Information Extraction

CIS, LMU München
Winter Semester 2015-2016

Dr. Alexander Fraser, CIS

Information Extraction – Administravia - I

- Vorlesung
  - Learn the basics of Information Extraction (IE)
- Seminar
  - Each student will present a Referat on IE (Powerpoint, LaTeX, Mac)
    - The group will discuss it
  - Also: three or so practical sessions (hopefully we have time)
  - There are two seminars! **You come to just one of the two sessions**, either Thursdays (starting tomorrow), or Wednesdays (starting next week)

Information Extraction – Administravia - II

- Registration:
  - If you are a CIS Student: check whether you are registered for *both* the Vorlesung and the Seminar (these are **two things** in LSF!)
  - There are a good number of people only in the Vorlesung
  - There are a few people only in the Seminar
- A word about space:
  - The seminars are very full in LSF
  - This may be because people registered who will not actually do a Referat – if this applies to you, please let me know (for the sake of your colleagues!)

Information Extraction – Administravia - III

- Vorlesung and Seminar are two separate courses (in same module for CIS people)
  - However, there may be some shifting around of slots depending on time constraints
- Vorlesung (Grade):
  - Klausur (probably 03.02, no discussion of this today please)
- Seminar (Grade):
  - Referat
  - Hausarbeit (write-up of the Referat) (6 pages, due **3 weeks** after you hold your Referat)
  - The Hausarbeit can also include practical exercises (optional, extra points)
- CIS-ler: No Notenverbesserung (everyone else: ask in your Fachschaft!)

Information Extraction - Administravia - IV

- NEXT SEMINAR - COME TOMORROW *OR* ON COMING WEDNESDAY!
  - Ungraded quiz (so that I can see what you already know)
    - Optionally anonymous (you either put your name, or you don't)
  - I will also collect information on who you are and your interests – PUT YOUR NAME ON THIS PAGE! (this page will be collected separately!)
  - **And I want to know what you want to learn in this class!**

Information Extraction – Administravia - V

- Syllabus: updated dynamically on my web page (see also WS last year, but there will be some differences)
  - Brief idea at end of this slide deck (if we finish, then today)
- List of Referatsthemen
  - This will be presented soon in the Seminar, probably in two weeks
- Literature:
  - Required: **Sunita Sarawagi. Information Extraction.** Foundations and Trends in Databases, 1(3):261–377, 2008. (good survey paper, somewhat brief)
    - **Please read the introduction for next week (it is available on the web page!)**
  - Optional: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schuetze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. (good information retrieval textbook, preview copies available from the book website: http://nlp.stanford.edu/IR-book/)

- Questions?

Information Extraction

- An introduction to the course
  - The topic "Information Extraction" means different things to different people
  - In this course we will look at several different perspectives
  - There is unfortunately no comprehensive textbook that includes all of these perspectives

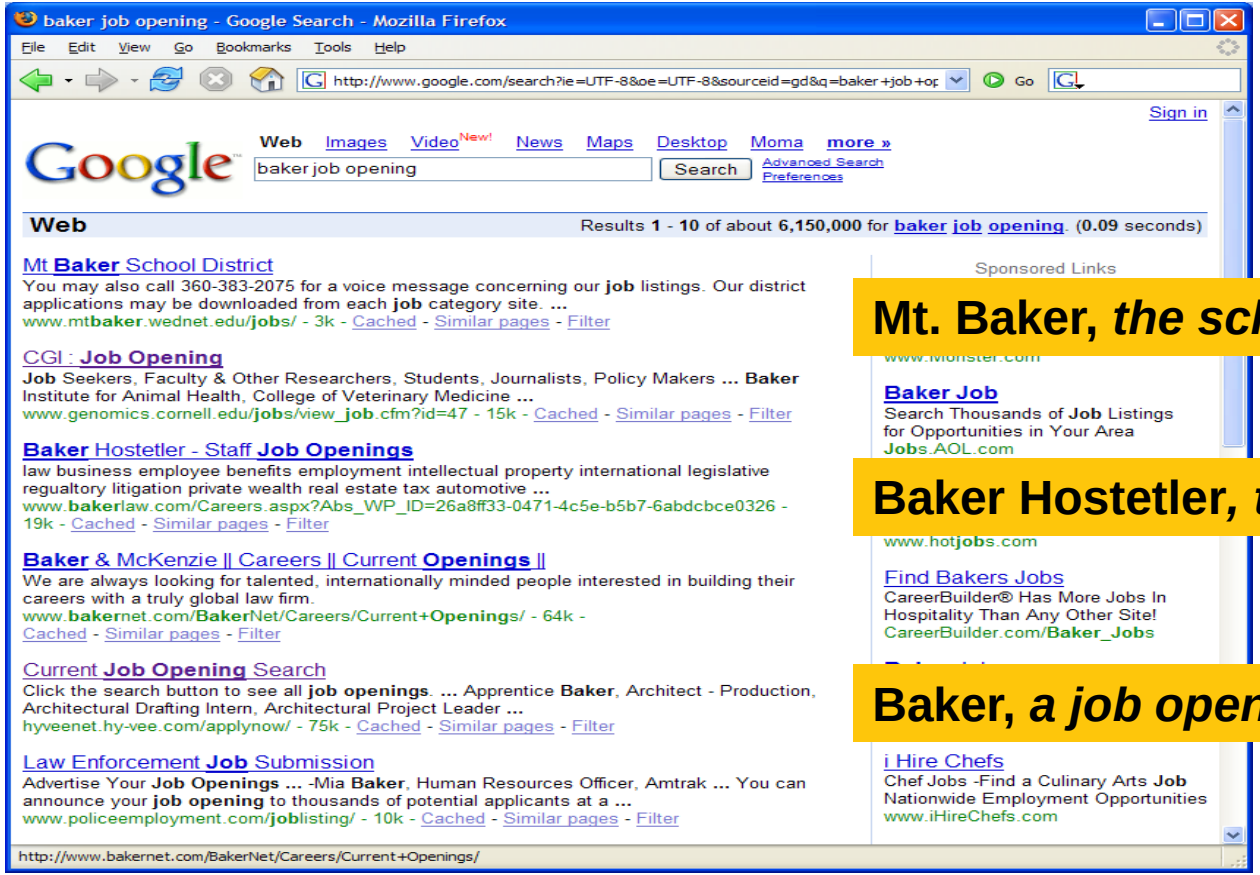
My Biases

- As you may have noticed by now: I am from the US (PhD in Computer Science from USC/ISI AI division)
- I am on permanent staff here at CIS
- I do research in the broad area of **statistical NLP**
  - I mostly work on **statistical machine translation**, and related structured prediction problems (e.g., treebank-based syntactic parsing, generation using sequence (tagging) models)
  - I also work on other multilingual problems such as cross-language information retrieval
- With respect to **rule-based NLP** (with manually written rules), I'll try to be as fair as (humanly) possible, I do use these techniques sometimes too

Outline for today

- Motivation
  - Problems requiring information extraction
  - Basic idea of the output
- Abstract idea of the core of an information extraction pipeline
- Course topics

A problem



Mt. Baker, *the school district*

Baker Hostetler, *the company*

Baker, *a job opening*

Slide from Cohen/Mccallum

Bakery Jobs on CareerBuilder.com
www.careerbuilder.com/**jobs**/keyword/**bakery** +1
Jobs 1 - 25 of 579 – Looking for **Bakery Jobs**? See currently available job **openings** on CareerBuilder.com. Browse the current listings and fill out job applications.

Baker Jobs, Employment | Indeed.com
www.indeed.com/q-**Baker-jobs**.html +1
Jobs 1 - 10 of 16047 – 16047 **Baker Jobs** available on Indeed.com. one search. all jobs.

Job Openings - **Baker** University
www.bakeru.edu › Jobs +1
If you are seeking employment in any of these areas, contact **Baker** University.

Baker, LA **Jobs** on CareerBuilder.com
www.careerbuilder.com/**Jobs**/**Baker**/ +1
Jobs 1 - 25 of 948 – Looking for **Baker**, LA **Jobs**? See currently available **job openings** on CareerBuilder.com. Browse the current listings and fill out **job** ...

Down Under **Bakery** Pies: **Job Openings** at DUB Pies
www.dubpies.com/**jobs**.php +1
Listing of **job openings** at DUB Pies. Down Under **Bakery** (DUB) Pies is looking for more staff - check out our list of vacancies.

Field Engineers | Geoscience | **Jobs** and Careers at **Baker** Hughes
jobs.**baker**hughes.com/ +1
... Oil and Natural Gas? **Baker** Hughes has career information for you on these, more. ... Search **Jobs**. **Baker** Hughes **Jobs** ... Recent **Job Openings**. Completion ...

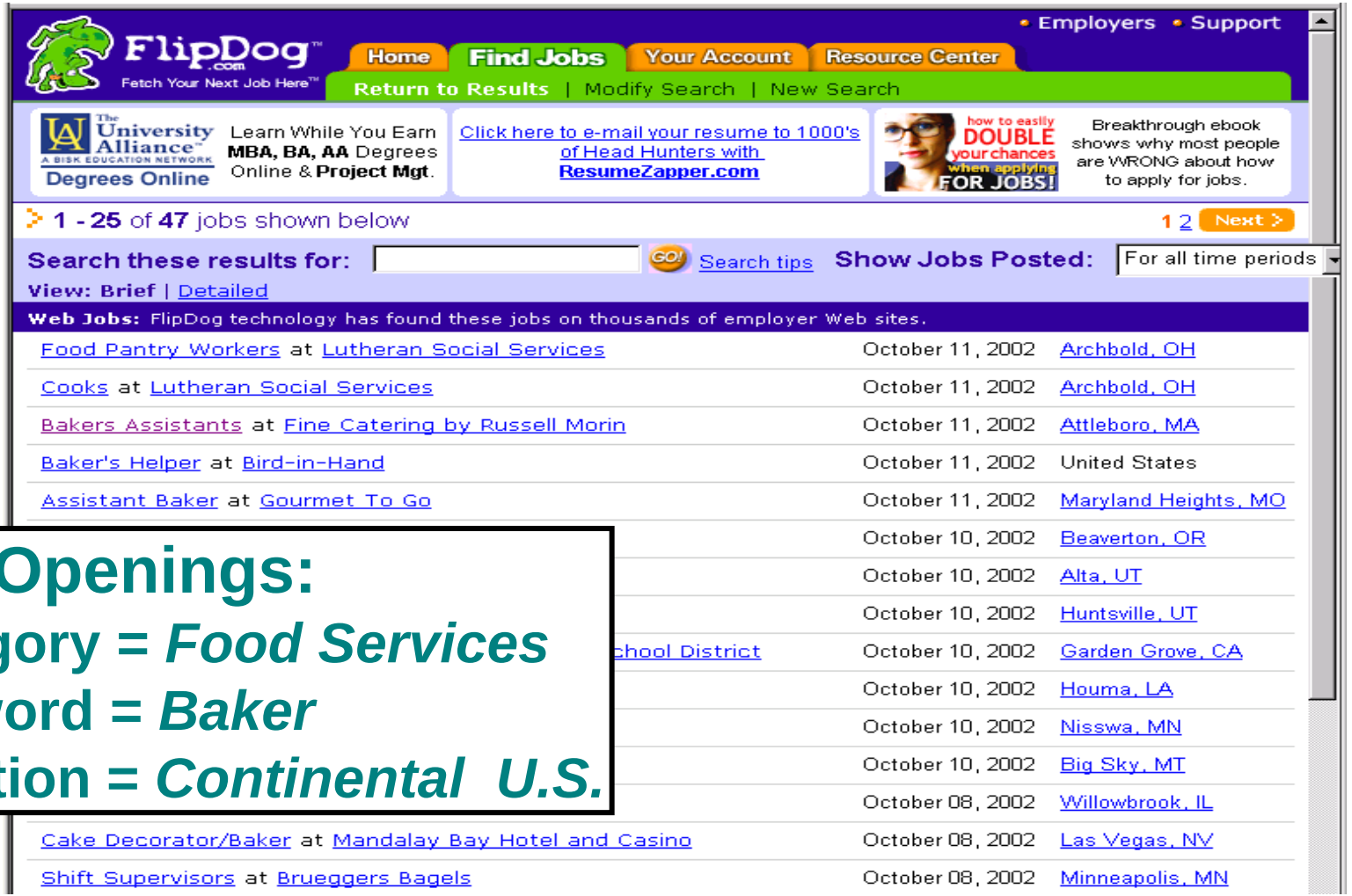
Corner **Bakery Job Openings** | Glassdoor
www.glassdoor.com/**Job**/Corner-**Bakery**-**Job**-**Openings**-E297310_P2... +1
45 Corner **Bakery job openings**. Search job openings, see if they fit - company salaries, reviews, and more posted by Corner Bakery employees.

Jobs - **Baker** University
www.**baker**u.edu/**jobs** +1
See links at left for a complete list of **Baker** University **job openings**. It is the policy of **Baker** University to afford equal opportunity for all persons without distinction ...

Slide from Kauchak

A solution

Slide from Cohen/McCallum



Job Openings:

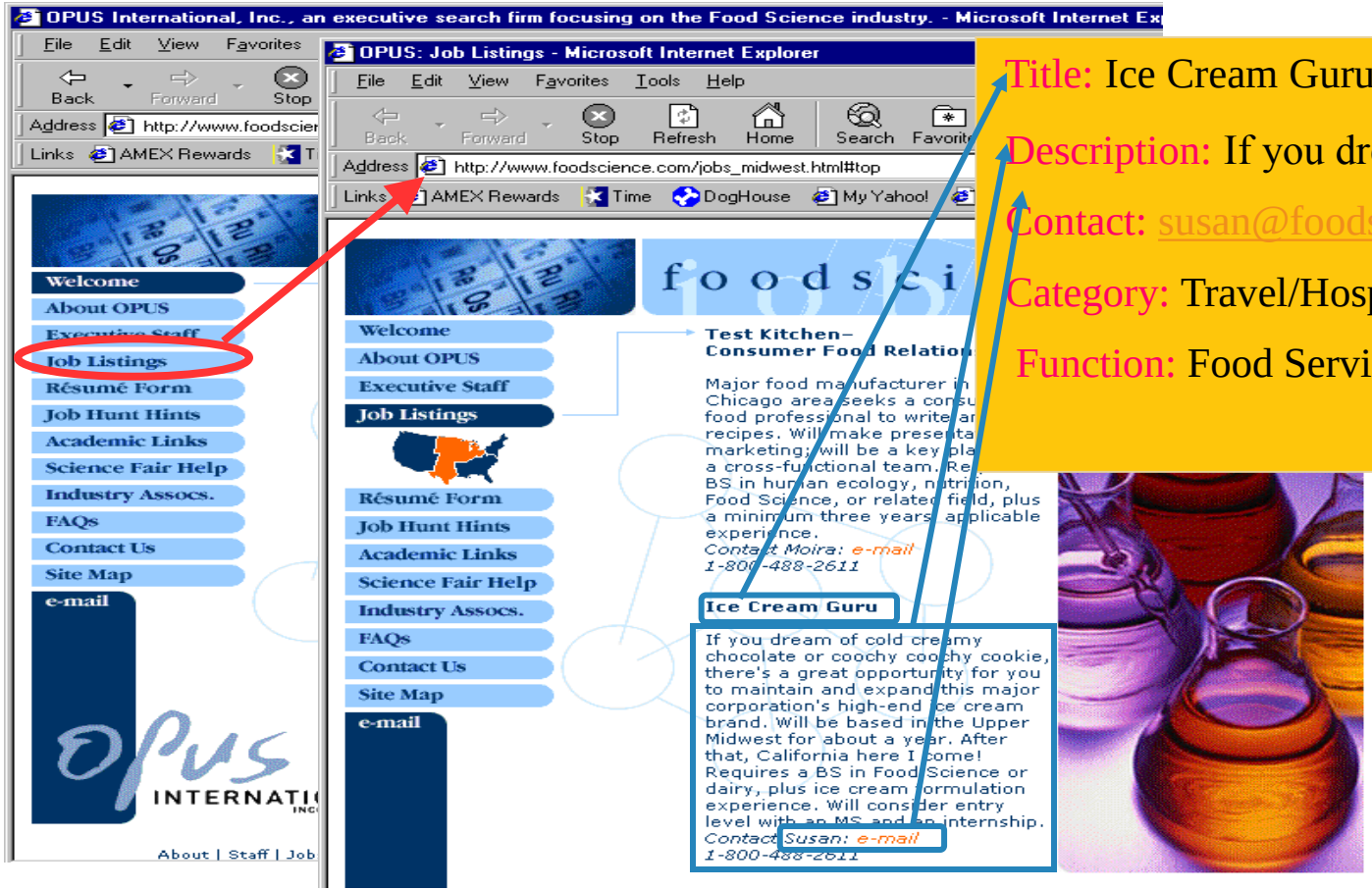
Category = *Food Services*

Keyword = *Baker*

Location = *Continental U.S.*

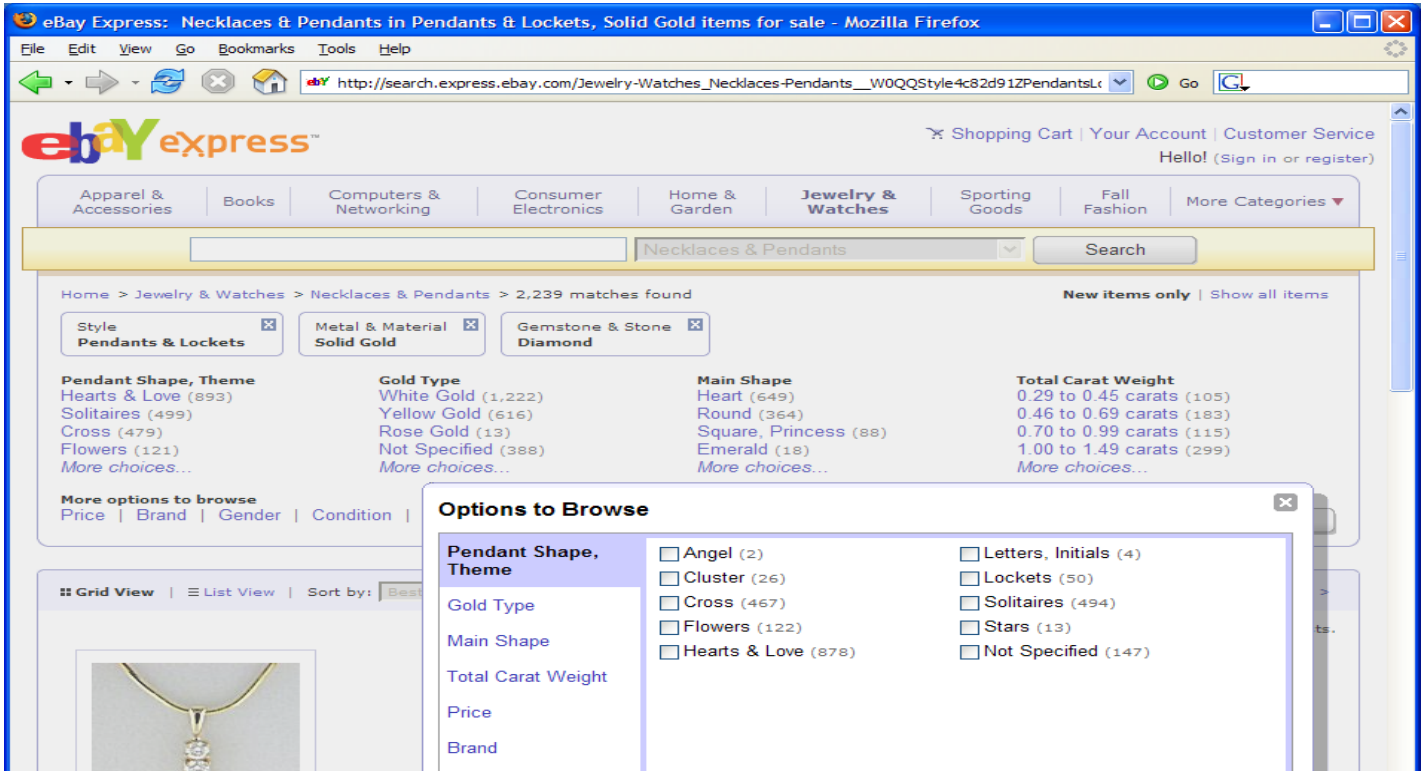
Slide from Cohen/McCallum

Extracting Job Openings from the Web



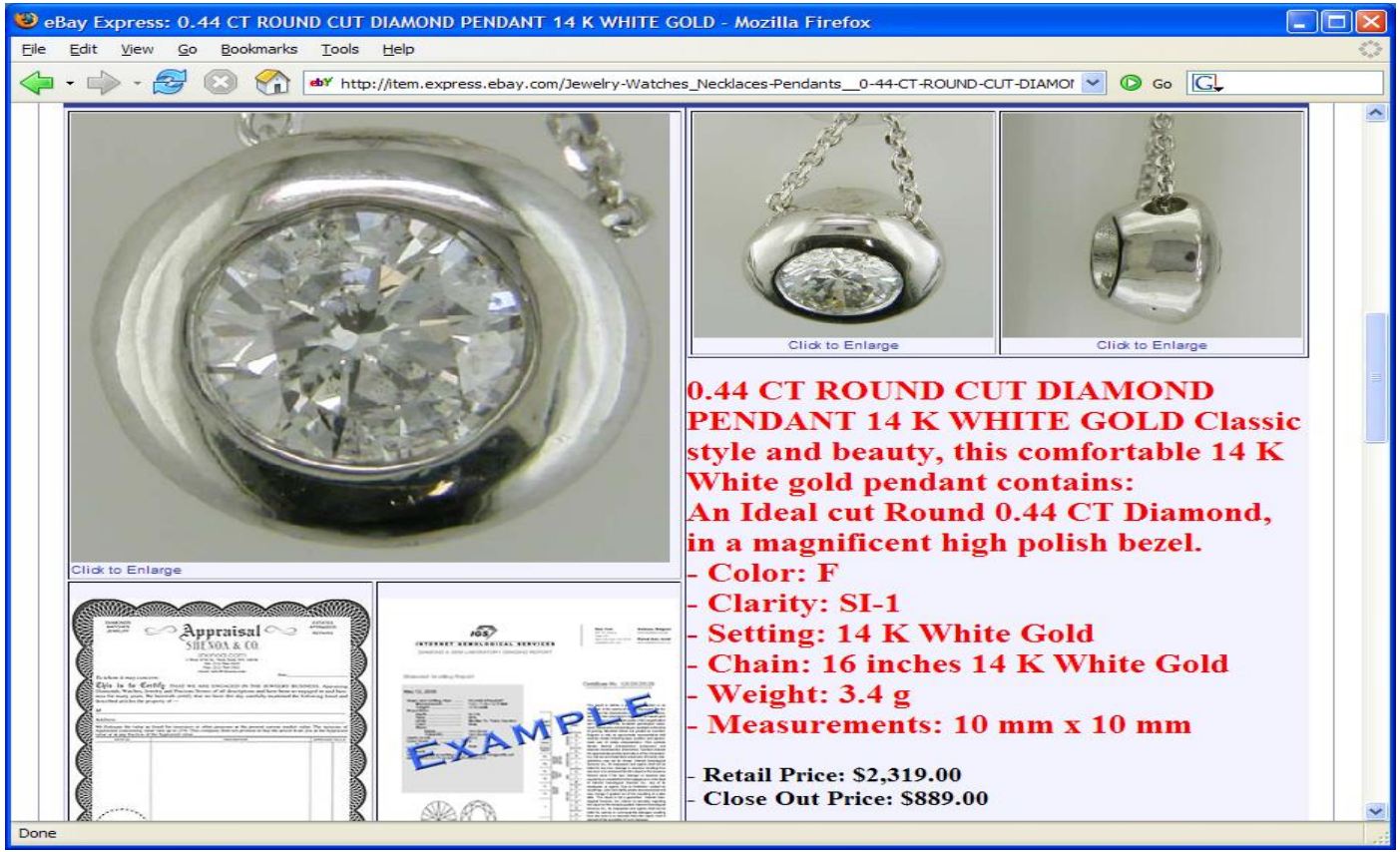
Slide from Cohen/McCallum

Another Problem



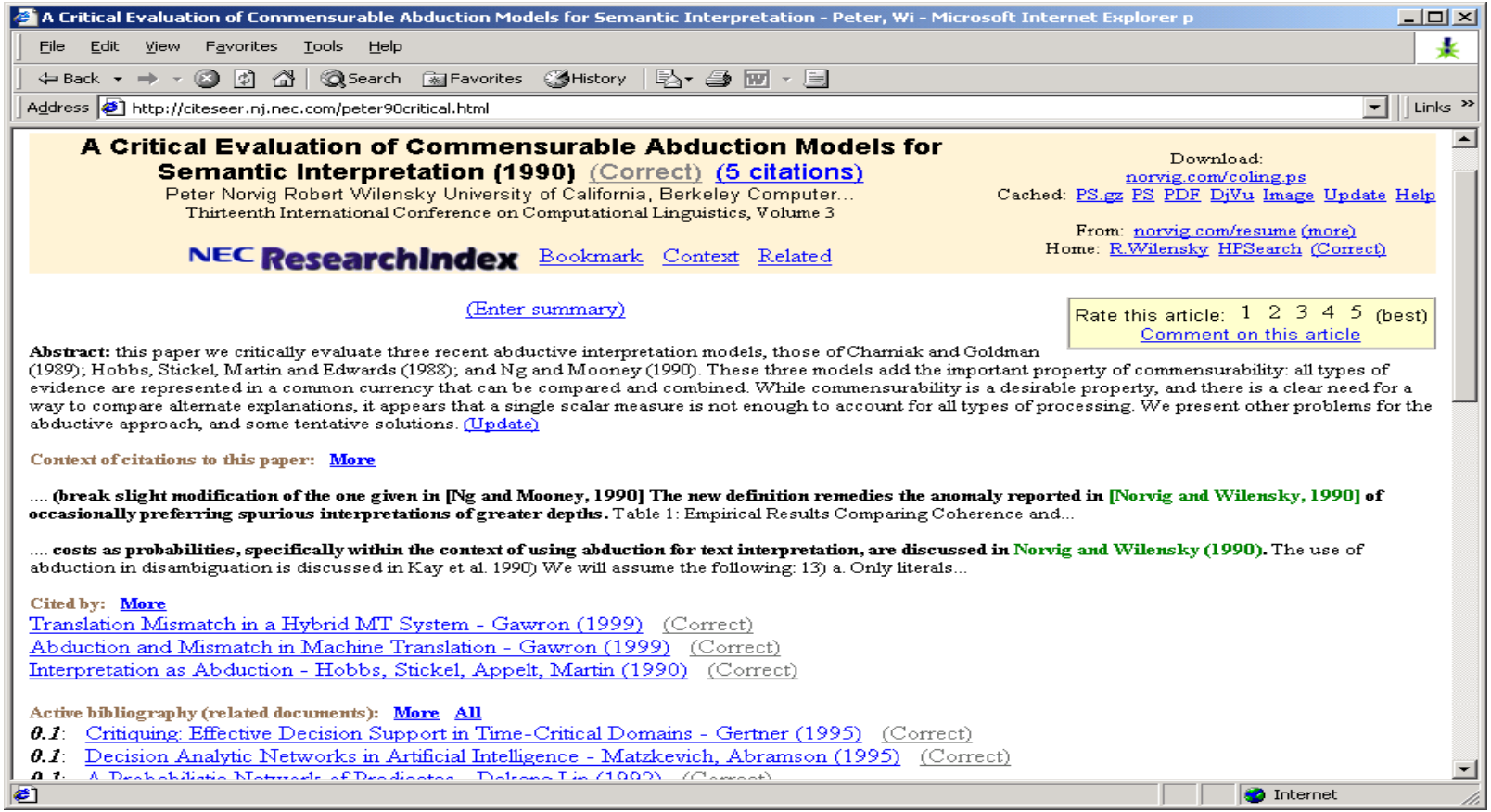
Slide from Cohen/McCallum

Often structured information in text



Slide from Cohen/McCallum

Another Problem



Slide from Cohen/McCallum

Definition of IE

Information Extraction (IE) is the process of extracting structured information (e.g., database tables) from unstructured machine-readable documents (e.g., Web documents).

Elvis Presley was a famous rock singer.
...

Mary once remarked that the only attractive thing about the painter Elvis Hunter was his first name.

Information Extraction

GName	FName	Occupation
Elvis	Presley	singer
Elvis	Hunter	painter
...	...	

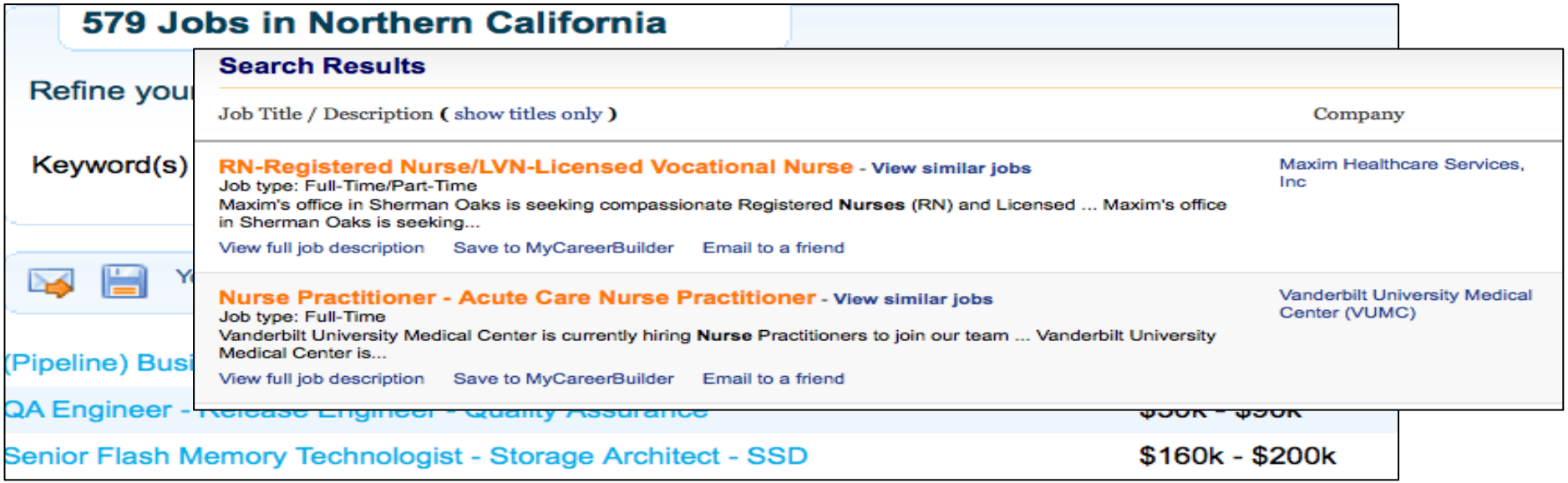
"Seeing the Web as a table"

Slide from Suchanek

Defining an IE problem

- In what I will refer to as "classic" IE, we are converting documents to one or more table entries
  - There are other kinds of IE, we will talk about those later
- The **design** of these tables is usually determined by some business need
- Let's look at the table entries for a similar set of examples to the ones we just saw

Motivating Examples



Title	Type	Location
Business strategy Associate	Part time	Palo Alto, CA
Registered Nurse	Full time	Los Angeles
…	…	

Slide from Suchanek

Motivating Examples

Biography for

Elvis Presley More at IMD

Date of Birth
8 January 1935, Tupelo, Mississippi, USA

Date of Death
16 August 1977, Memphis, Tennessee, USA (cardiac arrhythmia)

Birth Name
Elvis Aron Presley

Nickname
The Pelvis
The King
The King Of Rock 'n

Height
6' (1.83 m)

Mini Biography

Elvis Aaron Presley

Name	Birthplace	Birthdate
Elvis Presley	Tupelo, MI	1935-01-08
…	…	

Slide from Suchanek

Motivating Examples

Information Extraction:
Techniques and Challenges

Ralph Grishman

Information Integration Papers

Answering Queries Using Templates With Binding Patterns. In PODS 1995, specify binding patterns.

1

The TSIMMIS Approach to Mediation: Data Models and Languages. A surv
appears in *J. Intelligent Information Systems* **8**:2, pp. 117-132, March, 1997.

Author	Publication	Year
Grishman	Information Extraction...	2006
…	…	…

Slide from Suchanek

Motivating Examples

Ballroom Dance Shoe
1 new from $49.95
(5)
› Show only So Danca items

X-Strap Ba
1 new fr
› Show o

Compare

Dynex™ - 32" Class / 720p / 60Hz / LCD HDTV
Model: DX-32L150A11 | **SKU:** 9558089
3.8 of 5 (180 reviews)
Check Shipping & Availability ▸

Compare

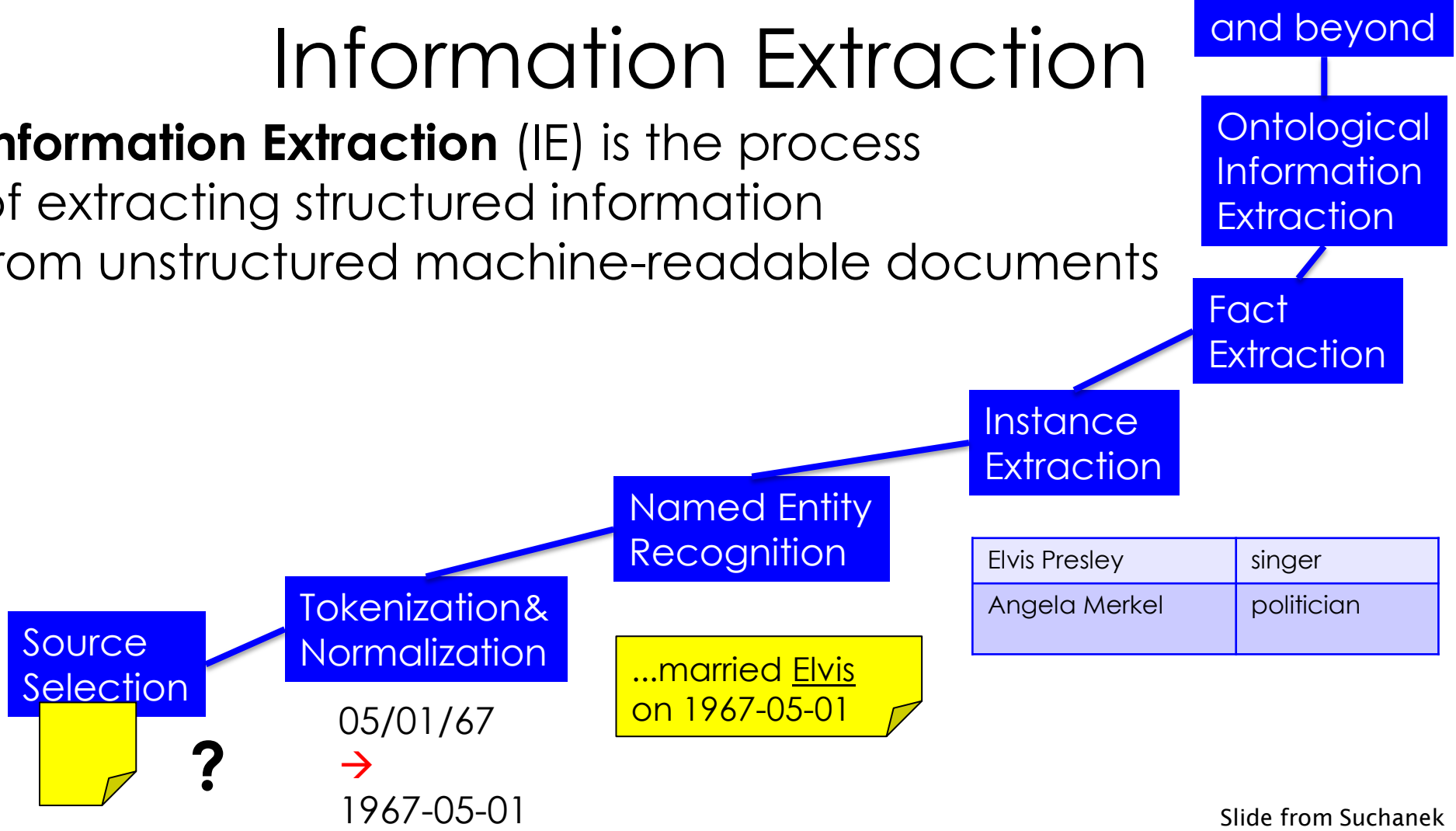
Dynex™ - 24" Class / 1080p / 60Hz / LCD HDTV
Model: DX-24L150A11 | **SKU:** 9848048
4.3 of 5 (54 reviews)
Check Shipping & Availability ▸

Product	Type	Price
Dynex 32"	LCD TV	$1000
…	…	

Slide from Suchanek

Information Extraction

Information Extraction (IE) is the process of extracting structured information from unstructured machine-readable documents

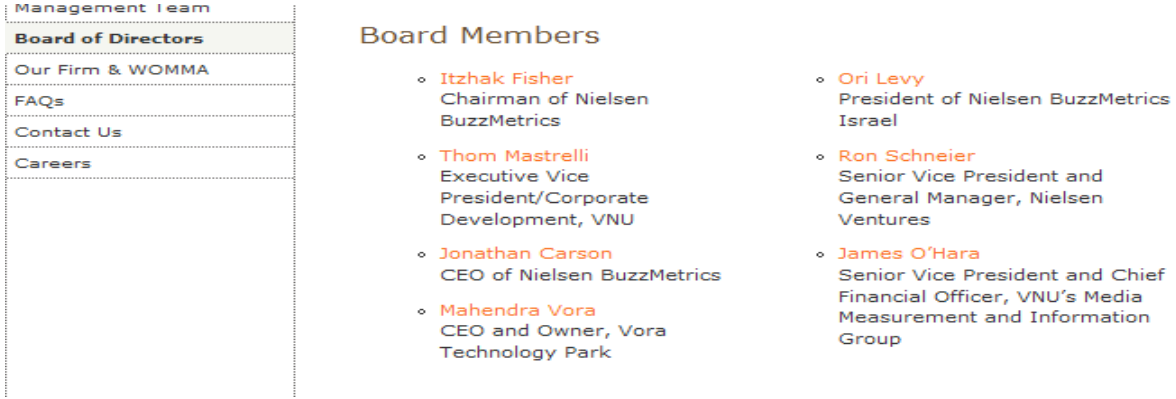


Slide from Suchanek

Information Extraction

Traditional definition: Recovering structured data from text

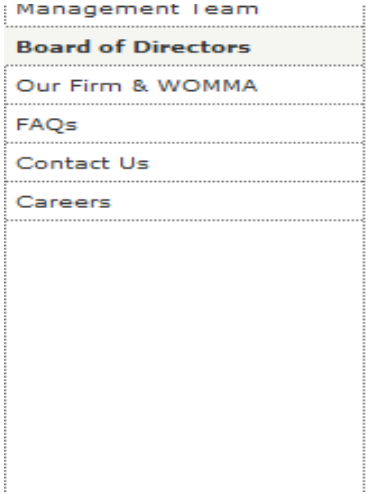
What are some of the sub-problems/challenges?

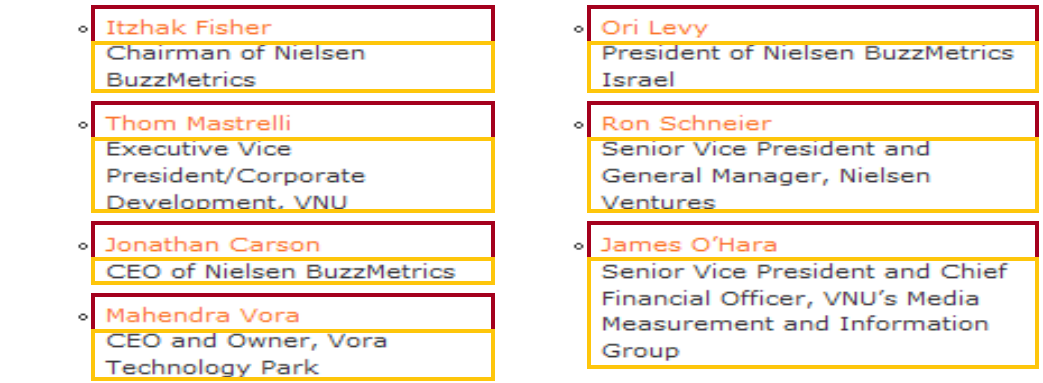


Slide from Nigam/Cohen/McCallum

Information Extraction?

- Recovering structured data from text
  - Identifying fields (e.g. named entity recognition)

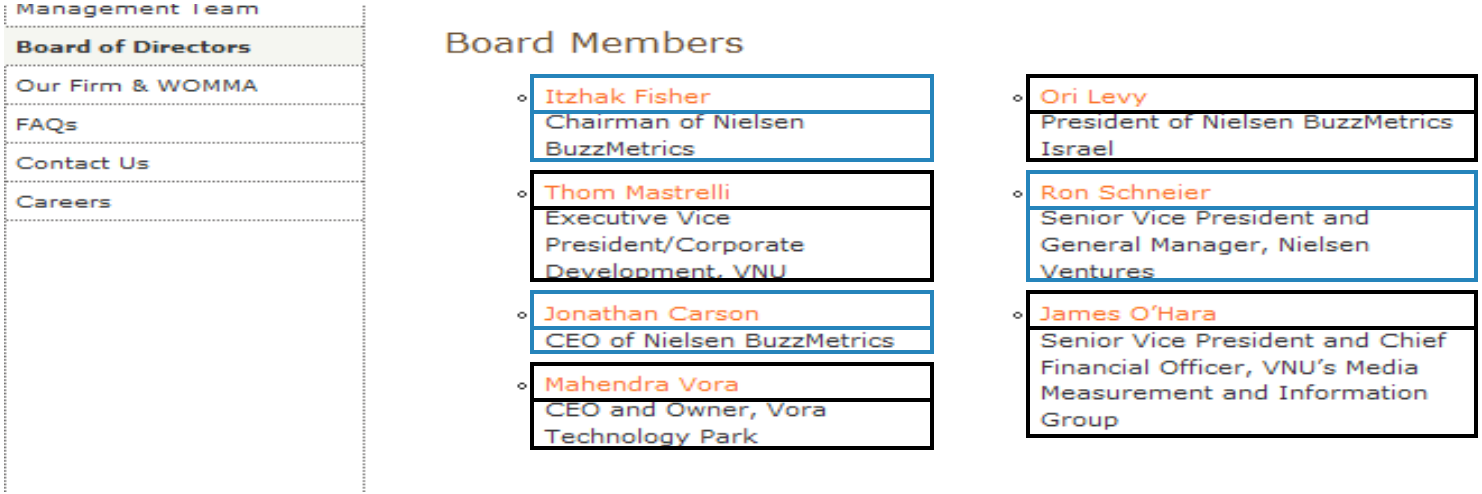




Slide from Nigam/Cohen/McCallum

Information Extraction?

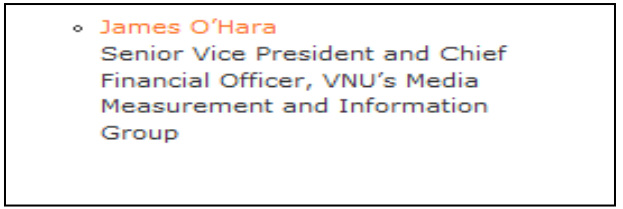
- Recovering structured data from text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)

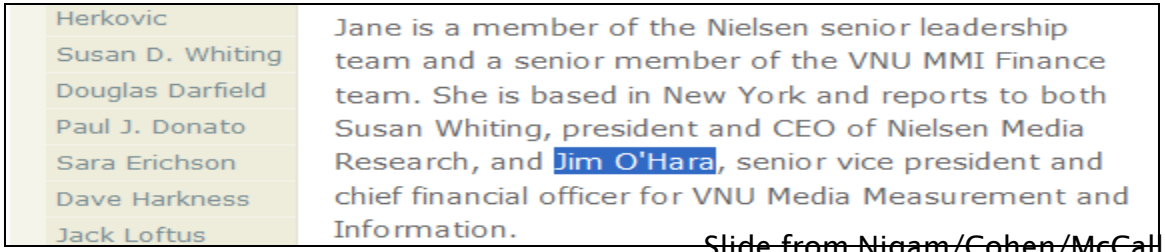


Slide from Nigam/Cohen/McCallum

Information Extraction?

- Recovering structured data from text
  - Identifying fields (e.g. named entity recognition)
  - Understanding relations between fields (e.g. record association)
  - Normalization and deduplication

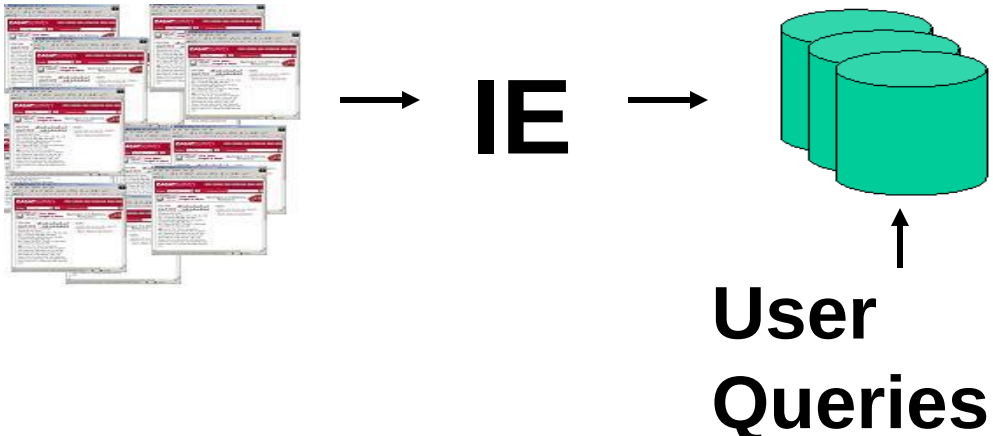




Slide from Nigam/Cohen/McCallum

Information extraction

- Input: Text Document
  - Various sources: web, e-mail, journals, …
- Output: Relevant fragments of text and relations possibly to be processed later in some automated way



Slide from McCallum

Not all documents are created equal…

- Varying regularity in document collections
- Natural or unstructured
  - Little obvious structural information
- Partially structured
  - Contain some canonical formatting
- Highly structured
  - Often, automatically generated

Slide from McCallum

Natural Text: MEDLINE Journal Abstracts

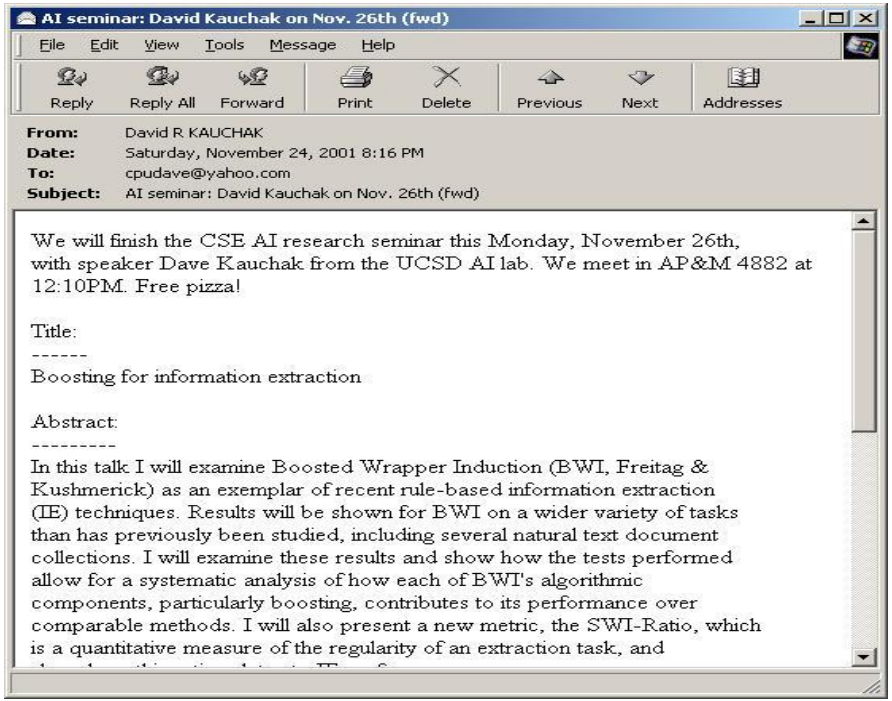
Extract number of subjects, type of study, conditions, etc.

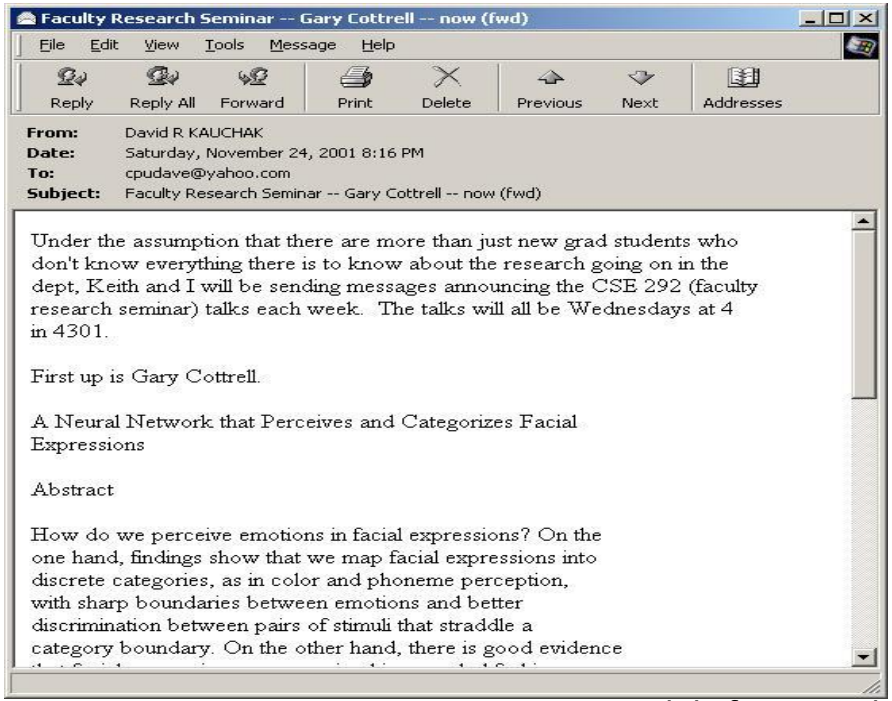
BACKGROUND: The most challenging aspect of revision hip surgery is the management of bone loss. A reliable and valid measure of bone loss is important since it will aid in future studies of hip revisions and in preoperative planning. We developed a measure of femoral and acetabular bone loss associated with failed total hip arthroplasty. The purpose of the present study was to measure the reliability and the intraoperative validity of this measure and to determine how it may be useful in preoperative planning. METHODS: From July 1997 to December 1998, forty-five consecutive patients with a failed hip prosthesis in need of revision surgery were prospectively followed. Three general orthopaedic surgeons were taught the radiographic classification system, and two of them classified standardized preoperative anteroposterior and lateral hip radiographs with use of the system. Interobserver testing was carried out in a blinded fashion. These results were then compared with the intraoperative findings of the third surgeon, who was blinded to the preoperative ratings. Kappa statistics (unweighted and weighted) were used to assess correlation. Interobserver reliability was assessed by examining the agreement between the two preoperative raters. Prognostic validity was assessed by examining the agreement between the assessment by either Rater 1 or Rater 2 and the intraoperative assessment (reference standard). RESULTS: With regard to the assessments of both the femur and the acetabulum, there was significant agreement ($p < 0.0001$) between the preoperative raters (reliability), with weighted kappa values of >0.75.

Slide from Kauchak

Partially Structured: Seminar Announcements

Extract time, location, speaker, etc.

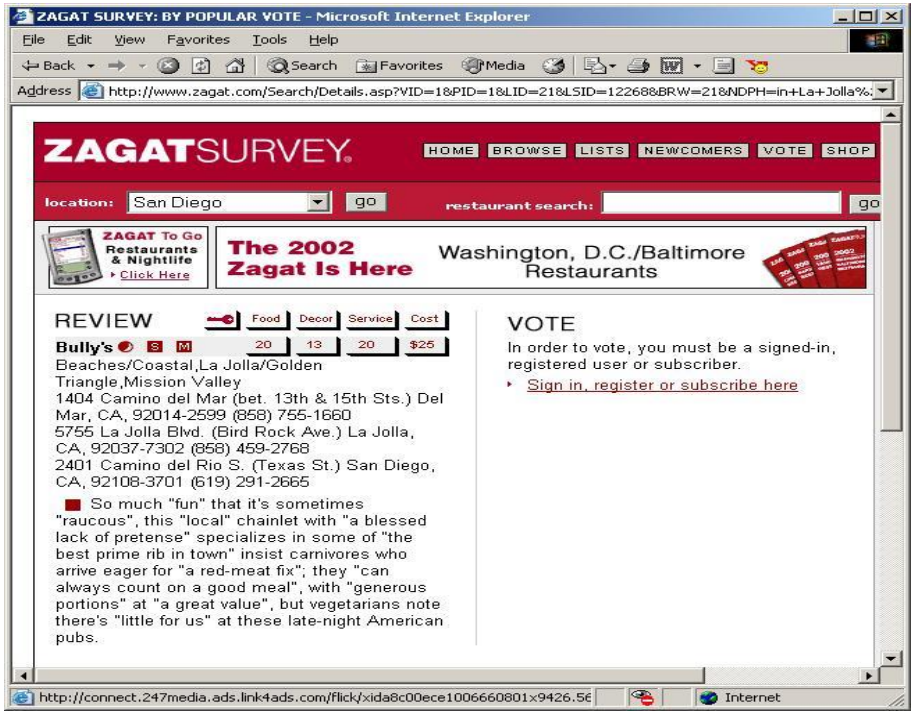


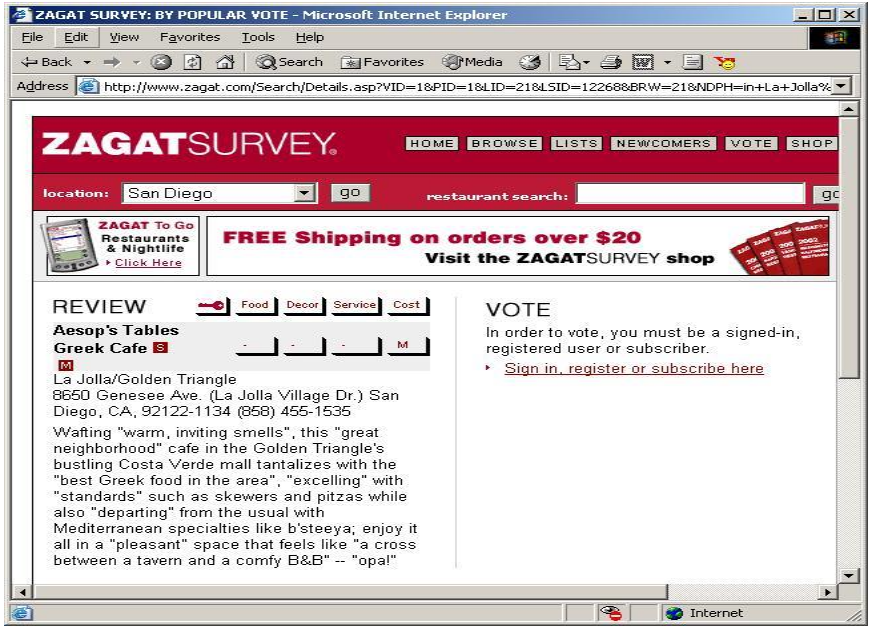


Slide from Kauchak

Highly Structured: Zagat's Reviews

Extract restaurant, location, cost, etc.





Slide from Kauchak

Information extraction pipeline

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "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 McCallum

The Full Task of Information Extraction

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

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.

Now 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."

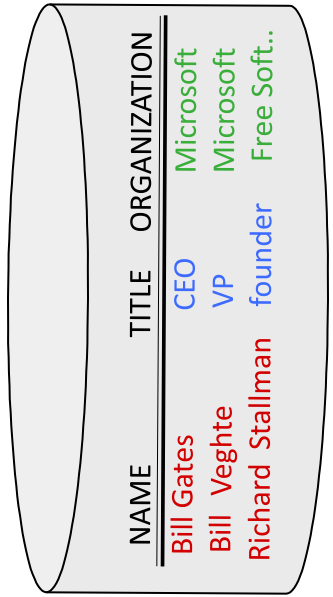
Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Gates
Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation



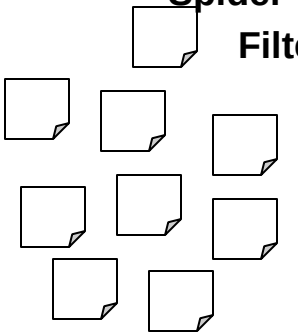
Slide from McCallum

An Even Broader View

Create ontology

Spider

Filter by relevance



Document collection

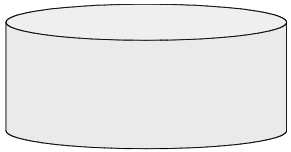
IE

Segment
Classify
Associate
Cluster

Train extraction models

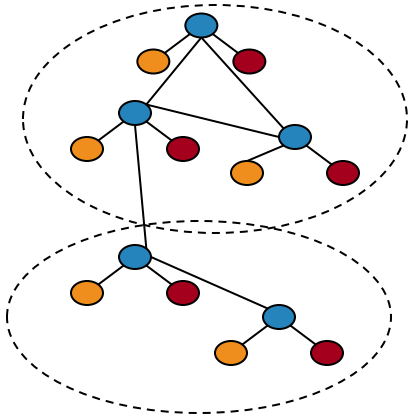
Label training data

Load DB



Database

Query, Search



Data mine

Slide from McCallum

Landscape of IE Tasks: Document Formatting

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University.

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*
9:30 - 10:00 AM	Coffee Break
10:00 - 11:30 AM	Technical Paper Sessions:

Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz and Gerhard Lakemeyer*	131: A Comparative Study of Logic Programs with Preference *Torsten Schaub and Kewen*	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation *Adnan Darwiche and Pierre Marquis*	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

Slide from McCallum

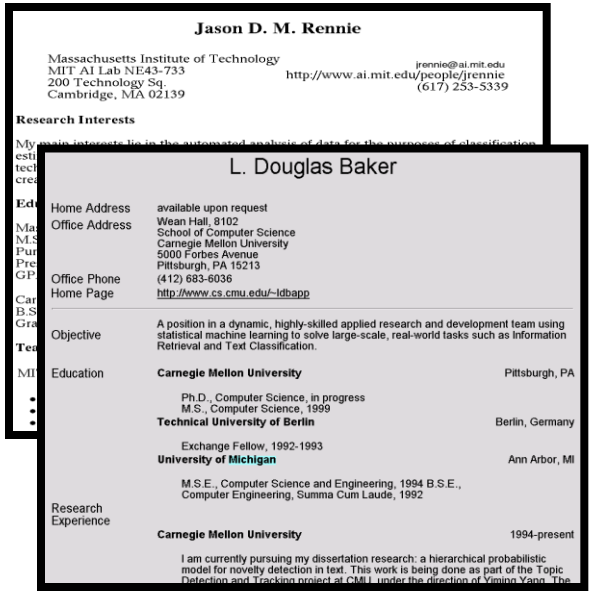
Landscape of IE Tasks
Intended Breadth of Coverage

Web site specific
Formatting

Amazon.com Book Pages

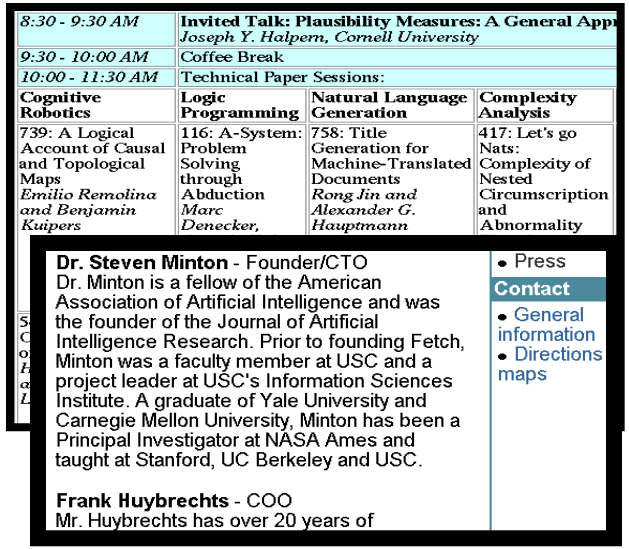
Genre specific
Layout

Resumes



Wide, non-specific
Language

University Names



Slide from McCallum

Landscape of IE Tasks :
Complexity of entities/relations

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office is 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, Al

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Slide from McCallum

Landscape of IE Tasks:
Arity of relation

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Slide from McCallum

Association task = Relation Extraction

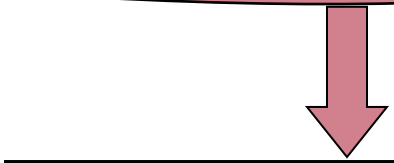
- Checking if groupings of entities are instances of a relation

1. Manually engineered rules
   - Rules defined over words/entities: “<company> located in <location>”
   - Rules defined over parsed text:
     - “((Obj <company>) (Verb located) (*) (Subj <location>))”
2. Machine Learning-based
   - Supervised: Learn relation classifier from examples
   - Partially-supervised: bootstrap rules/patterns from “seed” examples

Slide from Manning

Relation Extraction: 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…



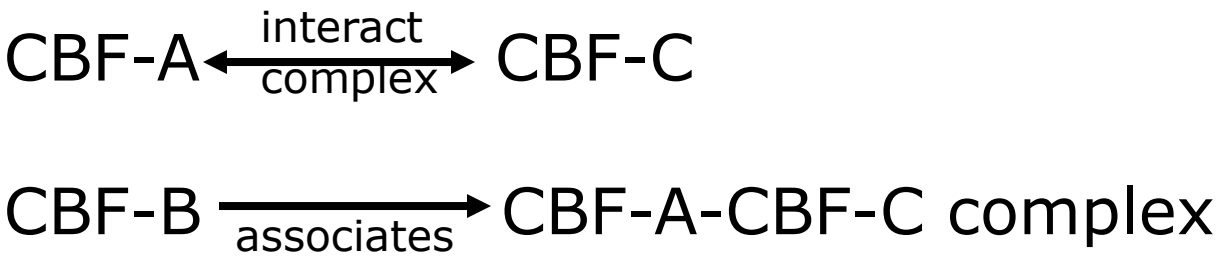
Information Extraction System

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

Slide from Manning

Relation Extraction: 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."



Slide from Manning

Binary Relation Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person Person Role

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Slide from Manning

Resolving coreference (both within and across documents)

John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tue
29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; R
turn, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political fi
was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline
years and attended Edward Devotion School, Noble and Greenough Lower School, and the Dexter
School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in
Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, N
where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become
President).[8] Kennedy spent summers with his family at their home in Hyannisport,
Massachusetts, and Christmas and Easter holidays with his family at their winter home i
Beach, Florida. For the 5th through 7th grade, Kennedy attended Riverdale Country School, a
private school for boys. For 8th grade in September 1930, the 13-year old Kennedy attended
Canterbury School in New Milford, Connecticut.

Slide from Manning

Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

- Errors cascade (error in entity tag → error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks

Slide from Manning

What we will cover in this class (briefly)

- History of IE, Related Fields
- Source Selection (which text?)
- Tokenization and Normalization
- Named Entity Recognition
- Instance Extraction
- Fact/Event Extraction
- Ontological IE/Open IE
- Probably: multilingual extraction
- Some of your suggestions, which you will give in the practical session

Seminar

- You attend EITHER Thusdays (starting tomorrow) or Wednesdays (starting next week)
- Survey: PUT YOUR NAME ON THIS
- Quiz/feedback: optionally *anonymous*

- Also, don't forget the reading for next week!
- Sarawagi: Information Extraction. Introduction

- Thank you for your attention!