Making Sense of Unstructured Data

Dan Roth
Department of Computer Science
University of Illinois at Urbana-Champaign

Cognitive Computation Group
University of Illinois at Urbana-Champaign

CCICADA

U.S. Department of Homeland Security

MIAS

ILLINOIS
University of Illinois at Urbana-Champaign

Information Access and Synthesis

- Most of the data today is unstructured
  - **books, newspaper articles, journal publications, reports, transcribed audio streams, images, and video streams.**

- How to deal with the huge amount of unstructured data?

- This is a collection of problems that everyone cares about:
  - Natural Scientists, social scientists, humanity scholars, engineers
  - Analysts of different sorts
  - All agencies
  - First responders
  - Business planning and intelligence

- A Natural Language Perspective

Information Access and Synthesis

In early November, American intelligence authorities say they learned from a communications intercept of Qaeda followers in Yemen that a man named "Umar Farouk" — the first two names of the jetliner suspect, Umar Farouk Abdulmutallab — had volunteered for a coming operation.

Different Names—same entity?

Different References—same entity?

Washington, Dec. 30 (ANI): Agency (CIA) had received a lead on a person dubbed "The Nigerian" suspected for meeting "terrorist elements" in Yemen as early as August 2009, but it failed to establish a link when Umar Farouk Abdulmutallab's father went to the U.S. Embassy in Nigeria in November to express concerns about his son's ties with al Qaeda.

Information Access and Synthesis

The Christmas Day Nigerian bomber's failed attack on Northwest's Flight 253 makes us thankful that the plot was foiled by alert passengers, but shows that security in the U.S. is also a failure at critical times. Abdul Mudallad was known to be a threat, with possible al-Qaeda ties, for at least two years, was on a federal "watch list" but not on the "no-fly list." Would anything have made a difference? Mudallad boarded the flight in Amsterda

Does it mean that….

Almost—Natural Language Understanding (in context)

In late December, more intercepts of Qaeda oper
focused their attacks in the region, mentioned the date of Dec. 25, and suggested that they were "looking for ways to get somebody out" or "for ways to move people to the West," one senior administration official said.

Information Access and Synthesis

But on Christmas Day, the final draft of the memorandum was still sitting in the computer of a junior C.I.A. analyst, waiting until a photo of the young Nigerian was located. Unbeknownst to the analyst, officials said, Mr. Abdulmutallab's photo had already been delivered to other counterterro

Can we integrate knowledge from TEXT with knowledge from IMAGES?

The American intelligence network was clearly listening in Yemen and sharing that

"The program not only can't connect the dots, it can't find the dots," Representative Brad Miller, Democrat of North Carolina and chairman of a House panel that oversees the program, said at the time.

Information Access and Synthesis

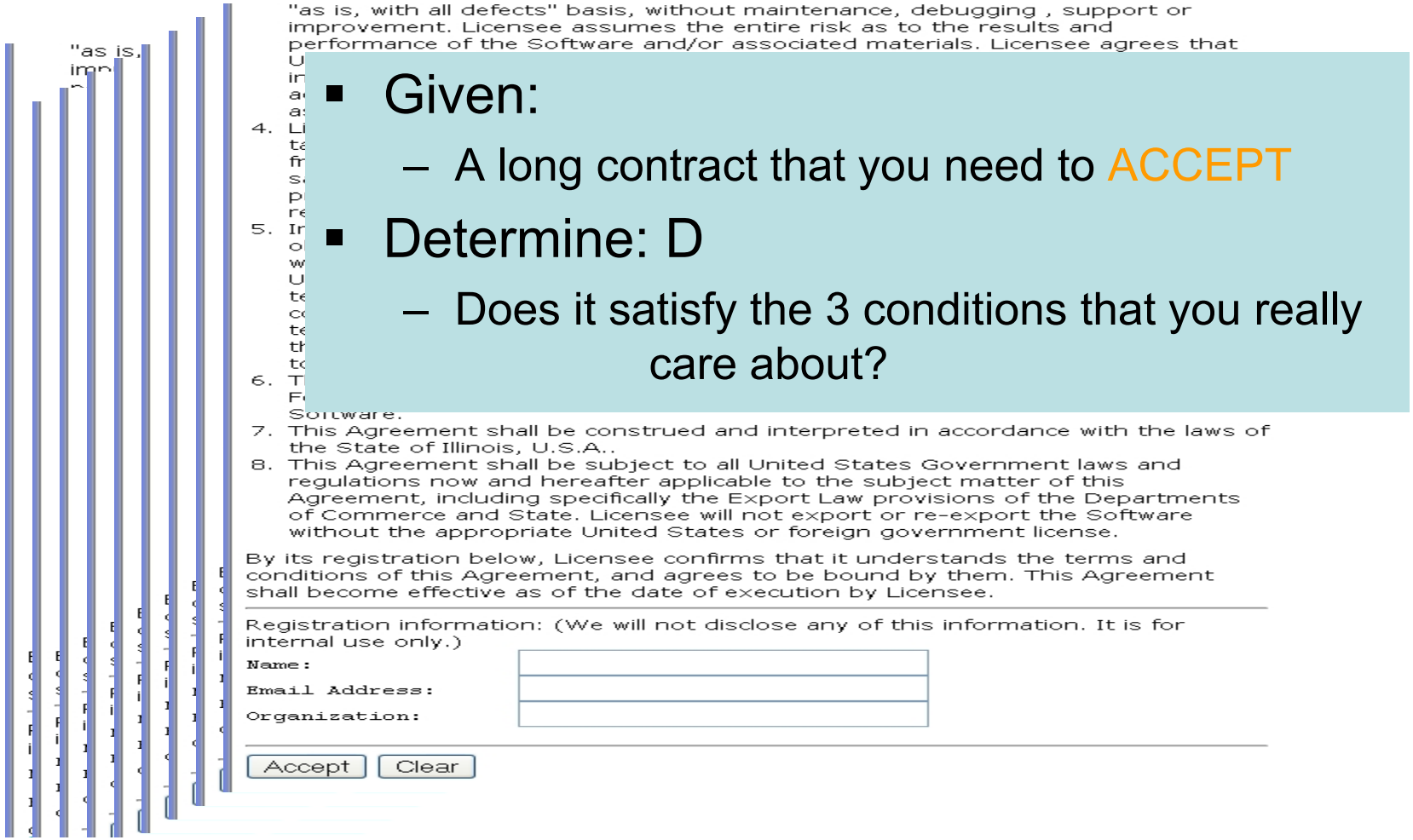
- Most of the data today is unstructured
  - **books, newspaper articles, journal publications, reports, images, and audio and video streams.**

- How to deal with the huge amount of unstructured data as if it was organized in a database with a known schema.
  - **how to locate, organize, access and analyze unstructured data.**

Research Program:

- develop the theories, algorithms, and tools for analysts to
  - **Intelligently access a variety of data formats and models**
  - **integrate them with existing resources**
  - **transform raw data into useful and understandable information.**

A view on Open Information Extraction


"as is, with all defects" basis, without maintenance, debugging , support or improvement. Licensee assumes the entire risk as to the results and performance of the Software and/or associated materials. Licensee agrees that

4. Li
5. In
6. T
Software.
7. This Agreement shall be construed and interpreted in accordance with the laws of the State of Illinois, U.S.A..
8. This Agreement shall be subject to all United States Government laws and regulations now and hereafter applicable to the subject matter of this Agreement, including specifically the Export Law provisions of the Departments of Commerce and State. Licensee will not export or re-export the Software without the appropriate United States or foreign government license.

By its registration below, Licensee confirms that it understands the terms and conditions of this Agreement, and agrees to be bound by them. This Agreement shall become effective as of the date of execution by Licensee.

Registration information: (We will not disclose any of this information. It is for internal use only.)

Name:

Email Address:

Organization:

Accept Clear

- Given:
  - A long contract that you need to ACCEPT
- Determine: D
  - Does it satisfy the 3 conditions that you really care about?

ACCEPT?

Large Scale Understanding: Massive & Deep

Determine if Jim Carpenter works for the government

topic—intelligence

Jim Carpenter works for the U.S. Government.

The American government employed Jim Carpenter.

Jim Carpenter was fired by the US Government.

Jim Carpenter worked in a number of important positions. …. As a press liaison for the IRS, he made contacts in the white house.

Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter

Former US … today…

Standard techniques cannot deal with the variability of expressing meaning nor with the ambiguity of interpretation

Needs:

- Relations and Semantic Classes, NOT keywords
- Exhaustive recovery of information needs
- Abstract over variability in natural language
- Integrate over large collections of text and DBs
- Track entities, events, etc.

This Talk

- Technical Approach
- Problems & Solutions
  - Semantic Role Labeling
  - NER
  - Reference & Co-reference
  - Textual Entailment
  - Events
  - Multilingual
  - Trust
  - Text & Images

- Describe a few key NLP and IE problems and sketch our solutions
- Can't be comprehensive nor deep

- Natural Language Processing is hard.
  - Not sufficient to do something; need to do it well, use common benchmarks and quantify performance.
- I will point to several state-of-the-art systems
  - all are publicly available for research purposes
  - Demo & Software page of MIAS & the Cognitive Computation Group
- Know about it & Use it – many have.
  - Recent customers: VACCINE

Machine Learning + Inference based NLP

- Modeling and learning algorithms for different phenomena
  - Classification models
  - Structured models
  - Supervised as well as semi-supervised and unsupervised models
  - Study of adaptation of modes to new domain

- Inference as a way to introduce domain & task specific constraints
  - Using constrained optimization

Extracting Relations via Semantic Analysis

Semantic Role Labeling Output

Screen shot from a CCG demo
http://L2R.cs.uiuc.edu/~cogcomp

Input Text:

A car bomb that exploded outside the U.S. military base in Beniji killed 11 Iraqi citizens.

Result: Complete!

⊞ General Explanation of Argument Labels

A	bomb [A1]		killer [A0]
car			
bomb			
that	bomb (Reference) [R-A1]		
exploded	V: explode		
outside	location [AM-LOC]		
the			
U.S.			
military	temporal [AM-TMP]		
base			
in	location [AM-LOC]		
Beniji			
killed			V: kill
11			corpse [A1]
Iraqi			
citizens			
.			

- Semantic parsing reveals several relations in the sentence along with their arguments.

Top system in the CoNLL Shared Task Competition 2005

This level of analysis, however, cannot abstract over the inherent variability in expressing the relations. Kill and Explode can be expressed in many different ways.

Extended Semantic Role Labeling

Who did what to whom, why, where, when…

Jim Carpenter **works** for the U.S. Government.

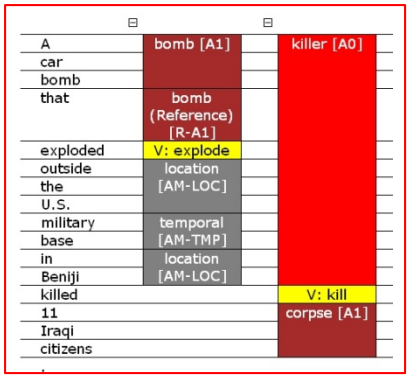
Jim Carpenter, the Dept. of State spokesperson, was seen in Mexico City on Wednesday

Jim Carpenter ***of*** *the US Department of State* said today that ….

Russian interior minister Yevgeny Topolov met yesterday with *his US counterpart, Jim Carpenter*.

Jim Carpenter ***'s*** *employment* with the American government …

Standard Semantic Role Labeling is defined in terms of verbs only.



Prepositions, commas, nominalizations and possessives also express relations.

Task:

- Generalize SRL to support more linguistic phenomena
- Joint inference across these tasks
- Address variability of training data across domains: essential given no unified resources for all phenomena

Named Entity Recognition.

SOCCER - [PER BLINKER] BAN LIFTED .
[LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was [illegible] for [ORG Feyenoord].

- **Identify**
  - Organizations
  - Locations
  - People
  - Quantities & Temporal (+reasoning)
  - …
- Main algorithmic questions:
  - Representing Entities
  - Exploiting non-local information.
  - Injecting world knowledge.

- **Key contribution**
  - Best system available today. [Ratinov & Roth, CoNLL-08]
  - Uses Wikipedia to augment entities lists and statistics
    - Other resources possible
  - Uses unlabeled text

The Reference Problem

Kennedy

The same problem exists with other types of entities

[Li, Morie, Roth, NAACL'04, AAAI'04, AI Magazine'05]

Document 1: *The Justice Department has officially ended its inquiry into the assassinations of* **John F. Kennedy** *and Martin Luther King Jr., finding ``no persuasive evidence'' to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that* **Kennedy** *was ``probably'' assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the* ***Warren Commission*** *'s belief that Lee Harvey Oswald acted alone in* ***Dallas*** *on Nov. 22, 1963.*

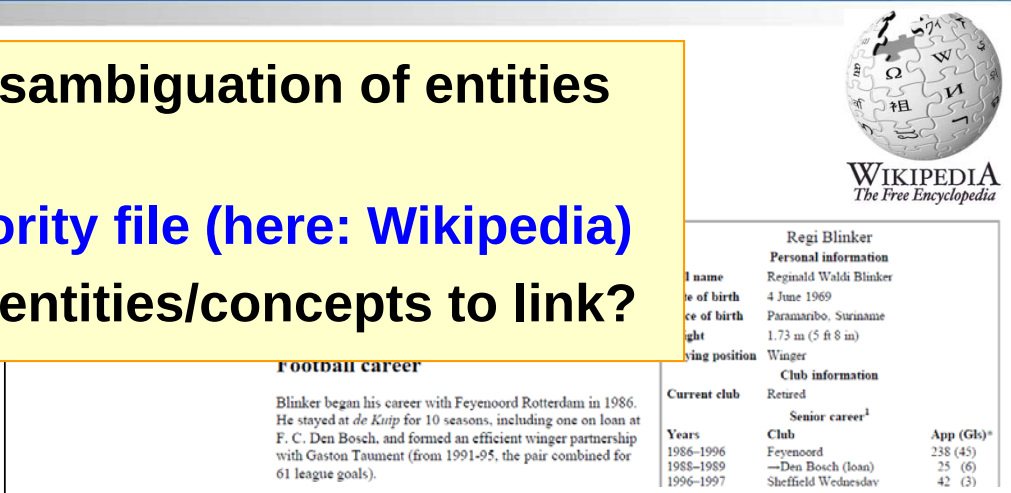
Document 2: *In 1953, Massachusetts* **Sen. John F. Kennedy** *married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate* **John F. Kennedy** *confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, ``I do not speak for my church on public matters, and the church does not speak for me.''*

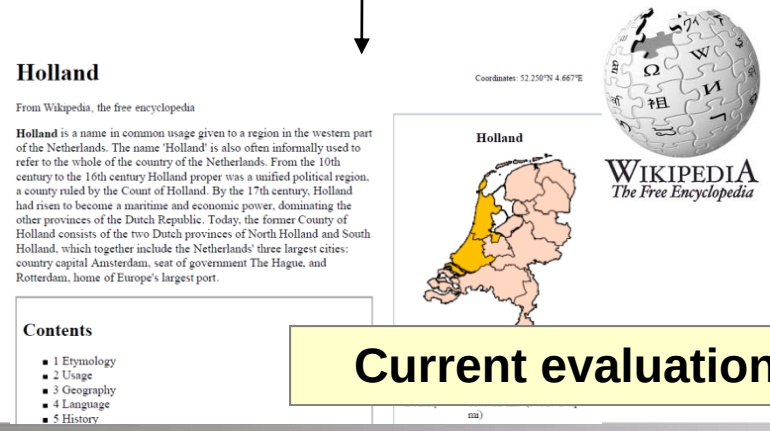
Document 3: David Kennedy *was born in Leicester, England in 1959. …***Kennedy** *co-edited The New Poetry (Bloodaxe Books 1993), and is the author of New Relations: The Refashioning Of British Poetry 1980-1994 (Seren 1996).*

Wikification: Linking Entities to Wikipedia

- **Context sensitive disambiguation of entities and concepts**
- **Mapping to an authority file (here: Wikipedia)**
- **Key question: what entities/concepts to link?**

SOCCER - [LOC LON] [PER Regg] lifted by [his [OR [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].





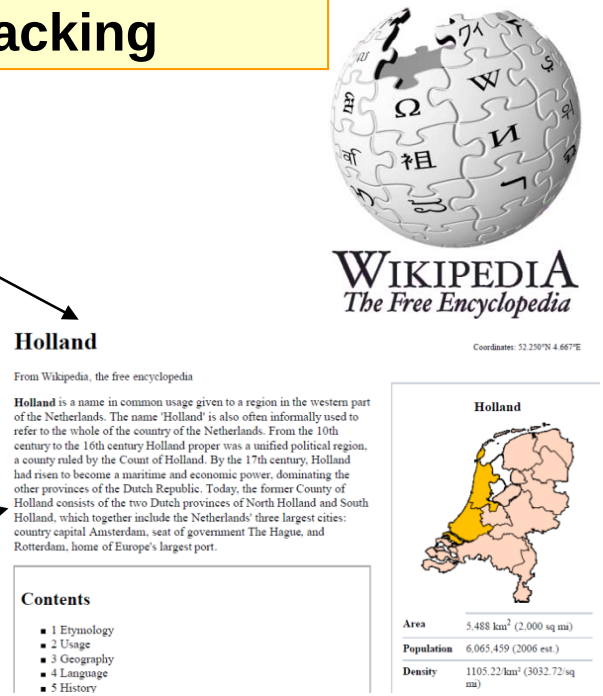
Current evaluation focuses on blogs data

Cross-document Co-reference: Dutch=Oranje?

- **Key Application: entity tracking**

SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Regi Blinker played three matches for Oranje. The left winger started his pro carreer at Feyenoord and played 400 official matches for Feyenoord, Celtic and Sparta. He retired from football in 2003. Where is he now?

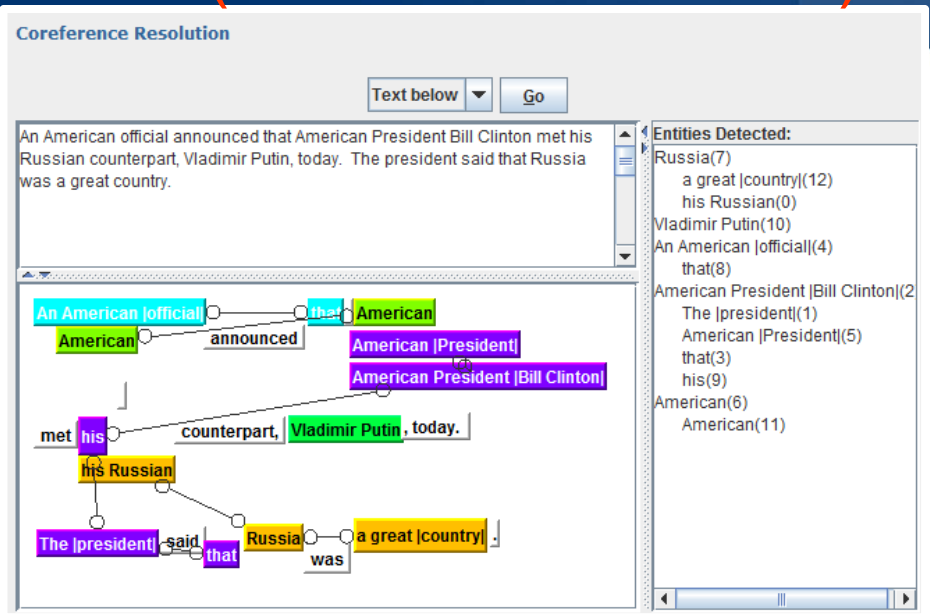


Co-reference Resolution (Within Documents)

The Problem:

identify phrases that refer to entities and cluster them according to those entities.

An American official announced that American President Bill Clinton met his Russian counterpart, Vladimir Putin, today. The president said that Russia was a great country.



The Solution:

- A Classification based pairwise model
- Followed by a decoding (inference) step with a single constraint,
- A thorough feature engineering

Details:

Coref classifier $C : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ takes two mentions a

We allow each mention to be linked to at most one previc
while maximizing the sum of linked scores.

Best system available today.
(Bengston & Roth, EMNLP'08)

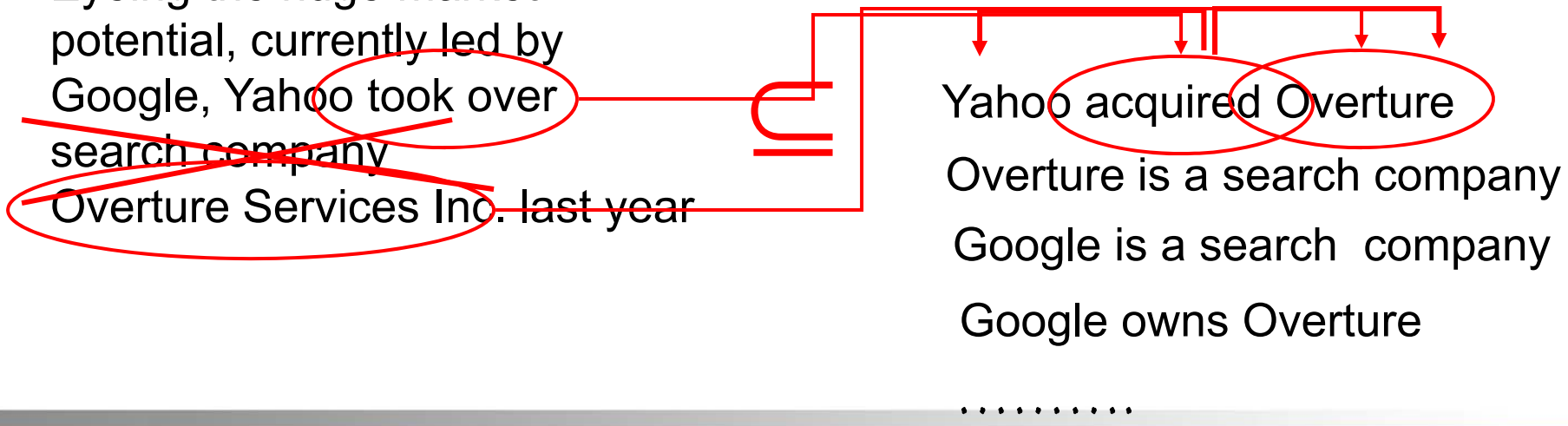
Current work:

- *Integrating within and across document co-ref*
- *Incorporating background knowledge into co-ref*

Recognizing Textual Entailment

- Alternate formulation of Natural Language Understanding
- Instead of mapping text to a canonical (FOL) representation, answer a different question:
  - Given two text fragments, does the meaning of one fragment follow from the meaning of the other? E.g.

Is it true that...?



Application: Discovering Sensitive Information

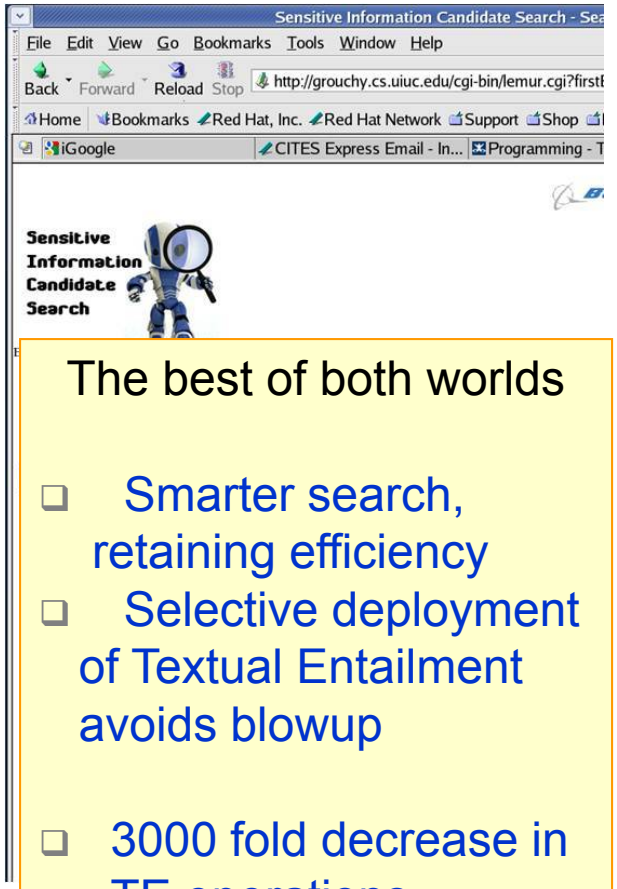
Recent work: Integrating Entailment and Search

Need: Exhaustive extraction of information need

Open-ended information need

- First Stage: use enhanced Information Retrieval techniques to return K most relevant text spans – Semantic Search (SR)
- Multiple NLP technologies support richer indexing
- Index a range of semantic units (Named Entities, Phrasal Verbs, Multi-word expressions…)
- Use specialized (Named Entity, phrasal verb, distributional) similarity measures to build multiple IR indexes, and expand query

- Second Stage: use Textual Entailment (TE) to filter results



The best of both worlds

- Smarter search, retaining efficiency
- Selective deployment of Textual Entailment avoids blowup
- 3000 fold decrease in TE operations

Event Recognition

Tasks:

- Recognizing and classifying events in texts.
- De-duplicating recognized events.
- Tracking events.

BBC, Nov. 6th, 2009: Thirteen people have been killed and 30 injured in a shooting at a military base in Texas. News of the shooting has dominated the US media.

CNN, Nov. 6th, [illegible]nd injured mor[illegible] his fellow s[illegible] fourteen ho[illegible] another ma[illegible] But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

Collaborations:

- START – Terrorism Events
- Democracy Institute: Societal Stability & Climate Change

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times. Munley was wounded in the exchange.

Event Recognition

Technical Challenges:

- Recognition of Entities
- Co-reference resolution
- Semantics of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen** people have been killed and **30** injured in a shooting at a military base in **Texas**. News of the shooting has dominated the **US** media.

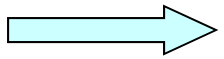
CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan**, killed **13** soldiers and injured more than **two dozen** on **November 5** by opening fire on **his** fellow soldiers at **Fort Hood** in a **US** military base. After fourteen hours, **two** people were killed and **six** were hurt in another mass shooting in the **United States** in **Orlando, Florida**. But the deadliest terror-event particular at the **US** military base is shocking for every one, belonging to any religious community.

CNN, Nov. 7th, 2009: It was **Munley** who arrived quickly **Thursday** at the scene of the worst massacre at an **Army** base in **U.S**. history, where **13** people were killed. **She** confronted the alleged gunman, **Maj. Nidal Malik Hasan**, and shot **him** **four** times. **Munley** was wounded in the exchange.

Entities:

- Number
- Date
- Person
- Organization
- Location

Co-reference:



Event Recognition

Technical Challenges:

- Recognition of Entities
- Co-reference resolution
- Semantics of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen people have been killed and 30 injured in a shooting at a military base in Texas.** News of the shooting has dominated the US media.

massacre

CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan, killed 13 soldiers and injured more than two dozen on November 5 by opening fire on his fellow soldiers at Fort Hood in a US military base. After fourteen hours, two people were killed and six were hurt in another mass shooting in the United States in Orlando, Florida.** But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

massacre

massacre

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. **She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times. Munley was wounded in the exchange.**

shooting

wound

Event Recognition

Technical Challenges:

- Recognition of Entities
- Co-reference resolution
- Semantics of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen people have been killed and 30 injured in a shooting at a military base in Texas.** News of the shooting has dominated the US media.

massacre

CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan, killed 13 soldiers and injured more than two dozen on November 5 by opening fire on his fellow soldiers at Fort Hood in a US military base. After fourteen hours, two people were killed and six were hurt in another mass shooting in the United States in Orlando, Florida.** But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

massacre

massacre

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. **She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times. Munley was wounded in the exchange.**

shooting

wound

Named Entity Transliteration

- The process of transcribing a NE to a different language
- Key step in
  - transferring resources across languages
  - supporting multilingual access to information

The Supreme Court has confirmed a narrow win for the centre-left opposition led by Romano Prodi.

But after a meeting with his advisers, Berlusconi let it be known he was considering a further legal challenge.

Премьер-министр Италии Берлускони отказался признать решение Верховного суда, который подтвердил победу левоцентристской коалиции на всеобщих выборах.

Find **Obama** in the Hebrew Wikipedia

- This process should capture phonetic and orthographic differences across the two languages
- Not easy. Context sensitive; time sens

Best system available, both unsupervised and supervised English—[Russian; Hebrew; Chinese] (NAACL'09, CIKM'09)

Application: Trust in on-line information

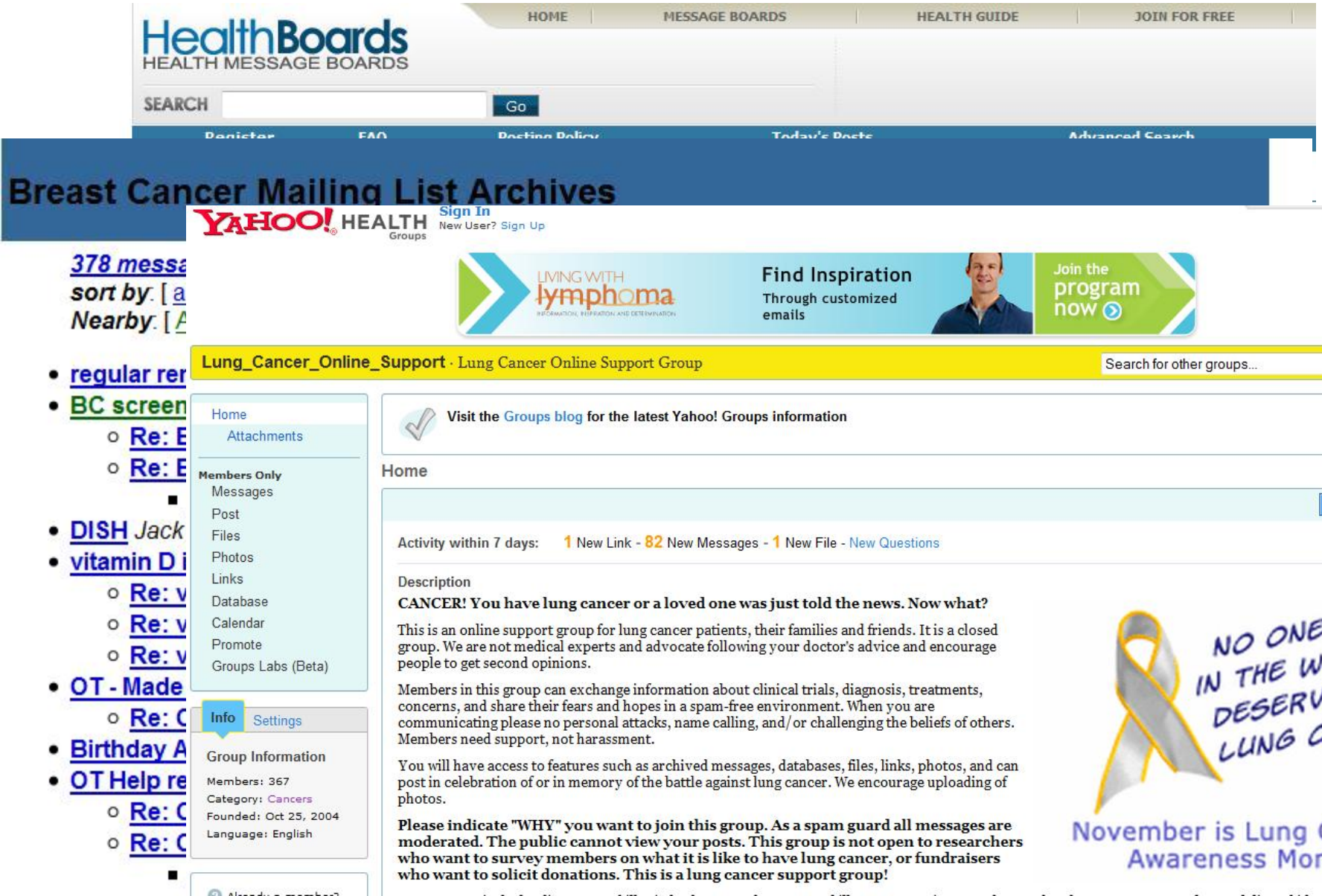
What to believe?

"A review article of the latest studies looking at red wine and cardiovascular health shows drinking two to three glasses of red wine daily is good for the heart."

Excerpt from http://www.sciencedaily.com/

- Is this information trustable?
- What do others think about this information?
- Is this supported by "reliable" sources?

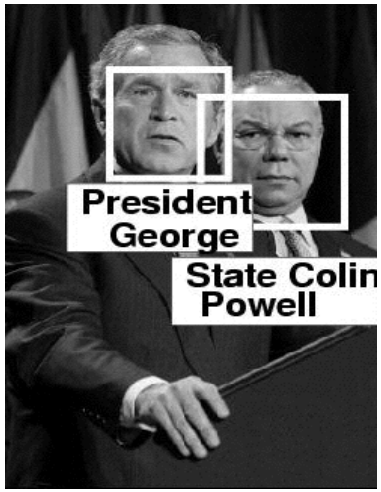
Many support groups and medical forums



Trustworthiness

- Given:
  - Multiple content sources: websites, blogs, forums, mailing lists
  - Some target relations ("facts")
    - E.g. [disease, treatments], [treatments, side-effects]
  - Prior beliefs and background knowledge
- We can:
  - Score trustworthiness of relations ("facts") based on
    - support across multiple (trusted) sources
    - source characteristics:
      - reputation, interest-group (commercial / govt. backed / public interest), verifiability of information (cited info)
  - Rate databases/sources as more/less trustworthy based on the verifiability scores of the facts in database
  - Track how the trustworthiness of fact / database varies with time as the text corpus grows over time

Semantic Data Enrichment: Text and Images



US President George W. Bush (L) makes remarks while Secretary of State Colin P... signing the U... HIV /AIDS , T... Act of 2003 a... State in Wash... year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/Luke Frazza)

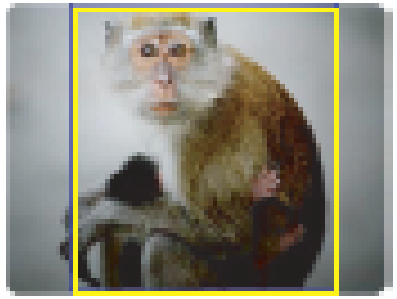
German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her ... Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

British director Sam Mendes and his partner actress Kate Winslet ... premiere of ... n', ... The films ... Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

The kinds of information contained in pictures may be very difficult to access in other ways.

Adding Annotation [Forsyth, Hoiem et. al]

- Problem: How to tag pictures with text-like summaries of what is important in the picture
  - Method: Analyze relations between text and images in illustrated documents
- Key Challenge:
  - make sensible statements about unknown objects
- Strategy
  - Represent objects as collections of attributes
    - "Red"; "has a head"; "furry"; "has a wing" ; "made of metal"
  - Build classifiers to detect attributes
    - Right bag of attributes → object
    - funny attributes? say so
    - Otherwise - report attributes



'Head'
'Ear'
'Snout'
'Nose'
'Mouth'

- State-of-the-art recognition for known objects
- Accurate reports of attributes of unknown objects Recognition from text description

Bird
"Leaf"

Motorbike
"cloth"

DiningTable
"skin"

Sofa
"wheel"

Bike
"Horn"

Aeroplane
No "wing"

Car
No "window"

- Predict ways in which recognized object is special
- I.e. extra/missing attributes

Bicycle
No "wheel"

Sheep
No "wool"

Train
No "window"

Information Access & Synthesis [Processes & Tools]

- Focused data retrieval and integration,
  - Identify and collect relevant data from multiple sources
- Semantic data enrichment,
  - Infer semantics from unstructured data and images;
  - Allow navigation and search across disparate data modalities;
- Entity identification and relations discovery,
  - Identify real-world entities and relations among them
  - Relate them to existing institutional resources for information i
- Knowledge discovery and hypotheses generation and verifi
  - Construct the rich semantic structure and hidden networks of e
- Foundations
  - Machine learning, database and data mining, natural language inference and optimization and computer vision techniques
  - Called for and driven by the aforementioned problems.

Tools

Text Processing& Analysis

Semantic Analysis & Information Extraction

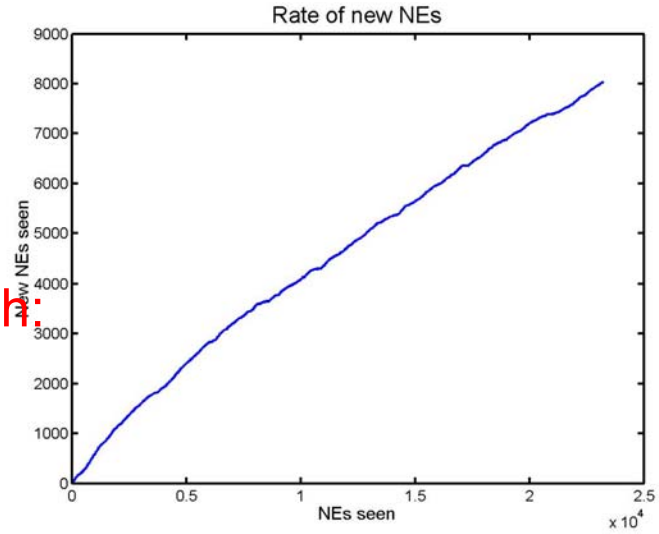
Information Integration

Machine Learning & Data Mining

Integrating Text & Images

Example: Named Entity Recognition

- Entities are inherently ambiguous (e.g. JFK can be both location and a person depending on the context)
  - Can appear in various forms ; Can be nested.
  - Using lists is not sufficient
  - New entities are always being introduced
- Necessary to use a Machine Learning approach:
- Significant over fitting
- Adaptation to:
  - New domains/corpora
  - Slightly new definition of an entity
  - New languages; New types of entities
- Need to:
  - reduce the requirements on resources (training)
  - Incorporate knowledge



Constraints Conditional Models (CCMs)
aka Integer Linear Programming for NLP

- Informally: Global decisions with learned models, in the presence of constraints
- Why Constraints?
  - A effective way to inject expressive prior knowledge into models.
- W
  -
  -
  -
- St
- Ha
  -

Issues to attend to:

- While we formulate the problem as an ILP problem, Inference can be done multiple ways
  - Search; sampling; dynamic programming; SAT; ILP
- The focus is on joint global inference
- Learning may or may not be joint.
  - Decomposing models is often beneficial

[ee tutorial on my web page and ILPNLP workshop]

Formal Model

$$\underset{y}{\operatorname{argmax}}\, \boldsymbol{\lambda} \cdot F(x, y)$$

Penalty for violating the constraint

Subject to constraints

oft) constraints
mponent

Weight Vector for "local" models

A collection of Classifiers; Log-linear models (HMM, CRF) or a combination

How far y is from a "legal" assignment

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an exact solution.

Search techniques are also possible

How to train?

How to decompose the global objective function?

Should we incorporate constraints in the learning process?