600.465
Natural Language Processing

Prof: Jason Eisner

Webpage: http://cs.jhu.edu/~jason/465

syllabus,
announcements,
slides,
homeworks

Goals of the field

Computers would be a lot more useful if they could handle our email, do our library research, talk to us …

But they are fazed by natural human language.

How can we tell computers about language? (Or help them learn it as kids do?)

A few applications of NLP

- Spelling correction, grammar checking …
- Better search engines
- Information extraction
- Psychotherapy; Harlequin romances; etc.

- New interfaces:
  - Speech recognition (and text-to-speech)
  - Dialogue systems (USS Enterprise onboard computer)
  - Machine translation (the Babel fish)

Goals of the course

- Introduce you to NLP problems & solutions
- Relation to linguistics & statistics

- At the end you should:
  - Agree that language is subtle & interesting
  - Feel some ownership over the formal & statistical models
  - Understand research papers in the field

Ambiguity: Favorite Headlines

- Iraqi Head Seeks Arms
- Is There a Ring of Debris Around Uranus?
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group

Ambiguity: Favorite Headlines

- British Left Waffles on Falkland Islands
- Never Withhold Herpes Infection from Loved One
- Red Tape Holds Up New Bridges
- Man Struck by Lightning Faces Battery Charge
- Clinton Wins on Budget, but More Lies Ahead
- Hospitals Are Sued by 7 Foot Doctors

Levels of Language

- Phonetics/phonology/morphology: what words (or subwords) are we dealing with?
- Syntax: What phrases are we dealing with? Which words modify one another?
- Semantics: What’ s the literal meaning?
- Pragmatics: What should you conclude from the fact that I said something? How should you react?

Subtler Ambiguity

- Q: Why does my high school give me a suspension for skipping class?
- A: Administrative error. They're supposed to give you a suspension for auto shop, and a jump rope for skipping class.   (*rim shot*)

What’s hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

To get a donut (spare tire) for his car?

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

store where donuts shop? or is run by donuts?
or looks like a big donut? or made of donut?
or has an emptiness at its core?

What's hard about this story?

I stopped smoking freshman year, but

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

What’s hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Describes where the store is? Or when he stopped?

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Well, actually, he stopped there from hunger and exhaustion, not just from work.

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

At that moment, or habitually?

(*Similarly:* Mozart composed music.)

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

That's how often he thought it?

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

But actually, a coffee only stays good for about 10 minutes before it gets cold.

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

Similarly: In America a woman has a baby every 15 minutes. Our job is to find that woman and stop her.

What’s hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

the particular coffee that was good every few hours? the donut store? the situation?

What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

too expensive for what? what are we supposed to conclude about what John did?

how do we connect "it" to "expensive"?

n-grams

- Letter or word frequencies: 1-grams (= unigrams)
  - useful in solving cryptograms: ETAOINSHRDLU…
- If you know the previous letter: 2-grams (= bigrams)
  - "h" is rare in English (4%; 4 points in Scrabble)
  - but "h" is common after "t" (20%)
- If you know the previous two letters: 3-grams (= trigrams)
  - "h" is really common after "(space) t"

etc. …

Some random n-gram text …