INTRODUCTION TO NLP AND WATSON

Heng Ji
jih@rpi.edu
January 22, 2014
Outline

• Course Overview
• Introduction to NLP
• Introduction to Watson
• Watson in NLP View
• Machine Learning for NLP
• Why NLP is Hard
Homepages

• Course homepage
  • http://nlp.cs.rpi.edu/course/spring14/nlp.html

• My homepage
  • http://nlp.cs.rpi.edu/hengji.html

• TA (Hongzhao Huang)’s homepage
  • https://sites.google.com/site/hongzhaohuang/

• My lab’s homepage
  • http://nlp.cs.rpi.edu
Goal and Topics of This Course

• Introduce you to Natural Language Processing (NLP) problems & solutions; and some effective machine learning tools, using Deep Question Answering (Watson) as a case study
  • We’re going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
  • We are also concerned with the insights that such computational work gives us into human processing of language.
  • It’s not a pure Watson engineering course; it’s a basic NLP science course.

• At the end you should:
  • Learn basic NLP techniques and approaches
  • Feel some ownership over the formal & statistical models
  • Learn technical paper writing and presentation skills
Course Organization

- Heng will teach most lectures
- Will also invite researchers from IBM Watson team and RPI Watson team to give three 40min-1hour guest lectures
Course Basics

• Course webpage (Check it for class notes):
  • [http://nlp.cs.rpi.edu/course/spring14/nlp.html](http://nlp.cs.rpi.edu/course/spring14/nlp.html)

• Office hour (Room 2148, Winslow Building)
  • 12:30pm-1:30pm Wednesday

• Raise up any questions and participate discussions at the mailing list/blackboard (will be setup this weekend)
• Feel free to drop me email for any questions and feedback: [jih@rpi.edu](mailto:jih@rpi.edu)
Recommended Textbooks

• The instructor will write most lecture notes based on recent papers

• But recommending the following books for background reading:
  • *Speech and Language Processing*, Daniel Jurafsky and James Martin, Prentice-Hall (second edition).
  • Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press.
  • Steven Abney. *Semi-supervised Learning for Computational Linguistics*.
  • Christopher M. Bishop. *Pattern Recognition and Machine Learning*.
  • Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning*. 
Assignments

• Written assignments will involve linguistics, math, and careful thinking
• Programming assignments: all of the above plus programming
  • Expect the programming assignments to take more time than the written assignments
• Final projects are up to your own devising
  • A good literature survey, or
  • An implementation of some novel and interesting research idea
  • You’ll need to come up with:
    • a model;
    • data to examine;
    • and a computer implementation of the model, fit to the data
• Start thinking about the project early, and start working on it early
• collaboration is encouraged
General Requirement

• Interests in languages and linguistics
• solid background in algorithms
• probabilities good programming skills
• sufficient mathematical background
Grading

• 6 Assignments (40%)
• Term Project (40%)
  • Proposal writing and presentation: 5%
  • Project implementation: 25%
  • Final Project Report writing 10%
• Final Exam (20%)
• Bonus Reading Assignment: 0-1 bonus credit if the student sends a good summary & critique about each lecture's reading materials
Class Policy

- Restricted Academic Integrity
- NO Incompleteness are accepted
- NO Late Assignments are accepted
- NO cell phones and internet surfing are allowed in the classroom
- Don't come late to the class, if you are late more than 10 minutes, simply skip it in order not to disturb the lecture
Outline

• Course Overview
  ➢ Introduction to NLP
• Introduction to Watson
• Watson in NLP View
• Machine Learning for NLP
• Why NLP is Hard
The Dream

• It’d be great if machines could
  • Process our email (usefully)
  • Translate languages accurately
  • Help us manage, summarize, and aggregate information
  • Use speech as a UI (when needed)
  • Talk to us / listen to us

• But they can’t:
  • Language is complex, ambiguous, flexible, and subtle
  • Good solutions need linguistics and machine learning knowledge

• So:
The mystery

- What’s now impossible for computers (and any other species) to do is effortless for humans
The mystery (continued)

• Patrick Suppes, eminent philosopher, in his 1978 autobiography:
  “…the challenge to psychological theory made by linguists to provide an adequate theory of language learning may well be regarded as the most significant intellectual challenge to theoretical psychology in this century.”

• So far, this challenge is still unmet in the 21st century

• Natural language processing (NLP) is the discipline in which we study the tools that bring us closer to meeting this challenge
What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
What is Natural Language Processing (NLP)

- Computers use (analyze, understand, generate) natural language

  - Text Processing
    - Lexical: tokenization, part of speech, head, lemmas
    - Parsing and chunking
    - Semantic tagging: semantic role, word sense
    - Certain expressions: named entities
    - Discourse: coreference, discourse segments

- Speech Processing
  - Phonetic transcription
  - Segmentation (punctuations)
  - Prosody
Why Should You Care?

Trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers
Any Light at The End of The Tunnel?

- Yahoo, Google, Microsoft → Information Retrieval
- Monster.com, HotJobs.com (Job finders) → Information Extraction + Information Retrieval
- Systran powers Babelfish → Machine Translation
- Ask Jeeves → Question Answering
- Myspace, Facebook, Blogspot → Processing of User-Generated Content
- Tools for “business intelligence”
- All “Big Guys” have (several) strong NLP research labs:
  - IBM, Microsoft, AT&T, Xerox, Sun, etc.
- Academia: research in an university environment
Commercial World

• Lot’s of exciting stuff going on…

Powerset

PEARSON Knowledge Technologies

Google™

Yahoo!

Nielsen BuzzMetrics

cymfony harnessing influence 2.0™

hakia® search for meaning

collective intellect

umbria
The Steps in NLP

Morphology

Syntax

Semantics

Pragmatics

Discourse

**we can go up, down and up and down and combine steps too!!

**every step is equally complex
Major Topics

1. Words
2. Syntax
3. Meaning
4. Discourse
5. Applications exploiting each
Simple Applications

- Word counters (wc in UNIX)
- Spell Checkers, grammar checkers
- Predictive Text on mobile handsets
Bigger Applications

- Intelligent computer systems
- NLU interfaces to databases
- Computer aided instruction
- Information retrieval
- Intelligent Web searching
- Data mining
- Machine translation
- Speech recognition
- Natural language generation
- Question answering
Other Popular NLP tasks: Machine Translation

<stoken id="S1">戏</stoken><ttoken id="T1" align="S5 S6">School</ttoken>
<stoken id="S2">剧</stoken><ttoken id="T2">of</ttoken>
<stoken id="S3">舞</stoken><ttoken id="T3">Theatre</ttoken>
<stoken id="S4">蹈</stoken><ttoken id="T4">and</ttoken>
<stoken id="S5">学</stoken><ttoken id="T5">Dance</ttoken>
<stoken id="S6">院</stoken><ttoken id="T6">presents</ttoken>
<stoken id="S7">演</stoken><ttoken id="T7">Wonderful</ttoken>
<stoken id="S8">出</stoken><ttoken id="T7">town</ttoken>
<stoken id="S9">奇</stoken><ttoken id="T7">妙的</ttoken>
<stoken id="S10">妙</stoken><ttoken id="T7">小镇</ttoken>
<stoken id="S11">的</stoken>
<stoken id="S12">小</stoken>
<stoken id="S13">镇</stoken>
<table>
<thead>
<tr>
<th>Section</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killing Palestinians and wounding nine in the raids Sector</td>
<td>Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.</td>
</tr>
<tr>
<td>Bashir meets Fraser, the Security Council will not impose forces Darfur</td>
<td>Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.</td>
</tr>
<tr>
<td>Rumsfield and Cheney insist on keeping the American forces in Iraq</td>
<td>Rumsfield and Cheney call Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.</td>
</tr>
<tr>
<td>Killing civilians and wounding officer suicide attack in Afghanistan</td>
<td>The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.</td>
</tr>
</tbody>
</table>
Other Popular NLP tasks: Weblog and Tweet Analytics

- Data-mining of Weblogs, discussion forums, message boards, user groups, tweets, and other forms of user generated media
  - Product marketing information
  - Political opinion tracking
  - Social network analysis
  - Buzz analysis (what’s hot, what topics are people talking about right now).
NLP for Big Data

- Huge Size
  - Google processes 20 PB a day (2008)
  - Wayback Machine has 3 PB + 100 TB/month (3/2009)
  - Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
  - eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 80% data is unstructured (IBM, 2010)
- More importantly, Heterogeneous
The Patient Protection and Affordable Care Act (PPACA),[1] commonly called the Affordable Care Act (ACA) or Obamacare, is a United States federal statute signed into law by President Barack Obama on March 23, 2010.

The U.S. Congress, still in partisan deadlock over Republican efforts to halt President Barack Obama's healthcare reforms, was on the verge of shutting down most of the U.S. government starting on Tuesday morning.

President Obama's 70-minute White House meeting late Wednesday afternoon with congressional leaders including House Speaker John Boehner, did nothing to help end the impasse.

Lockheed Martin to furlough 3,000 workers amid #USGovernmentShutdown

NSF and NIST are temporarily closed because the Government entered a period of partial shutdown.
From Big Data to Information: Go Beyond Google
From Information to Knowledge

Laurie A. Leshin, Dean of the School of Science, holds a scale model of the Mars Rover "Curiosity," 3D printed for her by an RPI grad student at her office on the RPI campus Tuesday Jan. 8, 2012, in Troy, N.Y. The NASA spacecraft is currently exploring the surface of Mars. (John Carl D'Annibale / Times Union)

Projects: MarsCuriosity

Full and freely available version of all the @MarsCuriosity #Science papers on #MarsDirt available here (right side):
Example: Depressed working mom seeking help…
What she hopes to get is something like…

<table>
<thead>
<tr>
<th>Source</th>
<th>Solution</th>
<th>Effectiveness</th>
<th>Efficacy</th>
<th>Price</th>
<th>Closest store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td></td>
<td>85%, 400</td>
<td>2 months</td>
<td>$2</td>
<td>Hongkong supermarket, NYC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>95%, 500</td>
<td>1 day</td>
<td>$10</td>
<td>Beijing Children’s Hospital</td>
</tr>
<tr>
<td>Western</td>
<td></td>
<td>50%, 100</td>
<td>1 week</td>
<td>$20</td>
<td>Duane Reade, NJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>55%, 150</td>
<td>1 week</td>
<td>$25</td>
<td>CVS, NJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>60%, 200</td>
<td>3 days</td>
<td>$30</td>
<td>CVS, NJ</td>
</tr>
</tbody>
</table>
Outline

• Course Overview
• Introduction to NLP
  ➢ Introduction to Watson
• Watson in NLP View
• Machine Learning for NLP
• Why NLP is Hard
What is Watson?

• A computer system that can compete in real-time at the human champion level on the American TV quiz show Jeopardy.

• [http://www.youtube.com/watch?v=WFR3lOM_xhE](http://www.youtube.com/watch?v=WFR3lOM_xhE)

• IBM computer designed to compete on the game show Jeopardy!

• Culmination of 4 years of dedicated research, and over a decade of advancements in algorithms pertaining to natural language parsing (NLP)
### Systems that think like humans

“The exciting new effort to make computers think... machines with minds, in the full and literal sense.” (Haugeland, 1985)

“[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning...” (Bellman, 1978)

- **Richard Bellman (1920-84)**

### Systems that think rationally

“The study of mental faculties through the use of computational models.” (Charniak and McDermott, 1985)

“The study of the computations that make it possible to perceive, reason, and act.” (Winston, 1972)

- **Aristotle (384BC -322BC)**

### Systems that act like humans

“The art of creating machines that perform functions that require intelligence when performed by people” (Kurzweil, 1990)

“The study of how to make computers do things at which, at the moment, people are better (Rich and Knight, 1991)

- **Alan Turing (1912-1954)**

### Systems that act rationally

“The branch of computer science that is concerned with the automation of intelligent behavior.” (Luger and Stubblefield, 1993)

“Computational intelligence is the study of the design of intelligent agents.” (Poole et al., 1998)

“AI... is concerned with intelligent behavior in artifacts.” (Nilsson, 1998)

- **Thomas Bayes (1702-1761)**
Chess and Go

Want to Play Chess or Just Chat?

**Chess**
- A finite, mathematically well-defined search space
- Limited number of moves and states
- All the symbols are completely grounded in the mathematical rules of the game

**Human Language**
- Words by themselves have no meaning
- Only grounded in human cognition
- Words navigate, align and communicate an infinite space of intended meaning
- Computers can **not** ground words to human experiences to derive meaning

Garry
Kasparov
The best player in the world shows no signs of slowing down

Deep Blue
This 1.4 ton 8-year-old sure plays a mean game of chess
Jeopardy Requires a Broad Knowledge Base

- http://j-archive.com/
- Factual knowledge
  - History, science, politics
- Commonsense knowledge
  - E.g., naïve physics and gender
- Vagueness, obfuscation, uncertainty
  - E.g., “KISS”ing music
The Questions: Solution Methods

• Factoid questions

  Category: Head North
  Clue: They’re the two states you could be reentering if you’re crossing Florida’s northern border.
  Answer: Georgia and Alabama

• Decomposition

  Category: “Rap” Sheet
  Clue: This archaic term for a mischievous or annoying child can also mean a rogue or scamp.
  Subclue 1: This archaic term for a mischievous or annoying child.
  Subclue 2: This term can also mean a rogue or scamp.
  Answer: Rapscallion

• Puzzles

  Category: Rhyme Time
  Clue: It’s where Pele stores his ball.
  Subclue 1: Pele ball (soccer)
  Subclue 2: where store (cabinet, drawer, locker, and so on)
  Answer: soccer locker
The Domain

- Example: castling is a *maneuver* in chess
Precision vs. Percentage Attempted

Upper line: perfect confidence estimation
Champion Human Performance

- Dark dots correspond to Ken Jenning’s games
Overarching Principles

- Massive parallelism
- Many experts
  - Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.
- Pervasive confidence estimation
- Integrate shallow and deep knowledge
Outline

• Course Overview
• Introduction to NLP
• Introduction to Watson
  ➢ Watson in NLP View
• Machine Learning for NLP
• Why NLP is Hard
An NLP view of Watson

- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring
Key NLP Components

- Baseline Search
  - Math basics, Information Retrieval
- Question Understanding
  - Lexical Analysis, Part-of-Speech Tagging, Parsing
- Deep Document Understanding
  - Syntactic Parsing
  - Semantic Role Labeling
  - DependencyParsing
  - Name Tagging
  - Coreference Resolution
  - Relation Extraction
  - Temporal Information Extraction
  - Event Extraction
- Answer Ranking, Confidence Estimation
- Knowledge Base Construction, Population and Utilization
What We have as Input

• Question: when and where did Billy Mays die?

• Source Collection: millions of news, web blogs, discussion forums, tweets, Wikipedia, …

• Question Parsing:
  • [Billy Mays, date-of-death]
  • [Billy Mays, place-of-death]
Question Analysis

Question classes

- Looked at a sample of 500 questions and refined over time
- Factoid is the default class
- QC have special candidate generation
- Some QC have different ML model (different training and features)

<table>
<thead>
<tr>
<th>QClass</th>
<th>Description</th>
<th>Example Questions (correct answer in parentheses)</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>A question that contains a definition of the answer</td>
<td>CONSTRUCTION: It can be the slope of a roof, or the gunk used to waterproof it. (pitch)</td>
<td>14.7%</td>
</tr>
<tr>
<td></td>
<td>CONSTRUCTION: The name of this large beam that supports the joist literally means “something that encircles” (a girder)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATEGORY-RELATION</td>
<td>The answer has a semantic relation to the question, where the relation is specified in the category.</td>
<td>FORMER STATE GOVERNORS: Nelson A. Rockefeller. (New York) COUNTRIES BY NEWSPAPER: Haaretz, Yedioth Ahronoth. (Israel)</td>
<td>7.2%</td>
</tr>
<tr>
<td>FITB</td>
<td>Fill-in-the-blank – question asks for completion of a phrase</td>
<td>COMPLETE IT: Attributed to Lincoln: “The ___ is stronger than the buller”. (ballot) SHAKESPEARE IN LOVE: “Not that I loved Caesar less”, says Brutus., “but that I loved this city ‘more’. (Rome)</td>
<td>3.8%</td>
</tr>
<tr>
<td>ABBREVIATION</td>
<td>The answer is an expansion of an abbreviation in the question</td>
<td>MILITARY MATTERS: Abbreviated SAS, this elite British military unit is similar to the USA’s Delta Force. (the Special Air Service)</td>
<td>2.9%</td>
</tr>
<tr>
<td>PUZZLE</td>
<td>A puzzle question - the answer requires derivation, synthesis, inference, etc.</td>
<td>BEFORE &amp; AFTER: 13th Century Venetian traveler who’s a Ralph Lauren short sleeve top with a collar. (Marlo Polo shirt) THE HIGHEST-SCORING SCRABBLE WORD: Zoom, quiz or haven. (quiz)</td>
<td>2.3%</td>
</tr>
<tr>
<td>ETYMOLGY</td>
<td>A question asking for an English word derived from a foreign word having a given meaning</td>
<td>ARE YOU A FOODIE?: From the Spanish for “to bake in pastry”, it’s South America’s equivalent of a calzone. (an empanada)</td>
<td>1.9%</td>
</tr>
<tr>
<td>VERB</td>
<td>Question asks for a verb</td>
<td>THE NOT-SO-DEADLY SINS: To capitalize all text in an email is an abomination that signifies the person is doing this. (shouting)</td>
<td>1.5%</td>
</tr>
<tr>
<td>TRANSLATION</td>
<td>A question asking for translation of a word or phrase from one language to another</td>
<td>FRUITS IN FRENCH: Pomme. (apple)</td>
<td>1.1%</td>
</tr>
<tr>
<td>NUMFRFR</td>
<td>The answer is a number</td>
<td>YOU NEED TO CONVERT: One eighth of a circle equals this many degrees. (45)</td>
<td>1.0%</td>
</tr>
<tr>
<td>BOND</td>
<td>The question asks for what is in common between a set of entities</td>
<td>EDIBLE COMMON BONDS: Mung, snap, string. (bean)</td>
<td>0.7%</td>
</tr>
<tr>
<td>MULTIPLE-CHOICE</td>
<td>The question contains multiple possible answers from which to choose the correct answer</td>
<td>THE SOUTHERNMOST CAPITAL CITY: Helsinki, Moscow, Bucharest (Bucharest) OSCAR, GRAMMY OR BOTH: Mickey Rooney. (Oscar)</td>
<td>0.5%</td>
</tr>
<tr>
<td>DATE</td>
<td>A question asking for a date or year</td>
<td>THE TEENS: World War I ended in November of this year. (1918)</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Table 1: Question Classes

© 2012 IBM Corporation
Part-of-Speech Tagging and Syntactic Parsing

School of Theatre presents Wonderful and Town

Tree:

- S
  - NP
    - NN: School
    - IN: of
  - NP
    - NN: and
    - CC: and
  - NP
    - NN: Dance
  - VP
    - VBZ: presents
    - NP
      - JJ: Wonderful
      - NN: Town
Semantic Role Labeling: Adding Semantics into Trees
Core Arguments

- Arg0 = agent
- Arg1 = direct object / theme / patient
- Arg2 = indirect object / benefactive / instrument / attribute / end state
- Arg3 = start point / benefactive / instrument / attribute
- Arg4 = end point
They hid the letter on the shelf
## Dependency Relations

<table>
<thead>
<tr>
<th>Argument Dependencies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj</td>
<td>nominal subject</td>
</tr>
<tr>
<td>csubj</td>
<td>clausal subject</td>
</tr>
<tr>
<td>dobj</td>
<td>direct object</td>
</tr>
<tr>
<td>iobj</td>
<td>indirect object</td>
</tr>
<tr>
<td>pobj</td>
<td>object of preposition</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modifier Dependencies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>tmod</td>
<td>temporal modifier</td>
</tr>
<tr>
<td>appos</td>
<td>appositional modifier</td>
</tr>
<tr>
<td>det</td>
<td>determinant</td>
</tr>
<tr>
<td>prep</td>
<td>prepositional modifier</td>
</tr>
</tbody>
</table>
• Since its inception in 2001, <name ID="1" type="organization">Red</name> has caused a stir in <name ID="2" type="location">Northeast Ohio</name> by stretching the boundaries of classical by adding multi-media elements to performances and looking beyond the expected canon of composers.

• Under the baton of <name ID="3" type="organization">Red</name> Artistic Director <name ID="4" type="person">Jonathan Sheffer</name>, <name ID="5" type="organization">Red</name> makes its debut appearance at <name ID="6" type="organization">Kent State University</name> on March 7 at 7:30 p.m.
But the little prince could not restrain admiration:

"Oh! How beautiful you are!"

"Am I not?" the flower responded, sweetly. "And I was born at the same moment as the sun . . ."

The little prince could guess easily enough that she was not any too modest--but how moving--and exciting--she was!

"I think it is time for breakfast," she added an instant later. "If you would have the kindness to think of my needs--"

And the little prince, completely abashed, went to look for a sprinkling-can of fresh water. So, he tended the flower.
**Relation Extraction**

**relation**: a semantic relationship between two entities

<table>
<thead>
<tr>
<th>ACE relation type</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent-Artifact</td>
<td>Rubin Military Design, the <strong>makers</strong> of the <strong>Kursk</strong></td>
</tr>
<tr>
<td>Discourse</td>
<td>each of whom</td>
</tr>
<tr>
<td>Employment/ Membership</td>
<td>Mr. Smith, a senior <strong>programmer</strong> at <strong>Microsoft</strong></td>
</tr>
<tr>
<td>Place-Affiliation</td>
<td><strong>Salzburg</strong> Red Cross <strong>officials</strong></td>
</tr>
<tr>
<td>Person-Social</td>
<td><strong>relatives</strong> of the <strong>dead</strong></td>
</tr>
<tr>
<td>Physical</td>
<td>a <strong>town</strong> some 50 miles south of <strong>Salzburg</strong></td>
</tr>
<tr>
<td>Other-Affiliation</td>
<td><strong>Republican senators</strong></td>
</tr>
</tbody>
</table>
In 1975, after being fired from Columbia amid allegations that he used company funds to pay for his son's bar mitzvah, Davis founded Arista.

Is ‘1975’ related to the employee_of relation between Davis and Arista?
If so, does it indicate START, END, HOLDS… ?

Each classification instance represents a temporal expression in the context of the entity and slot value.

We consider the following classes
- START Rob joined Microsoft in 1999.
- END Rob left Microsoft in 1999.
- HOLDS In 1999 Rob was still working for Microsoft.
- RANGE Rob has worked for Microsoft for the last ten years.
- NONE Last Sunday Rob’s friend joined Microsoft.
Event Extraction

- An event is a specific occurrence that implies a change of states
- **event trigger**: the main word which most clearly expresses an event occurrence
- **event arguments**: the mentions that are involved in an event (participants)
- **event mention**: a phrase or sentence within which an event is described, including trigger and arguments
- ACE defined 8 types of events, with 33 subtypes

<table>
<thead>
<tr>
<th>ACE event type/subtype</th>
<th>Event Mention Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life/Die</td>
<td>Kurt Schork <em>died</em> in Sierra Leone yesterday</td>
</tr>
<tr>
<td>Transaction/Transfer</td>
<td>GM <em>sold the company</em> in Nov 1998 to LLC</td>
</tr>
<tr>
<td>Movement/Transport</td>
<td>Homeless people <em>have been moved</em> to schools</td>
</tr>
<tr>
<td>Business/Start-Org</td>
<td>Schweitzer <em>founded</em> a hospital in 1913</td>
</tr>
<tr>
<td>Conflict/Attack</td>
<td>the <em>attack</em> on Gaza killed 13</td>
</tr>
<tr>
<td>Contact/Meet</td>
<td>Arafat’s cabinet <em>met</em> for 4 hours</td>
</tr>
<tr>
<td>Personnel/Start-Position</td>
<td>She later <em>recruited</em> the nursing student</td>
</tr>
<tr>
<td>Justice/Arrest</td>
<td>Faison was wrongly <em>arrested</em> on suspicion of murder</td>
</tr>
</tbody>
</table>
Mays, 50, had died in his sleep at his Tampa home the morning of June 28.
### Ranking Answers/Confidence Estimation

- **Problems:**
  - different information sources may generate claims with varied trustability
  - various systems may generate *erroneous, conflicting, redundant, complementary, ambiguously worded, or interdependent* claims from the same set of documents

<table>
<thead>
<tr>
<th>System</th>
<th>Source</th>
<th>Slot Filler</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Agence France-Presse, News</td>
<td>Los Angeles</td>
<td>The statement was confirmed by publicist Maureen O’Connor, who said <em>Dio</em> died in <em>Los Angeles</em>.</td>
</tr>
<tr>
<td>B</td>
<td>New York Times, News</td>
<td>Los Angeles</td>
<td><em>Ronnie James Dio</em>, a singer with the heavy-metal bands Rainbow, Black Sabbath and Dio, whose semioperatic vocal style and attachment to demonic imagery made him a mainstay of the genre, died on Sunday in <em>Los Angeles</em>.</td>
</tr>
<tr>
<td>C</td>
<td>Discussion Forum</td>
<td>Atlantic City</td>
<td><em>Dio</em> revealed last summer that he was suffering from stomach cancer shortly after wrapping up a tour in <em>Atlantic City</em>.</td>
</tr>
<tr>
<td>D</td>
<td>Associated Press Worldstream, News</td>
<td>Los Angeles</td>
<td><em>LOS ANGELES</em> 2010-05-16 20:31:18 UTC <em>Ronnie James Dio</em>, the metal god who replaced Ozzy Osbourne in Black Sabbath and later piloted the bands Heaven, Hell and Dio, has died, according to his wife and manager.</td>
</tr>
</tbody>
</table>
Ranking Answers

- Hard Constraints
  - *Used to capture the deep syntactic and semantic knowledge that is likely to pertain to the propositional content of claims.*

- Node Constraints
  - e.g., Entity type, subtype and mention type

- Path Constraints
  - Trigger phrases, relations and events, path length
  - Position of a particular node/edge type in the path

- Interdependent Claims
  - Conflicting slot fillers
  - Inter-dependent slot types
Ranking Answers

• Soft Features:
  • Cleanness
  • Informativeness
  • Local Knowledge Graph
  • Voting
pop recording artist, actress and fashion model. Born and raised in Houston, Texas, she enrolled in various performing is first exposed to singing and dancing competitions as a child. Knowles rose to fame in the late 1990s as the lead in group Destiny's Child, one of the world's best-selling girl groups of all time.

Destiny's Child, Knowles released her debut solo album Dangerously in Love (2003), which spawned the number one and "Baby Boy" and became one of the most successful albums of that year, earning her a then record-tying five [3] Following the group's disbandment in 2005, Knowles released BDay in 2006. It debuted at number one on the luded the anthemic "Single Ladies (Put a Ring on It)". The album and its singles earned her six Grammy Awards, for most Grammy Awards won by a female artist in one night.[4][5][6] Knowles is one of the most honored artists by the among female artists, with 16 awards—13 as a solo artist and three as a member of Destiny's Child.[7][8] acting career in 2001, appearing in the musical film Carmen: A Hip Hopera. In 2006, she starred in the lead role in the 1981 Broadway musical Dreamgirls, for which she earned two Golden Globe nominations. Knowles launched her House of Deréon, in 2004, and has endorsed such brands as Pepsi, Tommy Hilfiger, Armani and L'Oréal. In 2010, nes at number two on its list of the 100 Most Powerful and Influential Celebrities in the world;[9][10] she was also listed ul and influential musician in the world.[11] Time also included Knowles on its list of the "100 Most Influential People in the

• Knowledge Base (KB)
  • Attributes (a.k.a., “slots”) derived from Wikipedia infoboxes are used to create the reference KB

• Source Collection
  • A large corpus of unstructured texts
Knowledge Base Linking (Wikification)

Query = “James Parsons”
Jim Parsons, a graduate of the University of Houston School of Theatre and Dance, won the Emmy on Sunday for Lead Actor in a Comedy Series for his work on The Big Bang Theory.
<table>
<thead>
<tr>
<th>Person</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>per:alternate_names</td>
<td>per:title</td>
</tr>
<tr>
<td>org:alternate_names</td>
<td></td>
</tr>
<tr>
<td>per:date_of_birth</td>
<td>per:member_of</td>
</tr>
<tr>
<td>org:political/religious_affiliation</td>
<td></td>
</tr>
<tr>
<td>per:age</td>
<td>per:employee_of</td>
</tr>
<tr>
<td>org:top_members/employees</td>
<td></td>
</tr>
<tr>
<td>per:country_of_birth</td>
<td>per:religion</td>
</tr>
<tr>
<td>org:number_of_employees/members</td>
<td></td>
</tr>
<tr>
<td>per:stateorprovince_of_birth</td>
<td>per:spouse</td>
</tr>
<tr>
<td>org:members</td>
<td></td>
</tr>
<tr>
<td>per:city_of_birth</td>
<td>per:children</td>
</tr>
<tr>
<td>org:member_of</td>
<td></td>
</tr>
<tr>
<td>per:origin</td>
<td>per:parents</td>
</tr>
<tr>
<td>org:subsidiaries</td>
<td></td>
</tr>
<tr>
<td>per:date_of_death</td>
<td>per:siblings</td>
</tr>
<tr>
<td>org:parents</td>
<td></td>
</tr>
<tr>
<td>per:country_of_death</td>
<td>per:other_family</td>
</tr>
<tr>
<td>org:created_by</td>
<td></td>
</tr>
<tr>
<td>per:stateorprovince_of_death</td>
<td>per:charges</td>
</tr>
<tr>
<td>org:founded</td>
<td></td>
</tr>
<tr>
<td>per:city_of_death</td>
<td></td>
</tr>
<tr>
<td>org:dissolved</td>
<td></td>
</tr>
<tr>
<td>per:cause_of_death</td>
<td>org:country_of_headquarters</td>
</tr>
<tr>
<td>per:countries_of_residence</td>
<td>org:stateorprovince_of_headquarters</td>
</tr>
<tr>
<td>per:statesprovinces_of_residence</td>
<td>org:city_of_headquarters</td>
</tr>
<tr>
<td>per:cities_of_residence</td>
<td>org:shareholders</td>
</tr>
<tr>
<td>per:schools_attended</td>
<td>org:website</td>
</tr>
</tbody>
</table>
Text to Speech Synthesis

- ATT:

- IBM

- Cepstral
  - [http://www.cepstral.com/cgi-bin/demos/general](http://www.cepstral.com/cgi-bin/demos/general)

- Rhetorical (= Scansoft)
  - [http://www.rhetorical.com/cgi-bin/demo.cgi](http://www.rhetorical.com/cgi-bin/demo.cgi)

- Festival
Question Answering

Yahoo! Answers

Can't find it with search? Ask

Share knowledge Help others Earn points

What people think of Answers How does it work?

Best of Answers

How do I learn to ride a bike on my own?

32 Asked by Elliot - Cycling
Information Retrieval

① Relevant Document Set
② Sentence Set [Tree representation]
③ Extracted Paths
Outline

- Course Overview
- Introduction to NLP
- Introduction to Watson
- Watson in NLP View
  - Machine Learning for NLP
- Why NLP is Hard
Models and Algorithms

- By models we mean the formalisms that are used to capture the various kinds of linguistic knowledge we need.
- Algorithms are then used to manipulate the knowledge representations needed to tackle the task at hand.
Some Early NLP History

• 1950s:
  • Foundational work: automata, information theory, etc.
  • First speech systems
  • Machine translation (MT) hugely funded by military (imagine that)
    • Toy models: MT using basically word-substitution
  • Optimism!

• 1960s and 1970s: NLP Winter
  • Bar-Hillel (FAHQT) and ALPAC reports kills MT
  • Work shifts to deeper models, syntax
  • … but toy domains / grammars (SHRDLU, LUNAR)

• 1980s/1990s: The Empirical Revolution
  • Expectations get reset
  • Corpus-based methods become central
  • Deep analysis often traded for robust and simple approximations
  • Evaluate everything
Two Generations of NLP

- **Hand-crafted Systems – Knowledge Engineering [1950s–]**
  - Rules written by hand; adjusted by error analysis
  - Require experts who understand both the systems and domain
  - Iterative guess-test-tweak-repeat cycle

- **Automatic, Trainable (Machine Learning) System [1985s–]**
  - The tasks are modeled in a statistical way
  - More robust techniques based on rich annotations
  - Perform better than rules (Parsing 90% vs. 75% accuracy)
Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora

- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what
Resources: Existing Corpora

- Brown corpus, LOB Corpus, British National Corpus
- Bank of English
- Wall Street Journal, Penn Tree Bank, Nombank, Propbank, BNC, ANC, ICE, WBE, Reuters Corpus
- Canadian Hansard: parallel corpus English-French
- York-Helsinki Parsed corpus of Old Poetry
- Tiger corpus – German
- CORII/CODIS - contemporary written Italian
- MULTEX 1984 and The Republic in many languages
- MapTask
Paradigms

• In particular..
  • State-space search
    • To manage the problem of making choices during processing when we lack the information needed to make the right choice
  • Dynamic programming
    • To avoid having to redo work during the course of a state-space search
      • CKY, Earley, Minimum Edit Distance, Viterbi, Baum-Welch
  • Classifiers
    • Machine learning based classifiers that are trained to make decisions based on features extracted from the local context
Information Units of Interest - Examples

- Explicit units:
  - Documents
  - Lexical units: words, terms (surface/base form)

- Implicit (hidden) units:
  - Word senses, name types
  - Document categories
  - Lexical syntactic units: part of speech tags
  - Syntactic relationships between words – parsing
  - Semantic relationships
Data and Representations

• Frequencies of units
• Co-occurrence frequencies
  • Between all relevant types of units (term-doc, term-term, term-category, sense-term, etc.)
• Different representations and modeling
  • Sequences
  • Feature sets/vectors (sparse)
Learning Methods for NLP

- **Supervised**: identify hidden units (concepts) of explicit units
  - Syntactic analysis, word sense disambiguation, name classification, relations, categorization, …
  - Trained from labeled data

- **Unsupervised**: identify relationships and properties of explicit units (terms, docs)
  - Association, topicality, similarity, clustering
  - Without labeled data

- Semi-supervised: Combinations
Avoiding/Reducing Manual Labeling

• Basic supervised setting – examples are annotated manually by labels (sense, text category, part of speech)
• Settings in which labeled data can be obtained without manual annotation:
  • Anaphora, target word selection
    *The system displays the file on the monitor and prints it.*

• Unsupervised/Semi-supervised Learning approaches
  → Sometimes referred as unsupervised learning, though it actually addresses a supervised task of identifying an externally imposed class (“unsupervised” training)
Outline

• Course Overview
• Introduction to NLP
• Introduction to Watson
• Watson in NLP View
• Machine Learning for NLP
➢ Why NLP is Hard
Why NLP is Difficult: Ambiguity

• Favorite Headlines
  • Teacher Strikes Idle Kids
  • Stolen Painting Found by Tree
  • Kids Make Nutritious Snacks
  • Local HS Dropouts Cut in Half
  • Red Tape Holds Up New Bridges
  • Man Struck by Lightning Faces Battery Charge
  • Hospitals Are Sued by 7 Foot Doctors
How can a machine understand these differences?

- Get the cat with the gloves.
Ambiguity

• Computational linguists are obsessed with ambiguity
• Ambiguity is a fundamental problem of computational linguistics
• Resolving ambiguity is a crucial goal
Need Cross-source Linking to Reduce Uncertainty

- Medical Domain: 33% of abbreviations are ambiguous (Liu et al., 2001), major source of errors in medical NLP (Friedman et al., 2001)

<table>
<thead>
<tr>
<th>RA</th>
<th>“rheumatoid arthritis”, “renal artery”, “right atrium”, “right atrial”, “refractory anemia”, “radioactive”, “right arm”, “rheumatic arthritis”, …</th>
</tr>
</thead>
</table>

- Military Domain
  - “GA ADT 1, USDA, USAID, Turkish PRT, and the DAIL staff met to create the Wardak Agricultural Steering Committee.”
  - “DST” = “District Stability Team” or “District Sanitation Technician”? 
Uncertainty: Ambiguity Example

- Abdul Aziz Al-Omari
  - Member-of: Al-Qaeda
  - Attend-school: Saudi Arabia

- Abdulrahman al-Omari
  - Member-of: Al-Qaeda
  - Attend-school: Saudi Arabia

- Mohamed Atta
  - Supervisor: Imam Muhammad Ibn Saud Islamic University
  - Attend-school: JFK Airport

- Imam Muhammad Ibn Saud Islamic University
  - Nationality: Saudi Arabia

- Immigration
  - City-of-residence: Vero Beach, Florida
  - Flight School: Employee-of: Florida Flight School
Even More Uncertainty: “Morphing” in Vision
Steganography, Codes, Ciphers and Jargons

- Steganography (concealed writing)
  - Wooden tablets covered with wax
  - Tatooon it on the scalp of a messenger

- Codes
  - Box Seven Seeks Tiger5 At Red Coral

- Ciphers
  - BEWARE THE IDES OF MARCH

- Jargons
  - The supply drop will take place at 0100 hours tomorrow
To Ramzi bin al-Shibh

“The first semester commences in the World Trade Center and Pentagon and Capitol on September 11. This summer will surely be hot...”

19 hijackers for private education and four planes.

Regards to Bin Laden

Goodbye.

Abu Abdul Rahman
## Morphs in Social Media

<table>
<thead>
<tr>
<th>Morph</th>
<th>Target</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind Man (瞎子)</td>
<td>Chen Guangcheng (陈光诚)</td>
<td>Sensitive</td>
</tr>
<tr>
<td>First Emperor (元祖)</td>
<td>Mao Zedong (毛泽东)</td>
<td>Vivid</td>
</tr>
<tr>
<td>Kimchi Country (泡菜国)</td>
<td>Korea (韩国)</td>
<td>Vivid</td>
</tr>
<tr>
<td>Rice Country (米国)</td>
<td>United States (美国)</td>
<td>Pronunciation</td>
</tr>
<tr>
<td>Kim Third Fat (金三胖)</td>
<td>Kim Jong-un (金正恩)</td>
<td>Negative</td>
</tr>
<tr>
<td>Miracle Brother (奇迹哥)</td>
<td>Wang Yongping (王勇平)</td>
<td>Irony</td>
</tr>
</tbody>
</table>
Outline

• Course Overview
• Introduction to NLP
• Introduction to Watson
• Watson in NLP View
• Machine Learning for NLP
• Why NLP is Hard
  ➢ Some Demos
Question Answering

- OpenEphyra
- http://sourceforge.net/projects/openephyra/

- Demo show through server
- /m3/KBP/software/openephyra-0.1.1/scripts
Demo 1: Terrorism Networks Extraction

(Huang et al., 2012) Demo URL: http://blender2.cs.qc.cuny.edu/BlenderGraph/
Demo Video: http://nlp.cs.qc.cuny.edu/terrorism.m4v
Demo 2: Temporal Event Tracking

(Chen and Ji, 2009)

http://nlp.cs.qc.cuny.edu/demo/personvisual.html
Demo 3: News Gathering from Tweets

http://tt.zubiaga.org/ui/enhanced-2.0/?date=2012/02/25

(Zubiaga, Ji and Knight, 2013)
Homework

• Reading Assignment on Math & Information Retrieval
• By Friday 1/24, Send jih@rpi.edu, cc hongzhaohuang@gmail.com an email:
  • Your email address to be added to mailing list
  • Your brief background
    • Which year in which program
    • NLP, machine learning, math, linguistics etc. related courses that you have taken
  • Research topics
  • Comments/suggestions