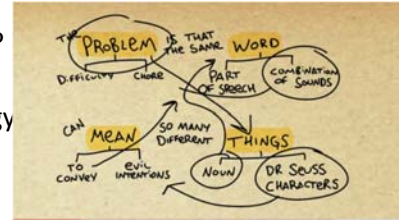


## Last class ...

# 600.465 Connecting the dots - II (NLP in Practice)

Delip Rao  
delip@jhu.edu

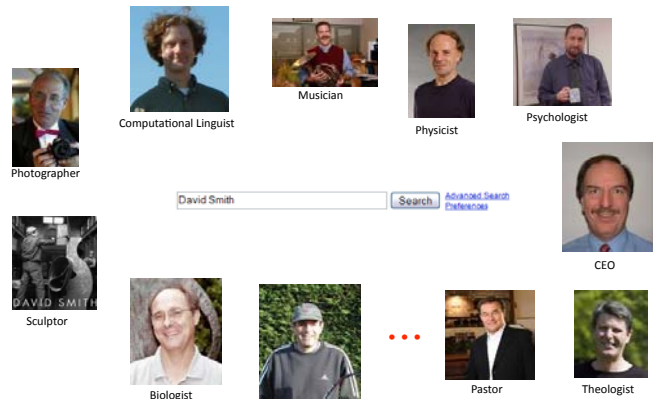
- Understood how to solve and ace in NLP tasks
- general methodology approaches
- End-to-End development using an example task
  - Named Entity Recognition



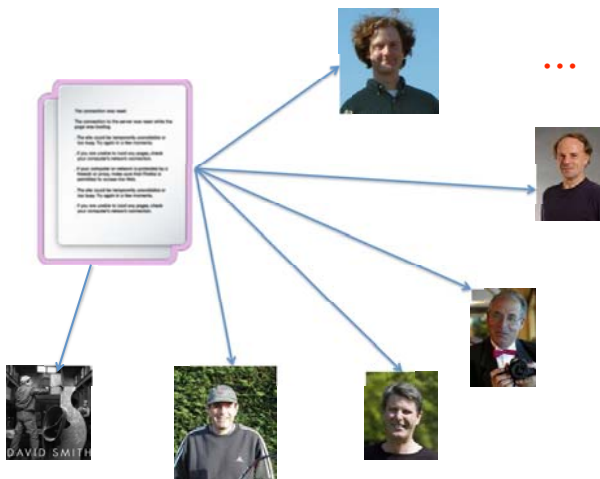
## Shared Tasks: NLP in practice

- Shared Task (aka Evaluations)
  - Everybody works on a (mostly) common dataset
  - Evaluation measures are defined
  - Participants get ranked on the evaluation measures
  - Advance the state of the art
  - Set benchmarks
- Tasks involve common hard problems or new interesting problems

## Person Name Disambiguation

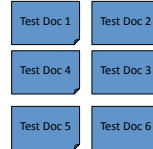


Rao, Garera & Yarowsky, 2007



## Clustering using web snippets

Goal: To cluster 100 given test documents for name "David Smith"



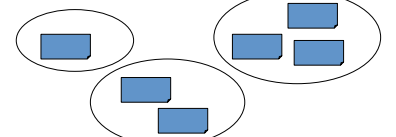
Step 1: Extract top 1000 snippets from Google



Step 2: Cluster all the 1100 documents together



Step 3: Extract the clustering of the test documents



Rao, Garera & Yarowsky, 2007

# Web Snippets for Disambiguation

Snippet

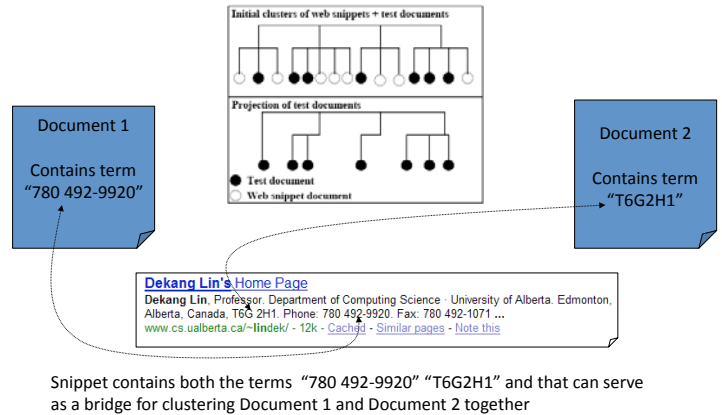
**Dekang Lin's Home Page**  
 Dekang Lin, Professor, Department of Computing Science - University of Alberta, Edmonton, Alberta, Canada, T6G 2H1. Phone: 780 492-9920. Fax: 780 492-1071 ...  
[www.cs.ualberta.ca/~lindel/](http://www.cs.ualberta.ca/~lindel/) - 12k - Cached - Similar pages - Note this

**Downloads**  
 Minipar, Minipar is a principle-based broad coverage parser. The version that is downloadable from here contains: HMM, A HMM Package in C++ ...  
[www.cs.ualberta.ca/~lindel/downloads.htm](http://www.cs.ualberta.ca/~lindel/downloads.htm) - 6k - Cached - Similar pages - Note this  
[\[ More results from www.cs.ualberta.ca \]](#)

- Snippets contain high quality, low noise features
- Easy to extract
- Derived from sources other than the document (e.g., link text)

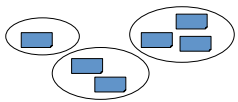
Rao, Garera & Yarowsky, 2007

# Term bridging via Snippets



Rao, Garera & Yarowsky, 2007

# Evaluating Clustering output



Dispersion: Inter-cluster

$$D_j = \sum_{\forall i \text{ in cluster } j} d(V_i, C_j)$$

$$D = \sum_j D_j$$

Silhouette: Intra-cluster

$$AvgSil = \frac{1}{n} \sum_i \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Other metrics:  
 Purity  
 Entropy  
 V-measure

# Entity Linking



## John Williams

Richard Kaufman goes a long way back with **John Williams**. Trained as a classical violinist, Californian Kaufman started doing session work in the Hollywood studios in the 1970s. One of his movies was Jaws, with **Williams** conducting his score in recording sessions in 1975...

## Michael Phelps

Debbie Phelps, the mother of swimming star **Michael Phelps**, who won a record eight gold medals in Beijing, is the author of a new memoir, ...

**Michael Phelps** is the scientist most often identified as the inventor of PET, a technique that permits the imaging of biological processes in the organ systems of living individuals. **Phelps** has ...

John Williams	author	1922-1994
J. Lloyd Williams	botanist	1854-1945
John Williams	politician	1955-
John J. Williams	US Senator	1904-1988
John Williams	Archbishop	1582-1650
<b>John Williams</b>	<b>composer</b>	<b>1932-</b>
Jonathan Williams	poet	1929-

Michael Phelps	swimmer	1985-
Michael Phelps	biophysicist	1939-

Identify matching entry, or determine that entity is missing from KB

# Challenges in Entity Linking

- Name Variation
  - Abbreviations: BSO vs. Boston Symphony Orchestra
  - Shortened forms: Osama Bin Laden vs. Bin Laden
  - Alternate spellings: Osama vs. Ussamah vs. Oussama
- Entity Ambiguity: Polysemous mentions
  - E.g., Springfield, Washington
- Absence: Open domain linking
  - Not all observed mentions have a corresponding entry in KB (NIL mentions)
  - Ability to predict NIL mentions determines KBP accuracy
  - Largely overlooked in current literature

# Entity Linking: Features

- Name-matching
  - acronyms, aliases, string-similarity, probabilistic FST
- Document Features
  - TF/IDF comparisons, occurrence of names or KB facts in the query text, Wikitology
- KB Node
  - Type (e.g., is this a person), Features of Wikipedia page, Google rank of corresponding Wikipedia page
- Absence (NIL Indications)
  - Does any candidate look like a good string match?
- Combinations
  - Low-string-match AND Acronym AND Type-is-ORG

## Entity Linking: Name Matching

- Acronyms
- Alias Lists
  - Wikipedia redirects, stock symbols, misc. aliases
- Exact Match
  - With and without normalized punctuation, case, accents, appositive removal
- Fuzzier Matching
  - Dice score (character uni/bi/tri-grams), Hamming, Recursive LCSubstring, Subsequences
  - Word removal (e.g., Inc., US) and abbrev. expansion
- Weighted FST for Name Equivalence
  - Trained models score name-1 as a re-writing of name-2

## Entity Linking: Document Features

- BoW Comparisons
  - TF/IDF & Dice scores for news article and KB text
  - Examined entire articles and passages around query mentions
- Named-Entities
  - Ran BBN's SERIF analyzer on articles
  - Checked for coverage of (1) query co-references and (2) all names/nominals in KB text
  - Noted type, subtype of query entity (e.g., ORG/Media)
- KB Facts
  - Looked to see if candidate node's attributes are present in article text (e.g., spouse, employer, nationality)
- Wikitology
  - UMBC system predicts relevant Wikipedia pages (or KB nodes) for text

## Question Answering

Why is the sky blue?

Input interpretation:  
Why is the sky blue?

Result:  
The sky's blue color is a result of the effect of Rayleigh scattering. Shorter-wavelength blue light is more strongly scattered in the earth's atmosphere than longer-wavelength red light; the human eye perceives the color blue when looking at the sky as a result.

Computed by: Wolfram Mathematica      Download as: PDF | Live Mathematica

## Question Answering: Ambiguity

Where is Washington?

Assuming "Washington" is a city | Use as a US state or a company instead  
Assuming Washington (District of Columbia, USA) | Use Washington (Pennsylvania, USA) or more instead

Input interpretation:  
Washington, District of Columbia location

Location:  
 World map | Show coordinates

Who founded Google?

Assuming "Who founded" is referring to internet domains | Use as a historical event instead | Use "Who" as a class of countries  
Assuming "Google" is an internet domain | Use as a financial entity instead

Input interpretation:  
google.com Corporate information

Corporate information:      More | Show map | Location details

company name	Google, Inc.
location	Mountain View, California, United States
market cap	\$177.7 billion
employees	23 331 people

Who founded Google?

Assuming "Who founded" is a historical event | Use as referring to internet domains instead | Use "Who" as a class of countries  
Assuming "Google" is an internet domain | Use as a financial entity instead

Input interpretation:  
The World Health Organization is established date  
at google.com (domain)

Result:  
April 7, 1948

## More complication: Opinion Question Answering

**Q:** What is the international reaction to the reelection of Robert Mugabe as President of Zimbabwe?

**A:** African observers **generally approved** of his victory while Western Governments **strongly denounced** it.

Stoyanov, Cardie, Wiebe 2005

Somasundaran, Wilson, Wiebe, Stoyanov 2007

## Subjectivity and Sentiment Analysis

- The linguistic expression of somebody's **opinions, sentiments, emotions, evaluations, beliefs, speculations (private states)**

	Sentiment Analysis	Subjectivity analysis	
obj:	Positive	Subjective	grammar of
Quii	Negative		
the	Neutral		

- Subjectivity analysis classifies content in **objective** or **subjective**

Thanks: Jan Wiebe

SUNDAY, MAY 13, 2007

### Seth McFarlane nailed to the wall

I don't like *Family Guy* but love *South Park*. Maybe it's because Seth McFarlane steals ideas from other shows and can't come up with his own. There are several comparison clips on YouTube showing how he outright stole ideas from *The Simpsons*. For years I've been saying that *Family Guy* was a poor man's *The Simpsons*. I stopped saying this when it became clear that fans of *Family Guy* didn't really care one way or the other. I've seen many examples of *Family Guy*'s second rate brand of comedy but now someone has gone that extra step.

Rent [Family Guy: Volume 5](#)  
Rent *Family Guy: Volume 5* on DVD.  
No late fees. Over 10,000 titles!  
Choose Netflix.com

[Family Guy on TBS](#)  
It's comedy for all tastes on TBS.  
View episode guide & message board  
www.tbs.com

[South Park](#)  
Tons of Videos, Pics & More Online.  
Watch *South Park* Wed at 10p / 9c!  
www.southparkstudios.com

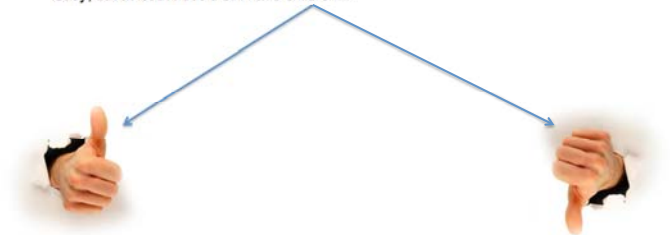
[South Park](#)  
Download Episodes of Your Favorite  
TV Shows Now! Fast & Easy!  
Video AOL.com

Rao & Ravichandran, 2009

## Subjectivity & Sentiment: Applications

### Cuisinart DGB-600BC Grind and Brew Coffee maker:

This coffee-maker does so much! It makes weak, watery coffee! It grinds beans if you want it to! It inexplicably floods the entire counter with half-brewed coffee when you aren't looking! Perhaps it could be used to irrigate crops... It is time-consuming to clean, but in fairness I should also point out that the stainless-steel thermal carafe is a durable item that has withstood being hurled onto the floor in rage several times. And if all these features weren't enough, it's pretty expensive too. If faced with the choice between having a car door repeatedly slamming into my genitalia and buying this coffee-maker, I'd unhesitatingly choose the Cuisinart! The coffee would be lousy, but at least I could still have children...



## Sentiment classification

- Document level
- Sentence level
- Product feature level
  - “For a heavy pot, the **handle** is not well designed.”
- Find opinion holders and their opinions

## Subjectivity & Sentiment: More Applications



“Who is the fairest one of all,  
and state your sources!”

**Product review mining:**  
Best Android phone in the market?

# Sentiment tracking

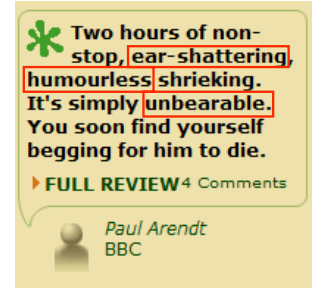
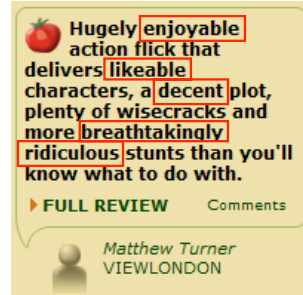


Tracking sentiments toward topics over time:

Is anger ratcheting up or cooling down?

Source: Research.ly

# Sentiment Analysis Resources : Lexicons



Rao & Ravichandran, 2009

# Sentiment Analysis Resources: Lexicons

amazing +	क्रूर -	جميل +
banal -	मोहित +	ممتاز +
bewilder -	शान्त +	قيبح -
divine +	शक्तिशाली +	سلمي +
doldrums -	ब्रेमजा -	فظيح -
...	...	...
aburrido -	magnifique +	
inocente +	céleste +	
mejor +	irrégulier -	
sabroso +	haine -	
odiar -	...	
....		

Rao & Ravichandran, 2009

# Sentiment Analysis Resources : Corpora

464 of 488 people found the following review helpful:

★★★★★ **Not the perfect "do-it-all" device, but very close to being the perfect e-reading device.** August 26, 2010  
 By **C. Vincent** (Rochester, NY) - See all my reviews

This review is from: **Kindle 3G Wireless Reading Device, Free 3G + Wi-Fi, 6" Display, 3G Works Globally - Latest Generation (Electronics)**

I woke up to a nice surprise this morning: a new kindle as a gift. I have an iPad and a Kindle DX, but I guess someone heard my complaints of them being too heavy and difficult to do extended-reading on. Don't get me wrong, I absolutely love my iPad and DX, but this new generation of Kindle is perfect for reading outside and for long

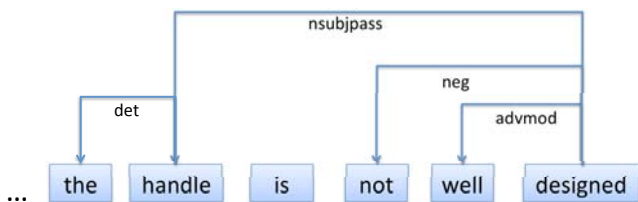
First, the good stuff. Kindle 3 has a very readable high contrast screen. The form factor is small and light enough to be a book replacement. Purchasing and downloading content is simple and convenient.

However, I had a series of problems which eventually led me to return my Kindle.

- Pang and Lee, Amazon review corpus
- Blitzer, multi-domain review corpus

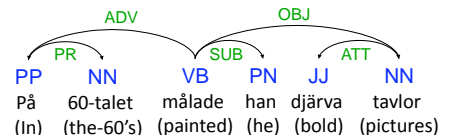
# Dependency Parsing

- Consider product-feature opinion extraction
  - “For a heavy pot, the **handle** is not well designed.”



# Dependency Representations

- Directed graphs:
  - V is a set of nodes (tokens)
  - E is a set of arcs (dependency relations)
  - L is a labeling function on E (dependency types)
- Example:



thanks: Nivre

## Dependency Parsing: Constraints

- Commonly imposed constraints:
  - Single-head (at most one head per node)
  - Connectedness (no dangling nodes)
  - Acyclicity (no cycles in the graph)
  - Projectivity:
    - An arc  $i \rightarrow j$  is projective iff, for every  $k$  occurring between  $i$  and  $j$  in the input string,  $i \rightarrow k$ .
    - A graph is projective iff every arc in  $A$  is projective.

thanks: Nivre

## Syntactic Variations versus Semantic Roles

*Yesterday, Kristina hit Scott with a baseball*

*Scott was hit by Kristina yesterday with a baseball*

*Yesterday, Scott was hit with a baseball by Kristina*

*With a baseball, Kristina hit Scott yesterday*

*Yesterday Scott was hit by Kristina with a baseball*

*The baseball with which Kristina hit Scott yesterday was hard*

*Kristina hit Scott with a baseball yesterday*



thanks: Jurafsky

## Dependency Parsing: Approaches

- Link grammar (Sleator and Temperley)
- Bilexical grammar (Eisner):
  - Lexicalized parsing in  $O(n^3)$  time
- Maximum Spanning Tree (McDonald)
- CONLL 2006/2007

## Semantic Role Labeling

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.
  - agent patient source destination instrument
  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”

thanks: Mooney

## SRL Datasets

- FrameNet:
  - Developed at UCB
  - Based on notion of Frames
- PropBank:
  - Developed at UPenn
  - Based on elaborating the Treebank
- Salsa:
  - Developed at Universität des Saarlandes
  - German version of FrameNet

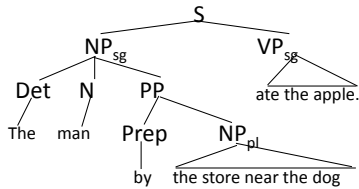
## SRL as Sequence Labeling

- SRL can be treated as an sequence labeling problem.
- For each verb, try to extract a value for each of the possible semantic roles for that verb.
- Employ any of the standard sequence labeling methods
  - Token classification
  - HMMs
  - CRFs

thanks: Mooney

## SRL with Parse Trees

- Parse trees help identify semantic roles through exploiting syntactic clues like “the agent is usually the subject of the verb”.
- Parse tree is needed to identify the true subject.



“The man by the store near the dog ate an apple.”

“The man” is the agent of “ate” not “the dog”.

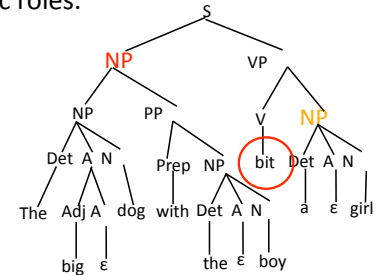
thanks: Mooney

## SRL with Parse Trees

- Assume that a syntactic parse is available.
- For each predicate (verb), label each node in the parse tree as either not-a-role or one of the possible semantic roles.

Color Code:

- not-a-role
- agent
- patient
- source
- destination
- instrument
- beneficiary



thanks: Mooney

## Selectional Restrictions

- Selectional restrictions** are constraints that certain verbs place on the filler of certain semantic roles.
  - Agents should be animate
  - Beneficiaries should be animate
  - Instruments should be tools
  - Patients of “eat” should be edible
  - Sources and Destinations of “go” should be places.
  - Sources and Destinations of “give” should be animate.
- Taxonomic abstraction hierarchies or ontologies (e.g. hypernym links in WordNet) can be used to determine if such constraints are met.
  - “John” is a “Human” which is a “Mammal” which is a “Vertebrate” which is an “Animate”

thanks: Mooney

## Word Senses

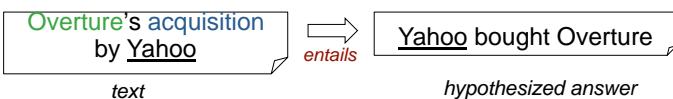
Beware of the burning **coal** underneath the **ash**.

Ash	Coal
<ul style="list-style-type: none"> <li><b>Sense 1</b> Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.</li> <li><b>Sense 2</b> The <b>solid</b> residue left when <b>combustible</b> material is thoroughly <b>burned</b> or oxidized.</li> <li><b>Sense 3</b> To convert into ash</li> </ul>	<ul style="list-style-type: none"> <li><b>Sense 1</b> A piece of glowing carbon or <b>burnt</b> wood.</li> <li><b>Sense 2</b> charcoal.</li> <li><b>Sense 3</b> A black <b>solid combustible</b> substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for <b>burning</b></li> </ul>

Self-training via Yarowsky’s Algorithm

## Recognizing Textual Entailment

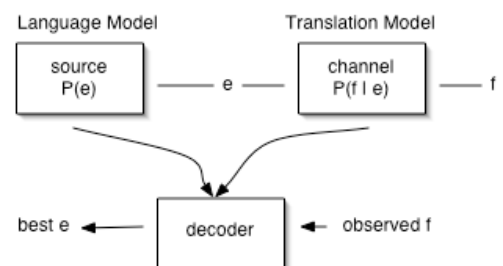
Question: Who bought Overture? >> Expected answer form: X bought Overture



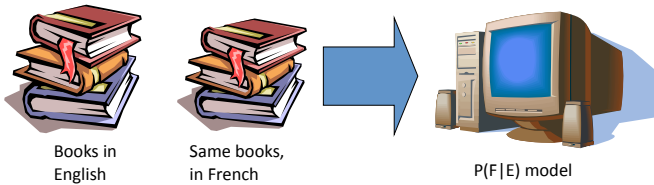
- Similar for IE: X acquire Y
- Similar for “semantic” IR
- Summarization (multi-document)
- MT evaluation

thanks: Dagan

## (Statistical) Machine Translation



## Where will we get $P(F|E)$ ?



We call collections stored in two languages **parallel corpora** or **parallel texts**

Want to update your system? Just add more text!

*thanks: Nigam*

## Machine Translation

- Systems
  - Early rule based systems
  - Word based models (IBM models)
  - Phrase based models (log-linear!)
  - Tree based models (syntax driven)
  - Adding semantics (WSD, SRL)
  - Ensemble models
- Evaluation
  - Metrics (BLEU, BLACK, ROUGE ...)
  - Corpora (statmt.org)

EGYPT  
GIZA++  
MOSES  
JOSHUA

## Allied Areas and Tasks

- Information Retrieval
  - TREC (Large scale experiments)
  - CLEF (Cross Lingual Evaluation Forum)
  - NTCIR
  - FIRE (South Asian Languages)

## Allied Areas and Tasks

- (Computational) Musicology
  - MIREX

Audio Test/Train tasks
Audio Cover Song Identification
Audio Tag Classification
Audio Music Similarity and Retrieval
Symbolic Music Similarity and Retrieval
Audio Onset detection
Audio Key detection
Real-time Audio to Score Alignment (a.k.a Score Following)
Query by Singing/Humming
Audio Melody Extraction
Multiple Fundamental Frequency Estimation & Tracking
Audio Chord Estimation
Query by Tapping
Audio Beat Tracking
Audio Structural Segmentation

## Where Next?

