CS 6961:
Information Extraction from Text

- Instructor: Prof. Ellen Riloff
- Time: Mondays and Wednesdays, 1:25-2:45pm
- Location: MEB 3105
- Credit Hours: 3
- Email: teach-cs6961@list.eng.utah.edu

ACTION ITEM: Subscribe to cs6961 mailing list!
https://sympa.eng.utah.edu/sympa

Grading

- Written assignments: 20%
- Programming assignments: 20%
- Midterm exam: 20%
- Project: 40%

The project will be individual, based on your own interests. Everyone will submit a project proposal, which I will discuss with you to ensure appropriate scope.

I want you to pick a topic that you will enjoy!

Project

- You are encouraged to choose a topic based on your personal or research interests.
- First, you will submit a proposal describing the IE task and approach you plan to explore, the data you will use, and your evaluation plans.
- I will then negotiate the details with you to ensure that your plan has appropriate scope and a good chance for success.
- Everyone will present their results to the entire class at the end of the semester and write-up your results in a report (like a research publication).

Course Objectives

- This course will study techniques for extracting different types of information from unstructured natural language text.
- We will explore a variety of information extraction tasks, methods, and applications. We will also study subproblems that are frequently used for IE.
- The project will give you hands-on experience designing, building, and evaluating an IE system.

(Some) IE Topics

- Named Entity Recognition
- Semantic Lexicon & Taxonomy Induction
- IE from the Web
- Semantic Role Labeling
- Relation Extraction
- Event Extraction
- Narrative Event Chains
- Opinion Extraction

IE Applications

- Biological Processes (Genomics)
- Clinical Medicine
- Question Answering / Web Search
- Query Expansion / Semantic Sets
- Extracting Entity Profiles
- Tracking Events (Violent, Diseases, Business, etc)
- Tracking Opinions (Political, Product Reputation, Financial Prediction, On-line Reviews, etc.)

What is Information Extraction?

- Information extraction (IE) is an umbrella term for NLP tasks that involve extracting pieces of information from natural language text.
- IE applications aim to turn unstructured information into a structured representation.
- IE problems typically involve:
  - identifying text snippets to extract
  - assigning semantic meaning to entities or concepts
  - finding relations between entities or concepts

General Techniques

- Syntactic Analysis
  - Phrase Identification
  - Feature Extraction
- Semantic Analysis
- Statistical Measures
- Machine Learning
  - Supervised & Weakly Supervised
- Graph Algorithms

Named Entity Recognition

NER typically involves extracting and labeling certain types of entities, such as proper names and dates.

The Wall Street Journal reports that Google plans to partner with Audi to develop Android-based software for new cars.

Mars One announced Monday that it has picked 1,058 aspiring spaceflyers to move on to the next round in its search for the first humans to live and die on the Red Planet.

Domain-specific NER

Clinical medical systems must recognize problems and treatments:

Adrenal-sparing surgery is safe and effective, and may become the treatment of choice in patients with hereditary phaeochromocytoma.

Biomedical systems must recognize genes and proteins:

IL–2 gene expression and NFkappa B activation through CD28 requires reactive oxygen production by 5lipoxygenase.

Semantic Class Identification

The Wall Street Journal reports that Google plans to partner with Audi to develop Android-based software for new cars.

Mars One announced Monday that it has picked 1,058 aspiring spaceflyers to move on to the next round in its search for the first humans to live and die on the Red Planet.

Semantic Lexicon Induction

- Although some general semantic dictionaries exist (e.g., WordNet), domain-specific applications often have specialized vocabulary.
- Semantic Lexicon Induction techniques learn lists of words that belong to a semantic class.

  Vehicles: car, jeep, helicopter, bike, tricycle, scooter, …

  Animal: tiger, zebra, wolverine, platypus, echidna, …

  Symptoms: cough, sneeze, pain, pu/pd, elevated bp, …

  Products: camera, laptop, iPad, tablet, GPS device, …

Domain-specific Vocabulary

A 14yo m/n doxy owned by a reputable breeder is being treated for IBD with pred.

doxy → ANIMAL
breeder → HUMAN
IBD → DISEASE
pred → DRUG

Domain-specific meanings: lab, mix, m/n = ANIMAL

Semantic Taxonomy Induction

- Ideally, we want semantic concepts to be organized in a taxonomy, to support generalization but distinguish different subtypes.

Animal
  Mammal
    Feline
      Lion, Panthera Leo
      Tiger, Panthera Tigris, Felis Tigris
      Cougar, Mountain Lion, Puma, Catamount
    Canine
      Wolf, Canis Lupus
      Coyote, Prairie Wolf, Brush Wolf, American Jackal
      Dog, Puppy, Canis Lupus Familiaris, Mongrel

Challenges in Taxonomy Induction

- But there are often many ways to organize a conceptual space!
- Strict hierarchies are rare in real data – graphs/networks are more realistic than tree structures.
- For example, animals could be subcategorized based on:
  - carnivore vs. herbivore
  - water-dwelling vs. land-dwelling
  - wild vs. pets vs. agricultural
  - physical characteristics (e.g., baleen vs. toothed whales)
  - habitat (e.g., arctic vs. desert)

Relation Extraction

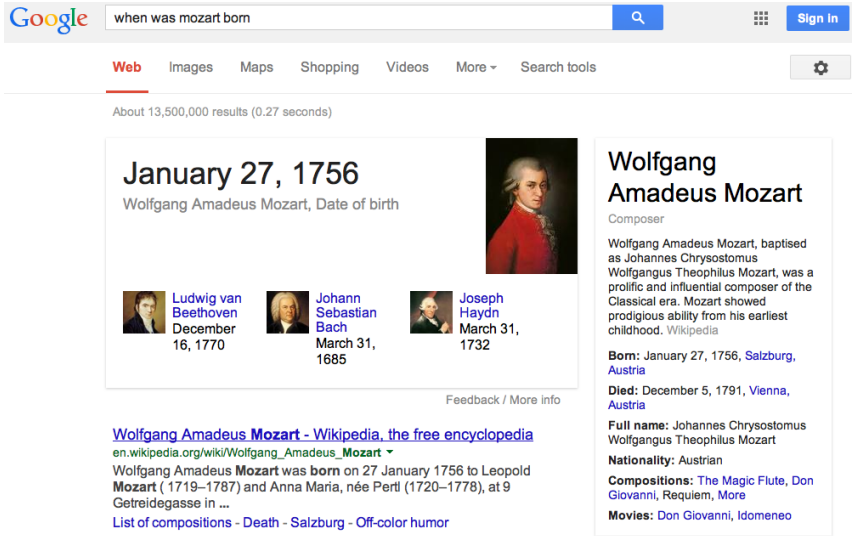
In **Salzburg**, little **Mozart** grew up in a loving middle-class environment.

Birthplace(Mozart, Salzburg)

Steve Ballmer is an American businessman who has been serving as the CEO of **Microsoft** since January 2000

Employed-By(Steve Ballmer, Microsoft)
CEO(Steve Ballmer, Microsoft)

Relations for Web Search



Paraphrasing

- Relations can often be expressed with a multitude of difference expressions.
- Paraphrasing systems try to explicitly learn phrases that represent the same type of relation.
- Examples:
  - X was born in Y
  - Y is the birthplace of X
  - X's birthplace is Y
  - X's hometown is Y
  - X grew up in Y

Thematic Roles

John broke the window with a hammer.
The hammer broke the window.
The window broke.

Agent = John
Theme = window
Instrument = hammer

I ate the spaghetti with tomato sauce with a fork with a friend.

Agent = I
Theme = spaghetti
Co-theme = tomato sauce
Instrument = fork
Co-Agent = friend

Semantic Role Labeling

She blamed the government for failing to do enough to help.

Judge: She
Evaluee: the government
Reason: failing to do enough to help

Julie argued with Bob about politics in French.

Protagonist1: Julie
Protagonist2: Bob
Topic: politics
Medium: French

Event Extraction

Goal: extract facts about events from unstructured documents

Example: extracting information about terrorism events in news articles:

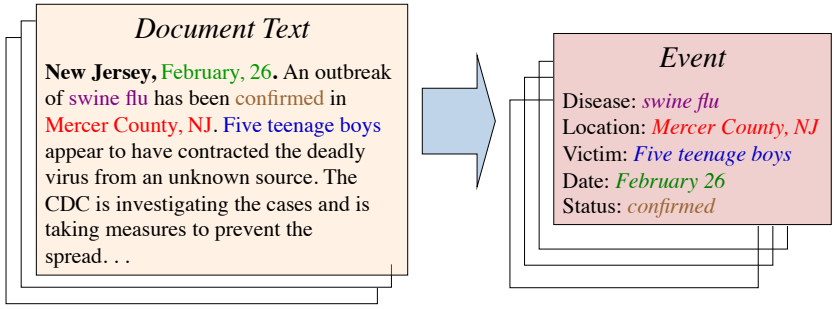
December 29, Pakistan - The U.S. embassy in Islamabad was damaged this morning by a car bomb. Three diplomats were injured in the explosion. Al Qaeda has claimed responsibility for the attack.

EVENT

Type: *bombing*
Target: *U.S. embassy*
Location: *Islamabad, Pakistan*
Date: *December 29*
Weapon: *car bomb*
Victim: *three diplomats*
Perpetrator: *Al Qaeda*

Event Extraction

Another example: extracting information about disease outbreak events.



After a brief lull, the avian flu is on the march again through Fraser Valley poultry farms.

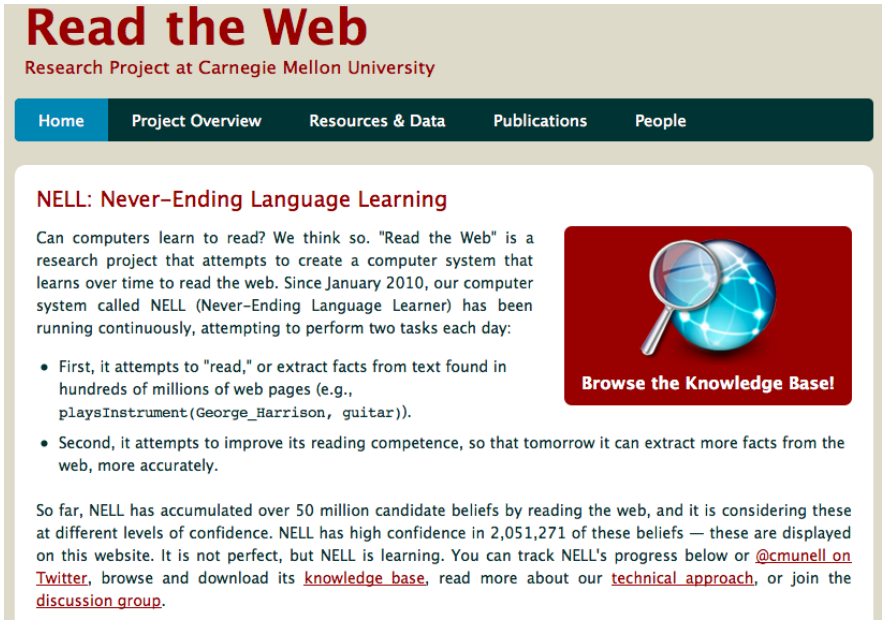
The Canadian Food Inspection Agency says ongoing surveillance efforts have led to the detection of bird flu on 36 premises.

The agency says it is continuing depopulation efforts on infected farms on a priority basis.

OUTBREAK
Disease: avian flu / bird flu
Victims: poultry
Location: Fraser Valley poultry farms / 36 premises
Country: Canada
Status: confirmed
Containment: depopulation

Large-Scale IE from the Web

- Some researchers have been developing IE systems for large-scale extraction of facts and relations from the Web.
- These systems exploit the massive amount of text and redundancy available on the Web and use weakly supervised, iterative learning to harvest information for automated knowledge base construction.
- The KnowItAll project at UW and NELL project at CMU are well-known research groups pursuing this work.



Opinion Extraction

I just bought a Powershot a few days ago. I took some pictures using the camera. Colors are so beautiful even when flash is used. Also easy to grip since the body has a grip handle. [Kobayashi et al., 2007]

Source: <writer>
Target: Powershot
Aspect: pictures, colors
Evaluation: beautiful, easy to grip

Opinion Extraction from News

[Wilson & Wiebe, 2009]

Italian senator Renzo Gubert praised the Chinese Government's efforts.

Source: Italian senator Renzo Gubert
Target: the Chinese Government
Evaluation: $\text{praised}_{\text{POSITIVE}}$

African observers generally approved of his victory while Western governments denounced it.

Source: African observers
Target: his victory
Evaluation: $\text{approved}_{\text{POSITIVE}}$

Source: Western governments
Target: it (his victory)
Evaluation: $\text{denounced}_{\text{NEGATIVE}}$

Summary

- Information extraction systems frequently rely on low-level NLP tools for basic language analysis, often in a pipeline architecture.
- There are a wide variety of applications for IE, including both broad-coverage and domain-specific applications.
- Some IE tasks are relatively well-understood (e.g., named entity recognition), while others are still quite challenging!
- We've only scratched the surface of possible IE tasks … nearly endless possibilities!