Toy
Australian Terrier		"Dedicated to the Advancement of Quality, Purebred **Australian Terriers**". Founded	12 - 14 lb	10 in	Terrier

Named Entity Recognition

- Labeling names of things in web pages:
  - An entity is a discrete thing like “IBM Corporation”
    - But often extended in practice to things like dates, instances of products and chemical/biological substances that aren’t really entities…
  - “Named” means called “IBM” or “Big Blue” not “it”
- E.g.,
  - Many web pages tag various entities
  - “Smart Tags” (Microsoft) inside documents
  - Reuters’ OpenCalais

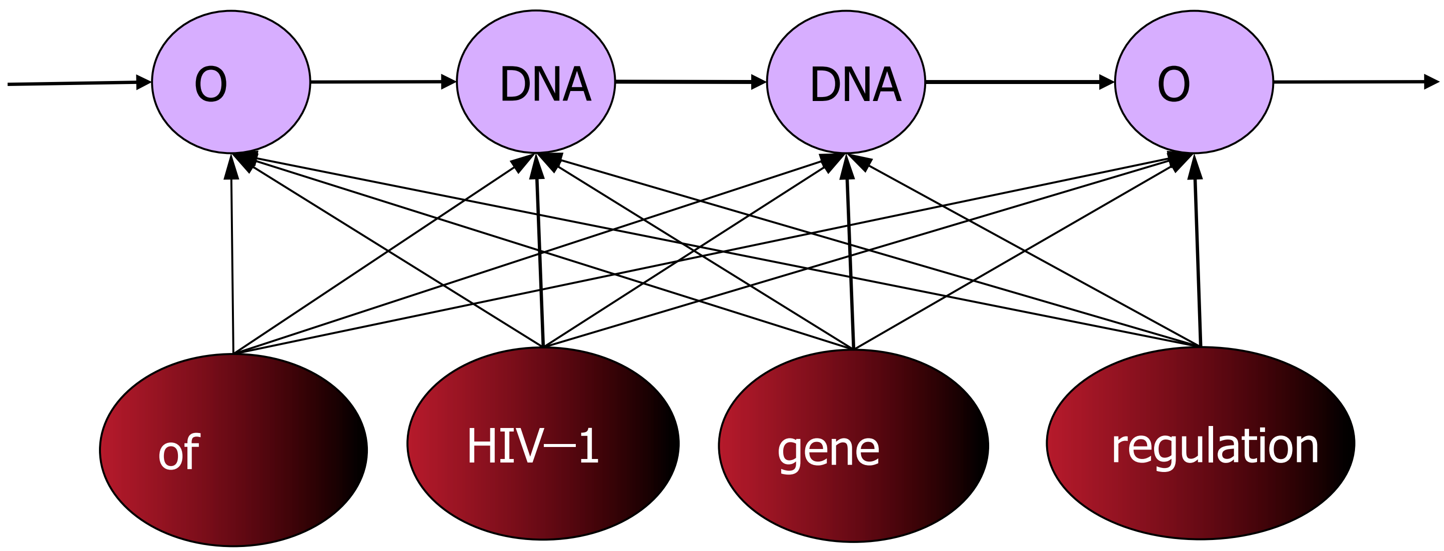
Named Entity Extraction

- The task: find and classify names in text, for example:

```
The European Commission [ORG] said on Thursday it disagreed
with German [MISC] advice.

Only France [LOC] and Britain [LOC] backed Fischler [PER]
's proposal .

"What we have to be extremely careful of is how other
countries are going to take Germany 's lead", Welsh
National Farmers ' Union [ORG] ( NFU [ORG] ) chairman John
Lloyd Jones [PER] said on BBC [ORG] radio .
```

- The purpose:
  - … a lot of information is really associations between named entities.
  - … for question answering, answers are usually named entities.
  - … the same techniques apply to other slot-filling classifications.

Maximum Entropy Markov Model



$$P(t \mid h) = \frac{\exp(\sum_{j=1}^{m} f_j(h,t)\lambda_j)}{\sum_{k=1}^{K} \exp(\sum_{j=1}^{m} f_j(h,t_k)\lambda_j)}$$

Slide from Chris Manning

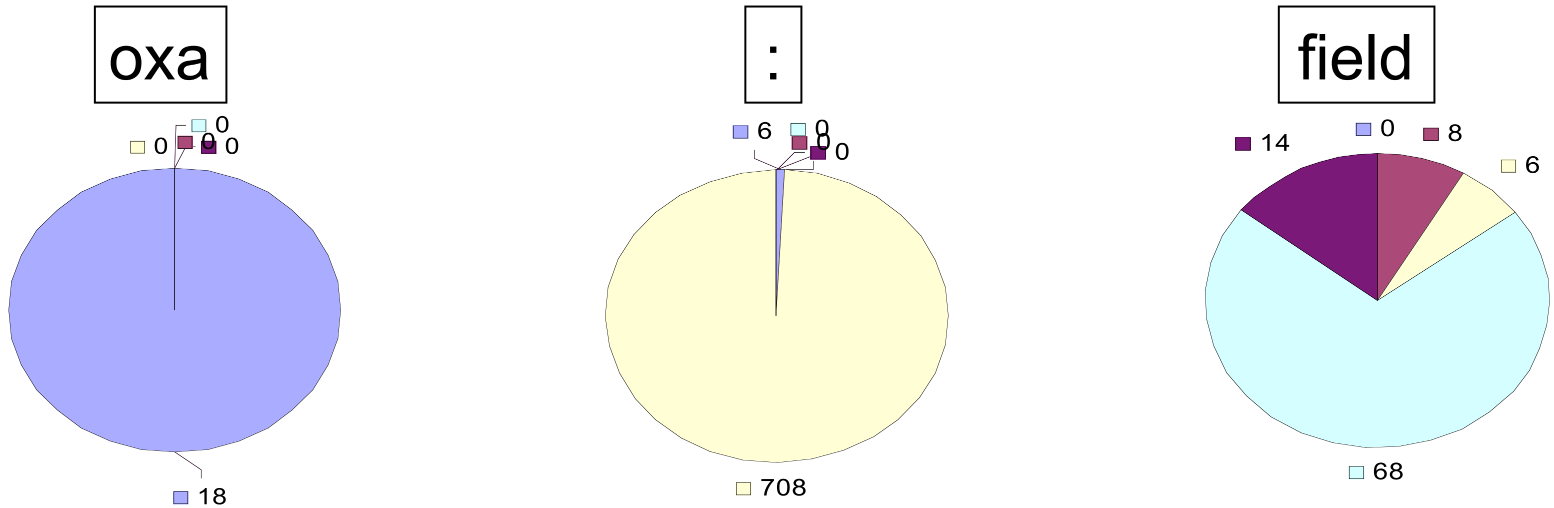
Interesting Features

- Words
- Word shapes
- Part−of−speech tags
- Parsing information
- Searching the web for the word in a given context
  - *X gene*, *X mutation*, *X antagonist*
- Gazetteer
  - list words whose classification is known
- Abbreviation extraction (Schwartz and Hearst, 2003)
  - Identify short and long forms when occurring together in text

… Zn finger homeodomain 2 (Zfh 2) …

Slide from Chris Manning

Orthographic (letter *n*-gram) features: what's in a name?



drug
company
movie
place
person

Cotrimoxazole

Wethersfield

Alien Fury: Countdown to Invasion

Slide from Chris Manning

Named entity recognition results

- NER is commonly thought of as a *solved* problem
- Accuracies of >90% are typical
  - (But very genre-dependent: BioMed NER is *much* harder)
- NER isn't usually considered part of NLU
- Reminiscent of "The AI Effect":

  "Every time we figure out a piece of it, it stops being magical; we say, *Oh, that's just a computation.*" —Rodney Brooks

Relation extraction example

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Relation types

For generic news texts...

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER $\longrightarrow$ PER
	Organizational	*spokesman for, president of*	PER $\longrightarrow$ ORG
	Artifactual	*owns, invented, produces*	(PER $\mid$ ORG) $\longrightarrow$ ART
Geospatial			
	Proximity	*near, on outskirts*	LOC $\longrightarrow$ LOC
	Directional	*southeast of*	LOC $\longrightarrow$ LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG $\longrightarrow$ ORG
	Political	*annexed, acquired*	GPE $\longrightarrow$ GPE

Slide from Jim Martin

Types of ACE Relations, 2003

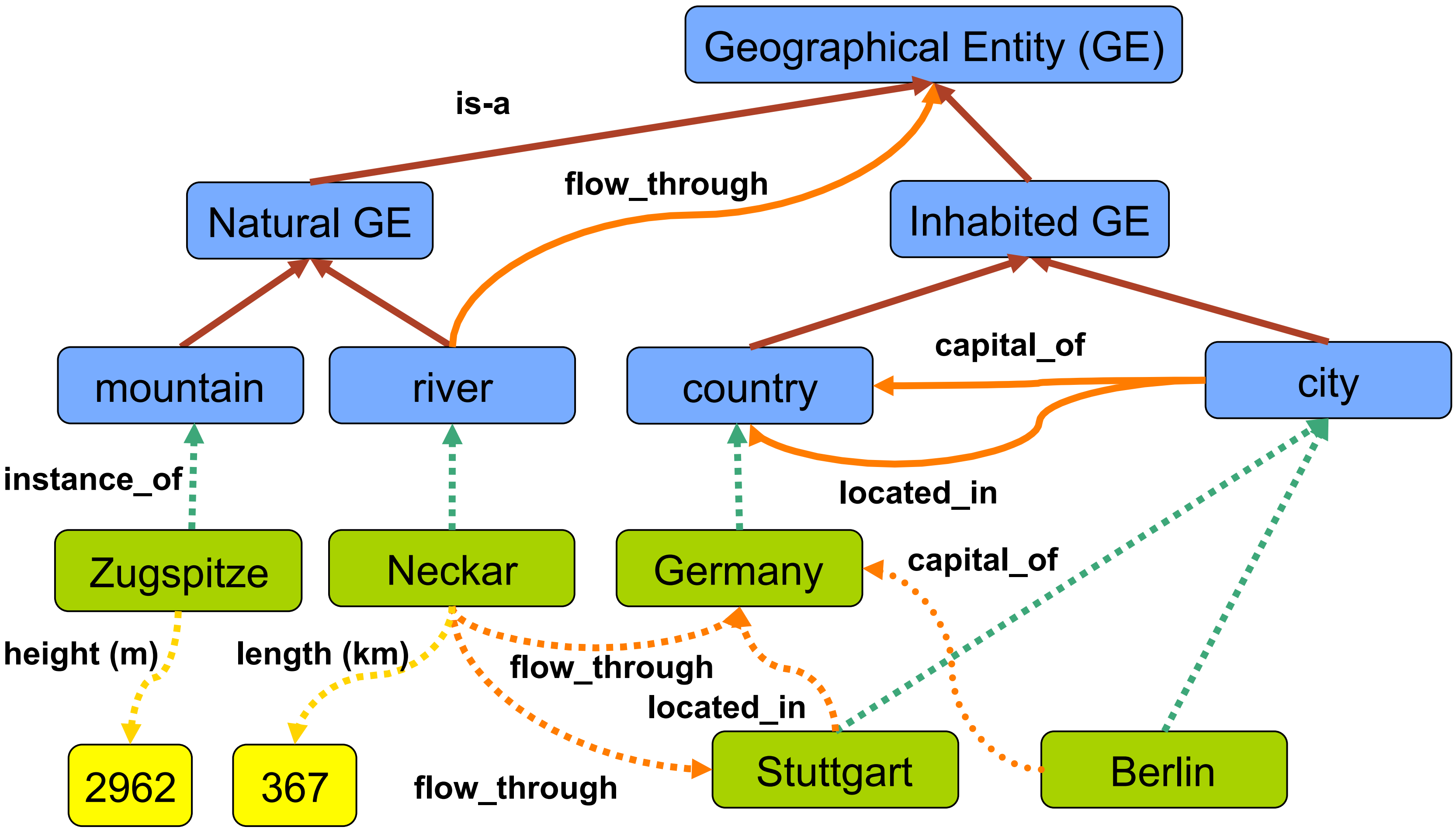
- **ROLE** - relates a person to an organization or a geopolitical entity
  - Subtypes: **member**, **owner**, **affiliate**, **client**, **citizen**
- **PART** - generalized containment
  - Subtypes: subsidiary, physical part-of, set **membership**
- **AT** - permanent and transient locations
  - Subtypes: **located**, **based-in**, **residence**
- **SOCIAL**- social relations among persons
  - Subtypes: **parent**, **sibling**, **spouse**, **grandparent**, **associate**

Slide from Doug Appelt

Frequent Freebase Relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Relations in ontologies: geographical



Design: Philipp Cimiano

Slide from Paul Buitelaar

Other relations: disease outbreaks

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly **Ebola** epidemic in **Zaire**, is finding itself hard pressed to cope with the crisis…

Information Extraction System (e.g., NYU's Proteus)

Disease Outbreaks in ***The New York Times***

Date	***Disease Name***	***Location***
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Slide from Eugene Agichtein

Other relations: protein interactions

„We show that CBF-A and CBF-C interact with each other to form a CBF-A-CBF-C complex and that CBF-B does not interact with CBF-A or CBF-C individually but that it associates with the CBF-A-CBF-C complex."

$$\text{CBF-A} \xleftrightarrow[\text{complex}]{\text{interact}} \text{CBF-C}$$

$$\text{CBF-B} \xrightarrow{\text{associates}} \text{CBF-A-CBF-C complex}$$

Slide from Rosario and Hearst

Other relations: UMLS

- **Unified Medical Language System**
  - integrates linguistic, terminological and semantic information
  - Semantic Network consists of 134 semantic types and 54 relations between types

Pharmacologic Substance	affects	Pathologic Function
Pharmacologic Substance	causes	Pathologic Function
Pharmacologic Substance	complicates	Pathologic Function
Pharmacologic Substance	diagnoses	Pathologic Function
Pharmacologic Substance	prevents	Pathologic Function
Pharmacologic Substance	treats	Pathologic Function

Slide from Paul Buitelaar

Relations in ontologies: GO (Gene Ontology)

- **GO (Gene Ontology)**
  - Aligns descriptions of gene products in different databases, including plant, animal and microbial genomes
  - Organizing principles are molecular function, biological process and cellular component

Accession:	GO:0009292
Ontology:	biological process
Synonyms:	broad: genetic exchange
Definition:	In the absence of a sexual life cycle, the processes involved in the introduction of genetic information to create a genetically different individual.
Term Lineage	all : all (164142)
	GO:0008150 : biological process (115947)
	GO:0007275 : development (11892)
	GO:0009292 : genetic transfer (69)

Why this is hard: Ambiguity!

Which relations hold between two entities?

Treatment

Cure?

Prevent?

Side Effect?

Disease

Relations between disease & treatment

- Cure

  These results suggest that con A-induced hepatitis was ameliorated by pretreatment with TJ-135.

- Prevent

  A two-dose combined hepatitis A and B vaccine would facilitate immunization programs.

- Vague

  … effect of interferon on hepatitis B.

Slide from Barbara Rosario and Marti Hearst

Relations between words

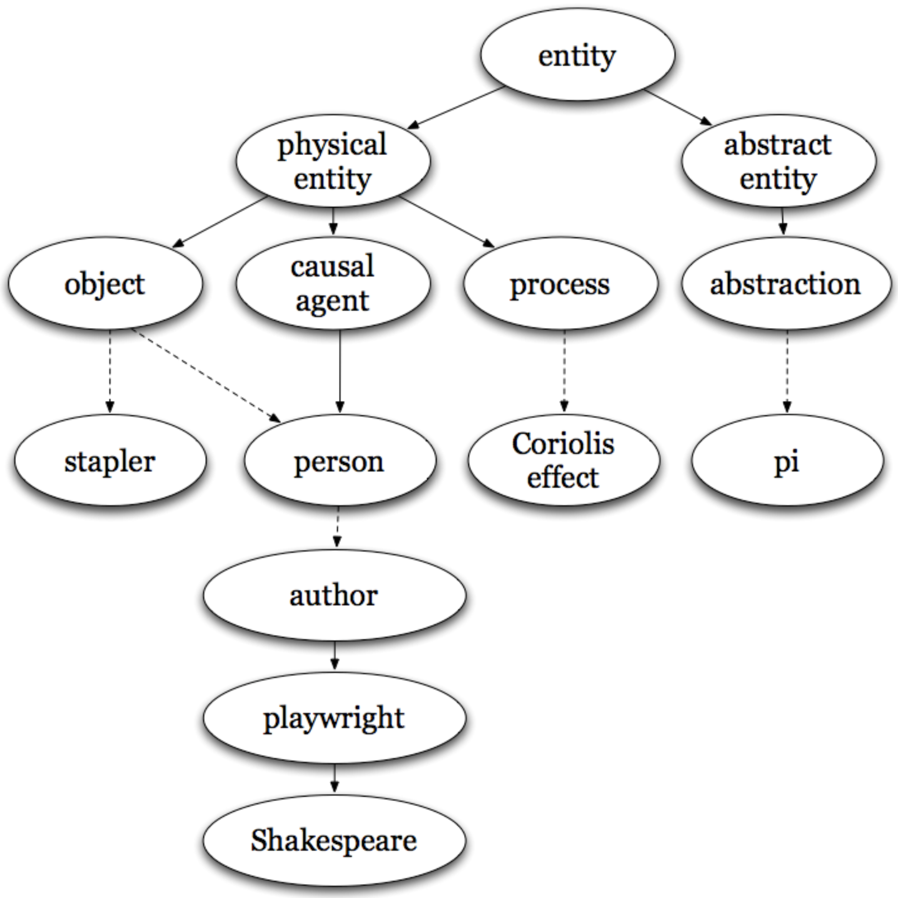
- Language understanding applications need word meaning!
  - Question answering
  - Conversational agents
  - Summarization
- One key meaning component: word relations
  - Hierarchical (hypernym/hyponym) relations
    - "San Francisco" is a "city"
  - Other relations between words
    - "alternator" is a part of a "car"

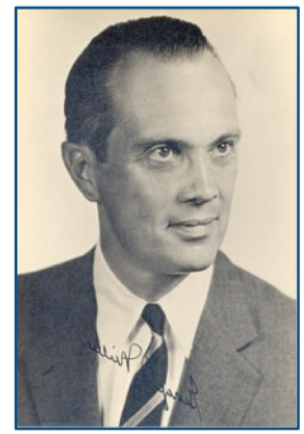
Hyponymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

superordinate	vehicle	fruit	furniture	mammal
hyponym	car	mango	chair	dog

The WordNet noun hierarchy





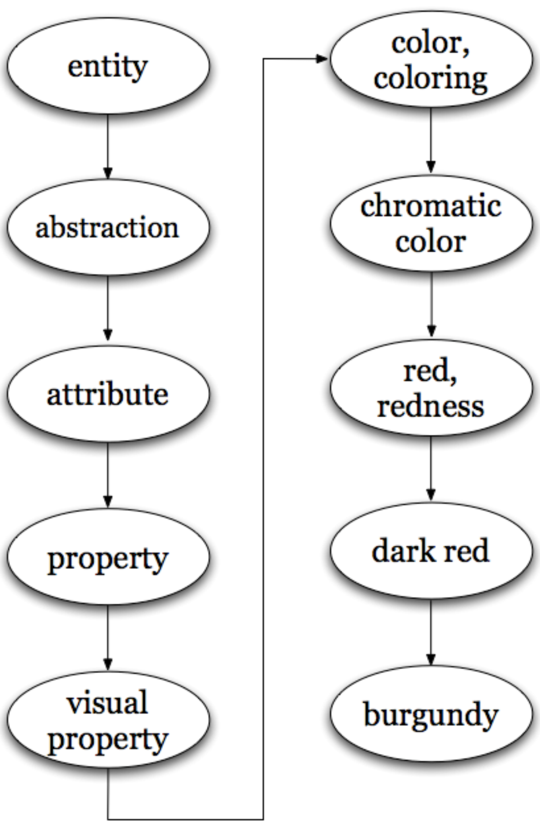
Properties:
Transitive, Acyclic

http://wordnetweb.princeton.edu/perl/webwn

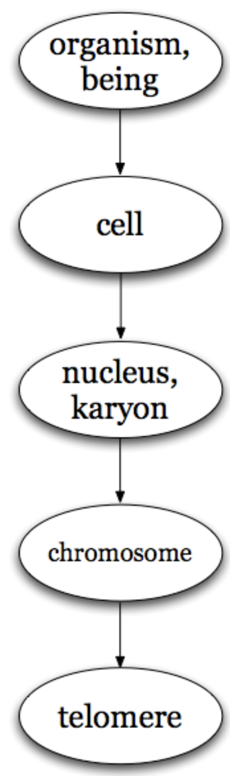
slide from Rion Snow

WordNet relations

X is-a-kind-of Y
(hyponym / hypernym)



X is-a-part-of Y
(meronym / holonym)



slide from Rion Snow

WordNet Noun Relations

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Instance Hypernym	Instance	From instances to their concepts	$Austen^1 \rightarrow author^1$
Instance Hyponym	Has-Instance	From concepts to concept instances	$composer^1 \rightarrow Bach^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Substance Meronym		From substances to their subparts	$water^1 \rightarrow oxygen^1$
Substance Holonym		From parts of substances to wholes	$gin^1 \rightarrow martini^1$
Antonym		Semantic opposition between lemmas	$leader^1 \Longleftrightarrow follower^1$
Derivationally Related Form		Lemmas w/same morphological root	$destruction^1 \Longleftrightarrow destroy^1$

WordNet is incomplete

Ontological relations are missing for many words:

In WordNet 3.1	Not in WordNet 3.1
insulin progesterone	leptin pregnenolone
combustibility navigability	affordability reusability
HTML	XML
Google, Yahoo	Microsoft, IBM

Esp. for specific domains: restaurants, auto parts, finance

Relation extraction: 5 easy methods

1. Hand-built patterns
2. Supervised methods
3. Bootstrapping (seed) methods
4. Unsupervised methods
5. Distant supervision

Relation extraction: 5 easy methods

1. **Hand-built patterns**
2. Supervised methods
3. Bootstrapping (seed) methods
4. Unsupervised methods
5. Distant supervision

A complex hand-built extraction rule

```
;;; For <company> appoints <person> <position>

(defpattern appoint
   "np-sem(C-company)?  rn?  sa?  vg(C-appoint) np-sem(C-person) ´,´?
    to-be?  np(C-position) to-succeed?:
    company-at=1.attributes, sa=3.span, lv=4.span, person-at=5.attributes
    position-at=8.attributes |
...
```

```
(defun when-appoint (phrase-type)
    (let ((person-at (binding ´person-at))
        (company-entity (entity-bound ´company-at))
        (person-entity (essential-entity-bound ´person-at ´C-person))
        (position-entity (entity-bound ´position-at))
        (predecessor-entity (entity-bound ´predecessor-at))
        new-event)
     (not-an-antecedent position-entity)
     ;; if no company is specified for position, use agent
...
```

NYU Proteus

Problems

- Have to write many new rules for each possible relation
  - hard to write
  - hard to maintain
  - there are a zillion of them
  - domain-dependent
- Can we do something more general?

Adding hyponyms to WordNet

- Intuition from Hearst (1992)
  - “Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”
- What does *Gelidium* mean?
- How do you know?`

Adding hyponyms to WordNet

- Intuition from Hearst (1992)
  - "Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use"
- What does *Gelidium* mean?
- How do you know?`

Predicting the hyponym relation

"...works by such **authors** as Herrick, Goldsmith, and **Shakespeare**."

"If you consider **authors** like **Shakespeare**..."

"Some **authors** (including **Shakespeare**)..."

"**Shakespeare** was the **author** of several..."

"**Shakespeare**, **author** of *The Tempest...*"

↓

Shakespeare IS-A ***author*** (0.87)

How can we capture the variability of expression of a relation in natural text from a large, unannotated corpus?

Hearst's lexico-syntactic patterns

"Y such as X ((, X)* (, and/or) X)"
"such Y as X…"
"X… or other Y"
"X… and other Y"
"Y including X…"
"Y, especially X…"

(Hearst, 1992): Automatic Acquisition of Hyponyms

Examples of Hearst patterns

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

Patterns for detecting part-whole relations (meronym-holonym)

Berland and Charniak (1999)


Berland pattern	**Example occurrences**
NP_Y's NP_X:	...building's basement...
NP_X of {the|a} NP_Y:	...basement of a building...
NP_X in {the|a} NP_X:	...basements in a building...
NP_X of NP_Y:	...basements of buildings...
NP_X in NP_Y:	...basements in buildings...

Results with hand-built patterns

- Hearst: hypernyms
  - 66% precision with "X and other Y" patterns
- Berland & Charniak: meronyms
  - 55% precision

Problem with hand-built patterns

- Requires that we hand-build patterns for each relation!
- Don't want to have to do this for all possible relations!
- Plus, we'd like better accuracy

Relation extraction: 5 easy methods

1. Hand-built patterns
2. **Supervised methods**
3. Bootstrapping (seed) methods
4. Unsupervised methods
5. Distant supervision

Supervised relation extraction

- Sometimes done in 3 steps:
  1. Find all pairs of named entities
  2. Decide if the two entities are related
  3. If yes, then classify the relation
- Why the extra step?
  - Cuts down on training time for classification by eliminating most pairs
  - Producing separate feature-sets that are appropriate for each task

Relation analysis

- Usually just run on named entities within the same sentence

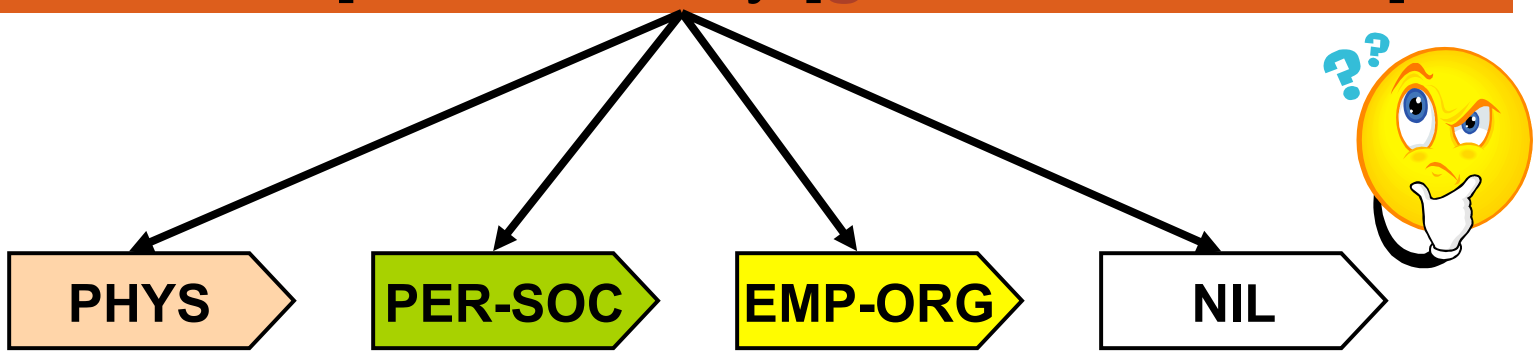
```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

Slide from Jim Martin

Relation extraction

- Task definition: to label the semantic relation between a pair of entities in a sentence (fragment)

…[leader arg-1] of a minority [government arg-2]…

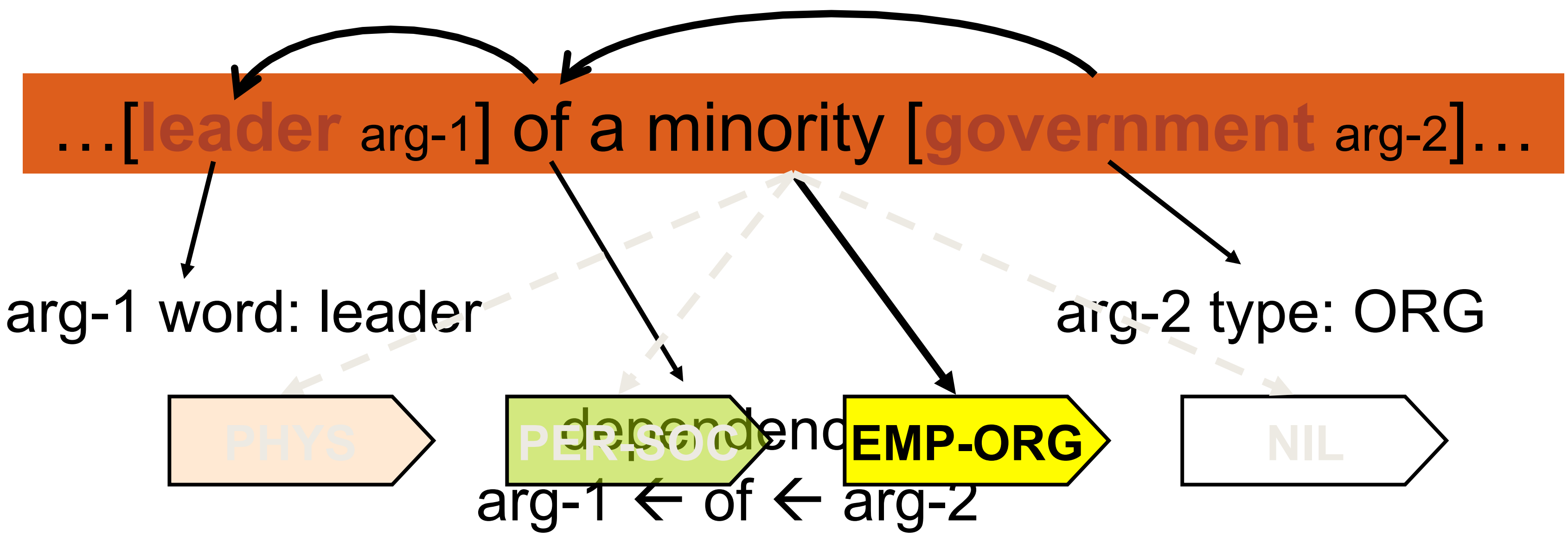


PHYS: Physical
PER-SOC: Personal / Social
EMP-ORG: Employment / Membership / Subsidiary

Slide from Jing Jiang

Supervised learning

- Supervised machine learning (e.g. [Zhou et al. 2005], [Bunescu & Mooney 2005], [Zhang et al. 2006], [Surdeanu & Ciaramita 2007])



- Training data is needed for each relation type

Slide from Jing Jiang

ACE 2008 tasks

- EDR (Entity Detection and Recognition)
  - within-document (“local”)
  - cross-document (“global”)
- RDR (Relation Detection and Recognition)
  - within-document (“local”)
  - cross-document (“global”)

ACE 2008

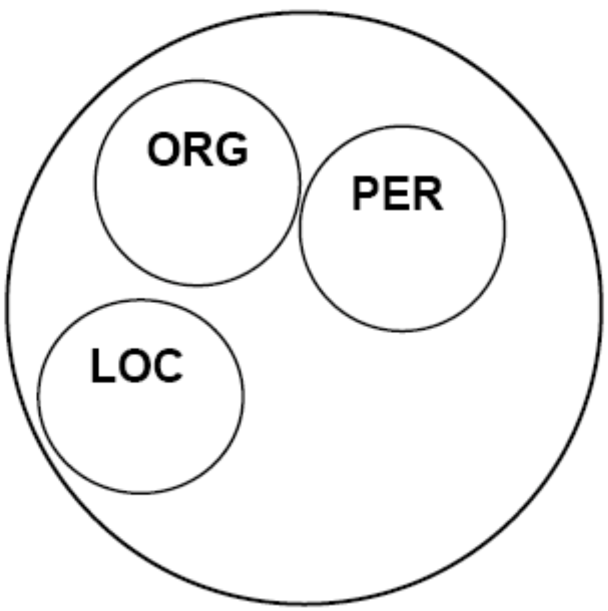
- An **entity** is an object or set of objects in the world.
- A **mention** is a reference to an entity.
  - Name Mention***: Joe Smith***
  - Nominal Mention***: the guy wearing a blue shirt***
  - Pronoun Mentions***: he, him***

ACE 2008: five entity types

- Person (PER) - Human individual or group.
  - PER.Individual [Bill Clinton], [The President of the U.S.]
  - PER.Group: [Analysts], [IBM's lawyers] [the house painters]
- Organization (ORG) - Corporation, agencies, etc. groups
  - ORG.GOV: *[KGB], [the administration]*
  - *ORG.COM, ORG.EDU, ORG.NONGOV "The Red Cross"*
  - *ORG.REL, ORG.SCI, ORG.SPO*
  - *ORG.ENT: [the Roundabout Theater Company]*

ACE 2008: five entity types

- Geo-political Entity (GPE) - GPE entities are geographical regions defined by political and/or social groups
  - NATION, CONTINENT, STATE, POPCENTER, etc
  - [France], [The people of France]



- **GPE.ORG** - France signed a treaty with Germany last week.
- **GPE.PER** - France vacations in August.
- **GPE.LOC** - The world leaders met in France yesterday.
- **GPE.GPE** - France produces better wine than New Jersey.

ACE 2008: five entity types

- Location (LOC) - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
  - ADDRESS, BOUNDARY, CELESTIAL, WATER-BODY, LAND-REGION-NATURAL, REGION-GENERAL
- Facility (FAC) - Buildings and other permanent man-made structures
  - AIRPORT, PLANT, PATH (street, bridge), etc.

ACE 2008: EDR

- For each entity, all mentions of the entity are recorded and coreferenced

ACE 2008: six relation types

Type	**Subtype**
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General 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-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

ACE Agent-Artifact Relation

- User-Owner-Inventor-Manufacturer

PER-FAC

[My house] is in West Philadelphia			
Class	***Type***	***Argument 1***	***Argument 2***
Possessive Asserted Unspecified	*Agent-Artifact.UOIM*	***My***	*My* ***house***

ACE General-Affiliation Relation

- Citizen-Resident-Religion-Ethnicity

PER-GPE

a sheep shearer from New Zealand			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Gen-Aff.CRRE*	*a sheep **shearer** from New Zealand*	***New Zealand***

- Org-Location-Origin

ORG-LOC

a small robotics company in a St. Louis suburb			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Gen-Aff.Loc-Origin*	*a small robotics **company** in a St. Louis suburb*	*a St. Louis **suburb***

ACE ORG-Affiliation Relation

- Employment

PER-ORG

the CEO of Microsoft			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Org-Aff.Employment*	*the* ***CEO*** *of Microsoft*	***Microsoft***

- Owner

PER-ORG

[Dallas Cowboys owner] Jerry Jones			
Class	***Type***	***Argument 1***	***Argument 2***
PreMod Asserted Unspecified	*Org-Aff.Ownership*	*Dallas Cowboys* ***owner***	***Dallas Cowboys***

- + Founder, Membership, Sports-Affiliation, Shareholder

ACE Part-Whole Relation

- GEO

FAC-FAC

St. Vartan's Cathedral, on Second Avenue			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Part-Whole.Geo*	***St. Vartan's Cathedral**, on Second Avenue*	***Second Avenue***

- SUBSIDIARY

ORG-ORG

Microsoft's accounting department			
Class	***Type***	***Argument 1***	***Argument 2***
Possessive Asserted Unspecified	*Part-Whole.Subsidiary*	*Microsoft's accounting **department***	***Microsoft***

ACE Personal-Social Relation

- Business

PER-PER

his lawyer			
Class	***Type***	***Argument 1***	***Argument 2***
Possessive Asserted Unspecified	*Per-Social.Business*	***his***	*his **lawyer***

- Family

PER-PER

relatives of the dead			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Per-Social.Family*	***relatives** of the dead*	*the **dead***

- Lasting

PER-PER

his friendship with some right-wing mayors			
Class	***Type***	***Argument 1***	***Argument 2***
Possessive Asserted Unspecified	*Per-Social.Lasting*	***his***	*some right-wing **mayors***

ACE Physical Relation

- LOCATED

PER-GPE

He was campaigning in his home state of Tennessee			
Class	***Type***	***Argument 1***	***Argument 2***
Verbal Asserted Past	*Physical.Located*	***He***	*his home **state** of Tennessee*

- NEAR

GPE-GPE

a town some 50 miles south of Salzburg in the central Austrian Alps			
Class	***Type***	***Argument 1***	***Argument 2***
Preposition Asserted Unspecified	*Physical.Near*	*a **town** some 50 miles south of Salzburg in the central Austrian Alps*	***Salzburg***

PER-FAC

Muslim youths recently staged a half dozen rallies in front of the embassy			
Class	***Type***	***Argument 1***	***Argument 2***
Other Asserted Past	*Physical.Near*	*Muslim **youths***	*the **embassy***

ACE 2008 Training data

Source	Training epoch	Approximate size
English Resources		
Broadcast News	3/03 – 6/03	55,000 words
Broadcast Conversations	3/03 – 6/03	40,000 words
Newswire	3/03 – 6/03	50,000 words
Weblog	11/04 – 2/05	40,000 words
Usenet	11/04 – 2/05	40,000 words
Conversational Telephone Speech	11/04-12/04 (differentiated by topic vs. eval)	40,000 words
Arabic Resources		
Broadcast News	10/00 – 12/00	30,000+ words
Newswire	10/00 – 12/00	55,000+ words
Weblog	11/04 – 2/05	20,000+ words

Features: words

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Bag-of-words features

WM1 = {American, Airlines}, WM2 = {Tim, Wagner}

Head-word features

HM1 = Airlines, HM2 = Wagner, HM12 = Airlines+Wagner

Words in between

WBNULL = false, WBFL = NULL, WBF = a, WBL = spokesman,
WBO = {unit, of, AMR, immediately, matched, the, move}

Words before and after

BM1F = NULL, BM1L = NULL, AM2F = said, AM2L = NULL

Word features yield good precision, but poor recall

Features: NE type & mention level

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Named entity types (ORG, LOC, PER, etc.)

ET1 = ORG, ET2 = PER, ET12 = ORG-PER

Mention levels (NAME, NOMINAL, or PRONOUN)

ML1 = NAME, ML2 = NAME, ML12 = NAME+NAME

Named entity type features help recall a lot
Mention level features have little impact

Features: overlap

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Number of mentions and words in between

#MB = 1, #WB = 9

Does one mention include in the other?

M1>M2 = false, M1<M2 = false

Conjunctive features

ET12+M1>M2 = ORG-PER+false
ET12+M1<M2 = ORG-PER+false
HM12+M1>M2 = Airlines+Wagner+false
HM12+M1<M2 = Airlines+Wagner+false

These features hurt precision a lot, but also help recall a lot

Features: base phrase chunking

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Parse using the Stanford Parser, then apply Sabine Buchholz's chunklink.pl:

```
 0 B-NP    NNP   American     NOFUNC   Airlines    1 B-S/B-S/B-NP/B-NP
 1 I-NP    NNPS  Airlines     NP       matched     9 I-S/I-S/I-NP/I-NP
 2 O       COMMA COMMA        NOFUNC   Airlines    1 I-S/I-S/I-NP
 3 B-NP    DT    a            NOFUNC   unit        4 I-S/I-S/I-NP/B-NP/B-NP
 4 I-NP    NN    unit         NP       Airlines    1 I-S/I-S/I-NP/I-NP/I-NP
 5 B-PP    IN    of           PP       unit        4 I-S/I-S/I-NP/I-NP/B-PP
 6 B-NP    NNP   AMR          NP       of          5 I-S/I-S/I-NP/I-NP/I-PP/B-NP
 7 O       COMMA COMMA        NOFUNC   Airlines    1 I-S/I-S/I-NP
 8 B-ADVP  RB    immediately  ADVP     matched     9 I-S/I-S/B-ADVP
 9 B-VP    VBD   matched      VP/S     matched     9 I-S/I-S/B-VP
10 B-NP    DT    the          NOFUNC   move       11 I-S/I-S/I-VP/B-NP
11 I-NP    NN    move         NP       matched     9 I-S/I-S/I-VP/I-NP
12 O       COMMA COMMA        NOFUNC   matched     9 I-S
13 B-NP    NN    spokesman    NOFUNC   Wagner     15 I-S/B-NP
14 I-NP    NNP   Tim          NOFUNC   Wagner     15 I-S/I-NP
15 I-NP    NNP   Wagner       NP       matched     9 I-S/I-NP
16 B-VP    VBD   said         VP       matched     9 I-S/B-VP
17 O       .     .            NOFUNC   matched     9 I-S
```

[$_{NP}$ American Airlines], [$_{NP}$ a unit] [$_{PP}$ of] [$_{NP}$ AMR], [$_{ADVP}$ immediately] [$_{VP}$ matched] [$_{NP}$ the move], [$_{NP}$ spokesman Tim Wagner] [$_{VP}$ said].

Features: base phrase chunking

[$_{NP}$ American Airlines], [$_{NP}$ a unit] [$_{PP}$ of] [$_{NP}$ AMR], [$_{ADVP}$ immediately] [$_{VP}$ matched] [$_{NP}$ the move], [$_{NP}$ spokesman Tim Wagner] [$_{VP}$ said].

Phrase heads before and after

CPHBM1F = NULL, CPHBM1L = NULL, CPHAM2F = said, CPHAM2L = NULL

Phrase heads in between

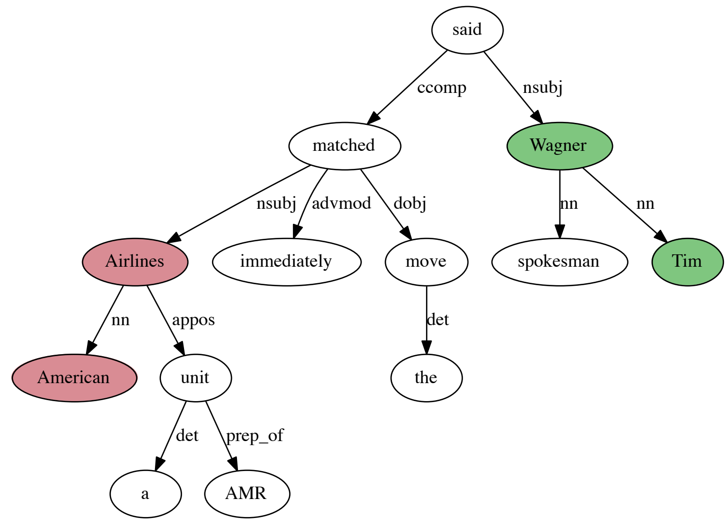
CPHBNULL = false, CPHBFL = NULL, CPHBF = unit, CPHBL = move
CPHBO = {of, AMR, immediately, matched}

Phrase label paths

CPP = [NP, PP, NP, ADVP, VP, NP]
CPPH = NULL

These features increased both precision & recall by 4-6%

Features: syntactic features



Features of mention dependencies

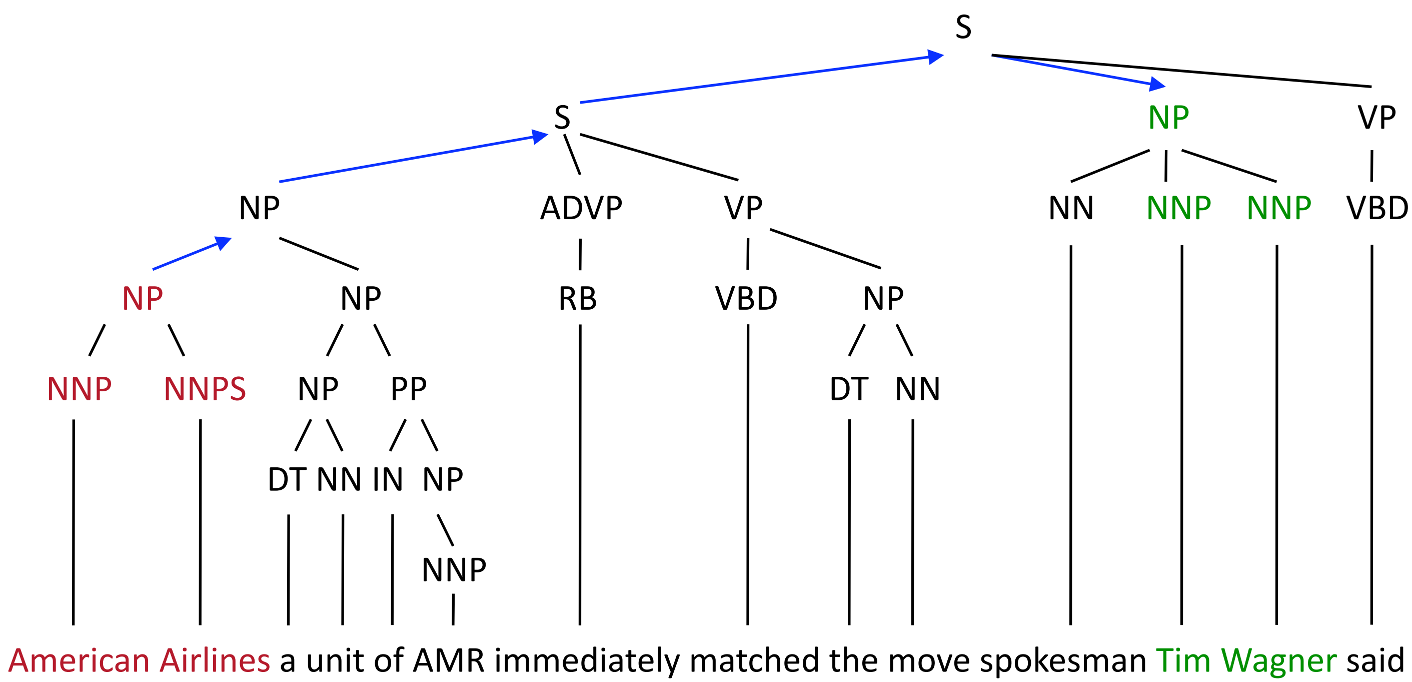
ET1DW1 = ORG:Airlines
H1DW1 = matched:Airlines
ET2DW2 = PER:Wagner
H2DW2 = said:Wagner

Features describing entity types and dependency tree

ET12SameNP = ORG-PER-false
ET12SamePP = ORG-PER-false
ET12SameVP = ORG-PER-false

These features had disappointingly little impact!

Features: syntactic features



Phrase label paths

PTP = [NP, S, NP]

PTPH = [NP:Airlines, S:matched, NP:Wagner]

These features had disappointingly little impact!

Feature examples

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

Entity-based features		
	Entity_1 type	ORG
	Entity_1 head	*airlines*
	Entity_2 type	PERS
	Entity_2 head	*Wagner*
	Concatenated types	ORGPERS
Word-based features		
	Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
	Word(s) before Entity_1	NONE
	Word(s) after Entity_2	*said*
Syntactic features		
	Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
	Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
	Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Classifiers for supervised methods

Now use any classifier you like:

- SVM
- Logistic regression
- Naïve Bayes
- etc.

[Zhou et al. used a one-vs-many SVM]

Sample results

	Count				Cost (%)						
	Ent	Detection		Rec	Detection		Rec	Value	Value-based		
	Tot	FA	Miss	Err	FA	Miss	Err	(%)	Pre	Rec	F
ART	261	38	157	84	9.1	63.9	2.5	**24.5**	74.2	33.6	**46.2**
GEN-AFF	235	28	120	92	9.1	51.5	5.0	**34.5**	75.6	43.6	**55.3**
ORG-AFF	503	71	216	237	9.6	45.4	4.0	**41.0**	78.9	50.6	**61.6**
PART-WHOLE	354	57	182	110	12.1	48.9	2.2	**36.8**	77.4	48.9	**59.9**
PER-SOC	213	24	90	116	5.6	38.5	2.4	**53.5**	88.0	59.1	**70.7**
PHYS	428	76	298	113	8.7	69.1	6.2	**16.0**	62.3	24.7	**35.4**
total	1994	294	1063	752	9.4	53.5	4.0	**33.1**	76.1	42.5	**54.5**

Surdeanu & Ciaramita 2007

Sample results

	Count				Cost (%)						
	Ent	Detection		Rec	Detection		Rec	Value	Value-based		
	Tot	FA	Miss	Err	FA	Miss	Err	(%)	Pre	Rec	F
Artifact	14	0	13	1	0.0	92.0	2.4	**5.6**	70.0	5.6	**10.4**
Business	63	4	39	24	2.2	63.8	3.4	**30.7**	85.6	32.8	**47.5**
Citizen...	171	23	83	73	10.5	49.6	5.7	**34.1**	73.3	44.6	**55.5**
Employment	344	61	113	189	12.1	34.8	4.0	**49.1**	79.1	61.2	**69.0**
Family	118	19	32	79	8.6	20.9	0.4	**70.1**	89.7	78.7	**83.8**
Founder	6	0	5	1	0.0	88.8	3.4	**7.8**	70.0	7.8	**14.1**
Geographical	223	33	102	71	10.4	42.0	1.9	**45.7**	82.1	56.1	**66.7**
Investor...	8	0	5	3	0.0	57.1	2.9	**40.0**	93.3	40.0	**56.0**
Lasting-Personal	32	1	19	13	1.9	50.6	7.8	**39.8**	81.2	41.6	**55.0**
Located	382	72	263	102	9.2	68.3	6.6	**15.9**	61.4	25.1	**35.6**
Membership	96	8	55	33	6.0	61.3	4.2	**28.5**	77.2	34.5	**47.7**
Near	46	4	35	11	4.9	75.2	3.2	**16.7**	72.8	21.6	**33.3**
Org-Location	64	5	37	19	5.9	55.6	3.2	**35.3**	82.0	41.2	**54.8**
Ownership	15	2	13	2	5.0	87.5	0.0	**7.5**	71.4	12.5	**21.3**
Sports-Affiliation	17	0	15	2	0.0	88.4	3.5	**8.1**	70.0	8.1	**14.6**
Student-Alum	17	0	10	7	0.0	60.0	7.5	**32.5**	81.2	32.5	**46.4**
Subsidiary	117	24	67	38	16.1	58.8	2.9	**22.2**	66.8	38.3	**48.7**
User-Owner...	261	38	157	84	9.1	63.9	2.5	**24.5**	74.2	33.6	**46.2**
total	1994	294	1063	752	9.4	53.5	4.0	**33.1**	76.1	42.5	**54.5**

Surdeanu & Ciaramita 2007

Relation extraction: summary

- Supervised approach can achieve high accuracy
  - At least, for *some* relations
  - If we have lots of hand-labeled training data
- But has significant limitations!
  - Labeling 5,000 relations (+ named entities) is expensive
  - Doesn't generalize to different relations
- Next time: beyond supervised relation extraction
  - Semi-supervised relation extraction
  - Distantly supervised relation extraction
  - Unsupervised relation extraction