Intro to Information Extraction -Sentiment Lexicon Induction as an example

Many slides adapted from Ellen Riloff and Dan Jurafsky

What is Information Extraction?

- Information extraction (IE) is an umbrella term for NLP tasks that involve extracting pieces of information from text and assigning some meaning to the information.
- Many IE applications aim to turn unstructured text 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

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.)

General Techniques

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

Named Entity Recognition (NER)

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 Toyota to develop Android software for their hybrid 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 Toyota to develop Android software for their hybrid 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 to distinguish different subtypes.

Animal
    Mammal
        Feline
            Lion, Panthera Leo
            Tiger, Panthera Tigris, Felis Tigris
            Cougar, Mountain Lion, Puma, Panther, 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

Google | when was mozart born | Sign in

Web | Images | Maps | Shopping | Videos | More | Search tools

About 13,500,000 results (0.27 seconds)

January 27, 1756

Wolfgang Amadeus Mozart, Date of birth

Ludwig van Beethoven
December 16, 1770

Johann Sebastian Bach
March 31, 1685

Joseph Haydn
March 31, 1732

Feedback / More info

Wolfgang Amadeus **Mozart** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Wolfgang_Amadeus_**Mozart**
Wolfgang Amadeus **Mozart** was **born** on 27 January 1756 to Leopold **Mozart** (1719–1787) and Anna Maria, née Pertl (1720–1778), at 9 Getreidegasse in ...
List of compositions - Death - Salzburg - Off-color humor

Wolfgang Amadeus Mozart

Composer

Wolfgang Amadeus Mozart, baptised as Johannes Chrysostomus Wolfgangus Theophilus Mozart, was a prolific and influential composer of the Classical era. Mozart showed prodigious ability from his earliest childhood. Wikipedia

Born: January 27, 1756, Salzburg, Austria

Died: December 5, 1791, Vienna, Austria

Full name: Johannes Chrysostomus Wolfgangus Theophilus Mozart

Nationality: Austrian

Compositions: The Magic Flute, Don Giovanni, Requiem, More

Movies: Don Giovanni, Idomeneo

Paraphrasing

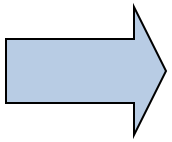
- Relations can often be expressed with a multitude of difference expressions.
- Paraphrasing systems try to explicitly learn phrases that represent the same type of relation.
- Examples:
  - X was born in Y
  - Y is the birthplace of X
  - X's birthplace is Y
  - X's hometown is Y
  - X grew up in Y

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.

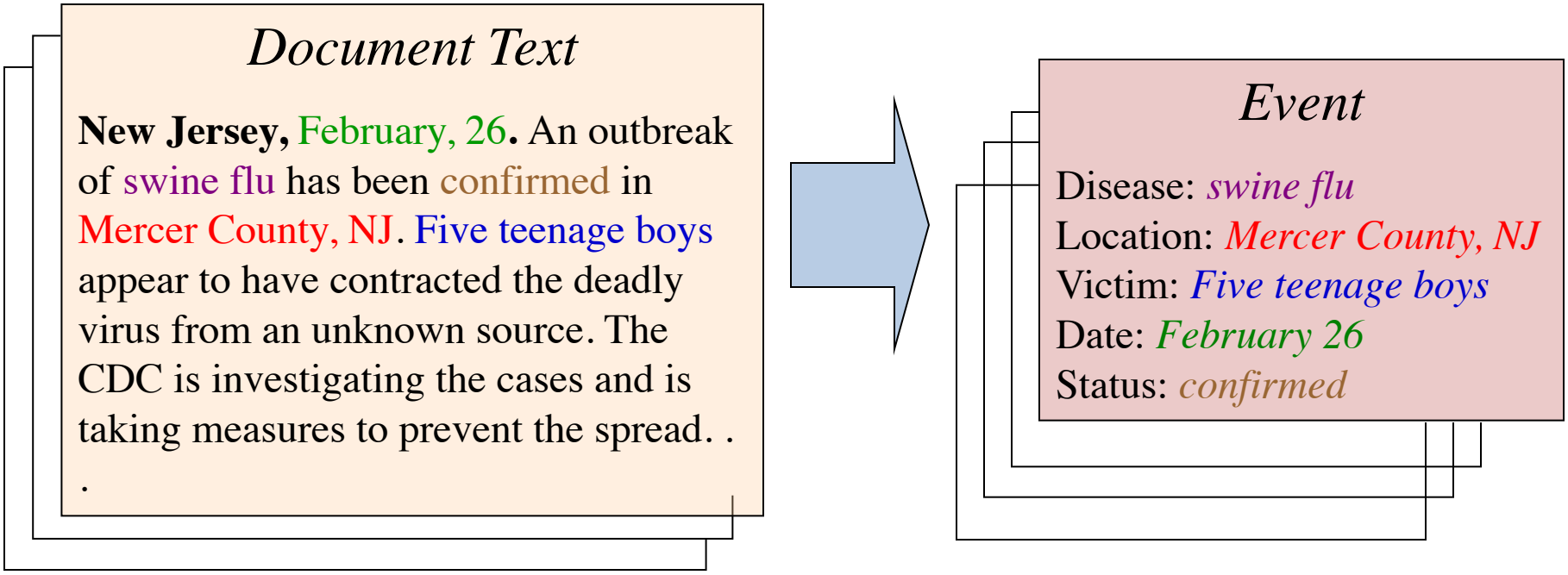


EVENT

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.



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.

Read the Web

Research Project at Carnegie Mellon University

Home | Project Overview | Resources & Data | Publications | People

NELL: Never-Ending Language Learning

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g., `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 2,051,271 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or @cmunell on Twitter, browse and download its knowledge base, read more about our technical approach, or join the discussion group.

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.

Turney Algorithm to learn a Sentiment Lexicon

Turney (2002): Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

1. Extract a *phrasal lexicon* from reviews
2. Learn polarity of each phrase
3. Rate a review by the average polarity of its phrases

Extract two-word phrases with adjectives

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN nor NNS
NN or NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

How to measure polarity of a phrase?

- Positive phrases co-occur more with *“excellent”* ->*”great”*
- Negative phrases co-occur more with *“poor”*
- But how to measure co-occurrence?

Pointwise Mutual Information

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Pointwise Mutual Information

- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **PMI between two words**:
  - How much more do two words co-occur than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

How to Estimate Pointwise Mutual Information

- Query search engine (Altavista)
  - P(word) estimated by `hits(word)/N`
  - P($word_1$,$word_2$) by `hits(word1 NEAR word2)/N`
    - (More correctly the bigram denominator should be kN, because there are a total of N consecutive bigrams (word1,word2), but kN bigrams that are k words apart, but we just use N on the rest of this slide and the next.)

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{\frac{1}{N} hits(word_1 \text{ NEAR } word_2)}{\frac{1}{N} hits(word_1) \frac{1}{N} hits(word_2)}$$

Does phrase appear more with “poor” or “excellent”?

$$\text{Polarity}(phrase) = \text{PMI}(phrase, \text{"excellent"}) - \text{PMI}(phrase, \text{"poor"})$$

$$= \log_2 \frac{\frac{1}{N}\text{hits}(phrase \text{ NEAR "excellent"})}{\frac{1}{N}\text{hits}(phrase)\frac{1}{N}\text{hits}(\text{"excellent"})} - \log_2 \frac{\frac{1}{N}\text{hits}(phrase \text{ NEAR "poor"})}{\frac{1}{N}\text{hits}(phrase)\frac{1}{N}\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(phrase \text{ NEAR "excellent"})}{\text{hits}(phrase)\text{hits}(\text{"excellent"})} \frac{\text{hits}(phrase)\text{hits}(\text{"poor"})}{\text{hits}(phrase \text{ NEAR "poor"})}$$

$$= \log_2 \left(\frac{\text{hits}(phrase \text{ NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(phrase \text{ NEAR "poor"})\text{hits}(\text{"excellent"})} \right)$$

Phrases from a thumbs-up review

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
…		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
Average		0.32

Phrases from a thumbs-down review

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
…		
virtual monopoly	JJ NN	-2.0
lesser evil	RBR JJ	-2.3
other problems	JJ NNS	-2.8
low funds	JJ NNS	-6.8
unethical practices	JJ NNS	-8.5
Average		-1.2

Results of Turney algorithm

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- Turney algorithm: 74%

- Phrases rather than words
- Learns domain-specific information