Advanced Language Technologies CS6740/INFO6300

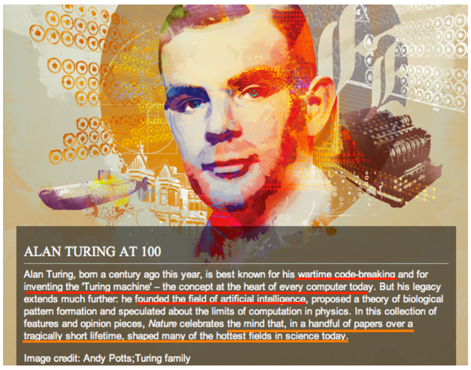
Professor Claire Cardie & Professor Lillian Lee

"I'm sorry, Dave, I'm afraid I can't do that":

Can computers really understand what we say?

the **dream** of language technologies

Why is this man smiling?



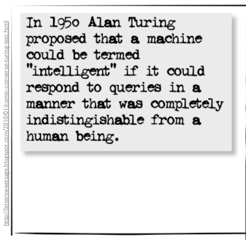
ALAN TURING AT 100

Alan Turing, born a century ago this year, is best known for his wartime code-breaking and for inventing the 'Turing machine' – the concept at the heart of every computer today. But his legacy extends much further: he founded the field of artificial intelligence, proposed a theory of biological pattern formation and speculated about the limits of computation in physics. In this collection of features and opinion pieces, *Nature* celebrates the mind that, in a handful of papers over a tragically short lifetime, shaped many of the hottest fields in science today.

Image credit: Andy Potts;Turing family

http://www.nature.com/nature/journal/v482/n7386/full/482440a.html

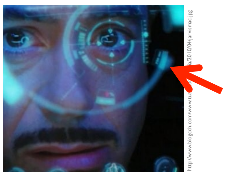
The Turing test: Intelligence ➔ human-level language use



In 1950 Alan Turing proposed that a machine could be termed "intelligent" if it could respond to queries in a manner that was completely indistinguishable from a human being.

Turing predicted we'd be close in about 50 years.

Do authors dream of electric speech?



"Jarvis", the A.I. system in *Iron Man*

Why is this man not smiling?

Open the pod bay doors, Hal.

I'm sorry, Dave, I'm afraid I can't do that.

from **sci-fi** to **science and engineering**

Natural-language processing (NLP)

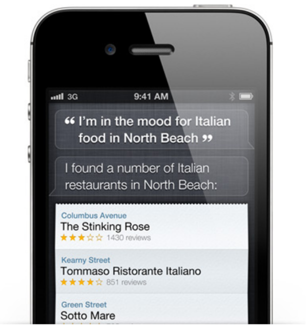
Goal: create systems that use human language as input/output

- speech-based interfaces
- information retrieval / question answering
- automatic summarization of news, emails, postings, etc.
- automatic translation

… and much more!

Interdisciplinary: computer science; linguistics, psychology, communication; probability & statistics, information theory…

Recently deployed (in beta): Siri



3G 9:41 AM

“ I'm in the mood for Italian food in North Beach ”

I found a number of Italian restaurants in North Beach:

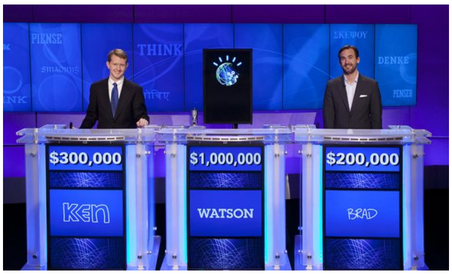
Columbus Avenue
The Stinking Rose
★★★☆☆ 1430 reviews

Kearny Street
Tommaso Ristorante Italiano
★★★★☆ 851 reviews

Green Street
Sotto Mare

http://www.apple.com/iphone/features/siri.html

State of the art: Watson



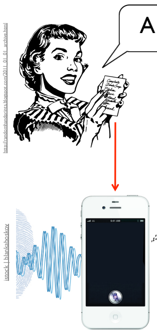
Credit: AP Photo/Jeopardy Productions Inc.

The Watson system beat human Jeopardy! champions (and didn't have internet access; it learned by "reading" before the match)

Why is this woman smiling?

But we're **not all the way there yet**

Real-life error (1)

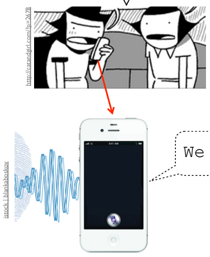


A bunch of grapes.

```
Hey bunch of grapes
```

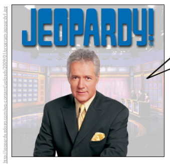
Real-life error (2)

We can email you when we're back.



```
We can email you when you're fat.
```

Real-life error (3)



[This U.S. city's] largest airport …

What is Toronto???

why is **understanding language** so **hard**?

Challenge: ambiguity

List all flights on Tuesday

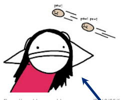
List all flights on Tuesday = *List all the flights leaving on Tuesday.*

List all flights on Tuesday = *Wait 'til Tuesday, then list all flights.*

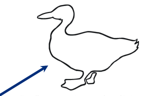
More realistic example

Retrieve all the local patient files

Baroque example



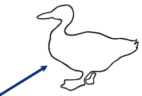
[http://casablancapa.blogspot.com/2010/05/fore.html]



[http://www.supercoloring.com/pages/duck-outline/]

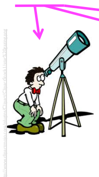
I saw her duck with a telescope.

Baroque example



[http://www.supercoloring.com/pages/duck-outline/]

I saw her duck with a telescope.



www.geocities.ws/looneyebay/dell/bb040.jpg

http://pokerfoldingtable.com/wp-content/uploads/2009/02/three-men-gambling-sitting-at-poker-table-playing-cards-betting-party-pen-ink-drawing-300x234.png

Conversation complications

: Do you know when the train to Boston leaves?

: Yes.

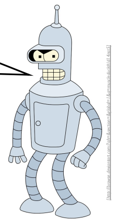
: I want to know when the train to Boston leaves.

: I understand.

[Grishman 1986]

Images: http://3.bp.blogspot.com/_o4kq5TNL0Z4/TUx0j6E5BLI/AAAAAAAAA5k/J7xjhvrcNIU/s1600/Trillian-hitchhikers-guide-to-the-galaxy-the-2005.jpg, http://www.tvacres.com/images/robots_androids_marvin_movie.jpg

I'm sorry, Dave, I'm afraid I can't do that.



I'm afraid you might be right.

Meeting these challenges: a brief history

1940s – 50s: From language to probability

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point …

[The] semantic aspects of communication are irrelevant to the engineering problem.

The significant aspect is that the actual message is one selected from a set of possible messages."

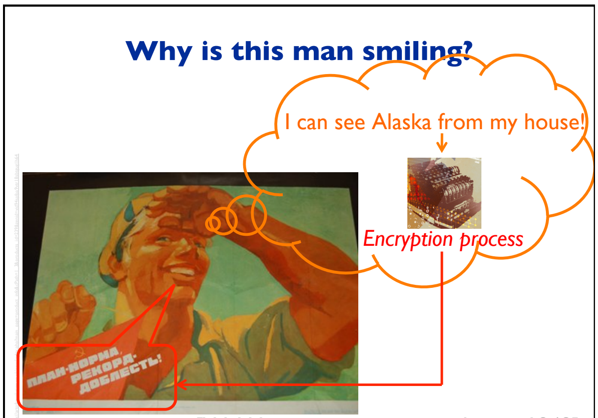
--C. Shannon, 1948 Bell Laboratories

Language, statistics, cryptography

WWII: Turing helps break the German "Enigma" code

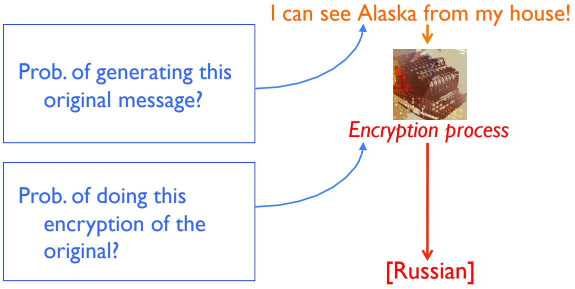
(An original Enigma machine for encrypting messages is on display now in the Kroch Library in Olin.)

Why is this man smiling?



[W. Weaver memo on translation, 1949]

Two probabilities to infer



Prob. of generating this original message?

Prob. of doing this encryption of the original?

I can see Alaska from my house!

Encryption process

[Russian]

Another use of message probs: speech recognition

(1) It's hard to recognize speech

(2) It's hard to wreck a nice beach

Both messages have almost the same acoustics, but different likelihoods.

1950s-1980s: Breaking with statistics

N. Chomsky (1957):

(a) Colorless green ideas sleep furiously

(b) Furiously sleep ideas green colorless

The argument: Neither sentence has ever occurred in the history of English. So any statistical model would given them the same probability (zero).

The field moved to sophisticated non-probabilistic models of language.

1990s: The empiricists strike back

- Huge amounts of data start coming online

Google!

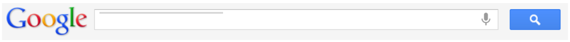
- Advances in algorithms and computational power

"Every time I fire a linguist, my [system's] performance goes up" -- F. Jelinek (apocryphal)

2000s and beyond: integrating language insights and statistical techniques

Is Snooki on stork watch?

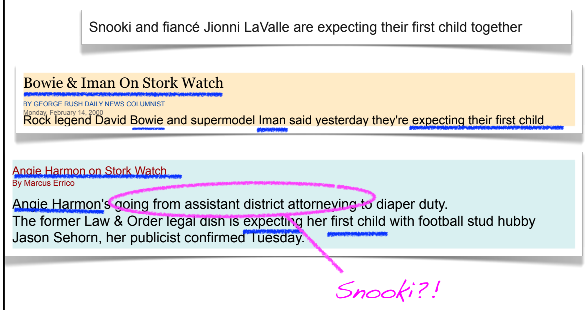
(wondered in March 2012)



[All 8 results were from March 2011 or earlier]

Integrating lang and stats (cont)

Is Snooki on stork watch?



Snooki and fiancé Jionni LaValle are expecting their first child together

Bowie & Iman On Stork Watch

BY GEORGE RUSH DAILY NEWS COLUMNIST

Monday, February 14, 2000

Rock legend David Bowie and supermodel Iman said yesterday they're expecting their first child

Angie Harmon on Stork Watch

By Marcus Errico

Angie Harmon's going from assistant district attorneying to diaper duty.
The former Law & Order legal dish is expecting her first child with football stud hubby Jason Sehorn, her publicist confirmed Tuesday.

Snooki?!

the **game-changers:**

- **data-driven approaches**
- **models of language**

Why is this man smiling?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision.... I do not know what the right answer is, but I think [different] approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

What topics might we cover?

Information retrieval

Text categorization

Information extraction

Summarization

Question-answering systems

NL generation

Machine translation

Dialog systems

Part-of-speech tagging

Word sense disambiguation

Language models

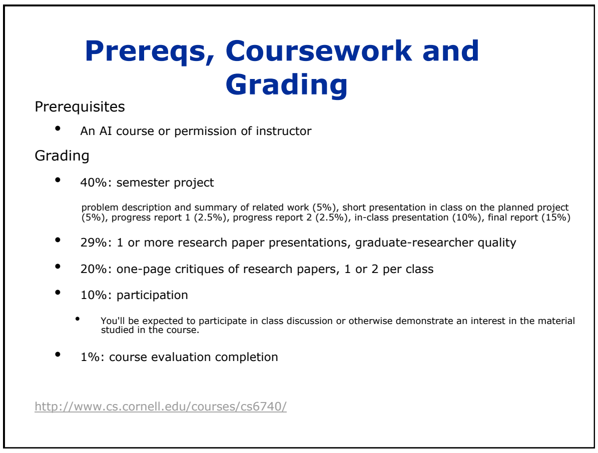
Topic models

Parsing

Semantic analysis

Discourse processing

Coreference analysis



Prereqs, Coursework and Grading

Prerequisites

- An AI course or permission of instructor

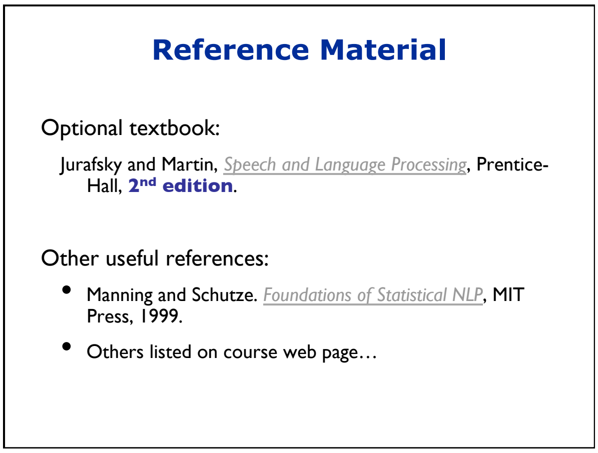
Grading

- 40%: semester project

  problem description and summary of related work (5%), short presentation in class on the planned project (5%), progress report 1 (2.5%), progress report 2 (2.5%), in-class presentation (10%), final report (15%)

- 29%: 1 or more research paper presentations, graduate-researcher quality
- 20%: one-page critiques of research papers, 1 or 2 per class
- 10%: participation
  - You'll be expected to participate in class discussion or otherwise demonstrate an interest in the material studied in the course.
- 1%: course evaluation completion

http://www.cs.cornell.edu/courses/cs6740/



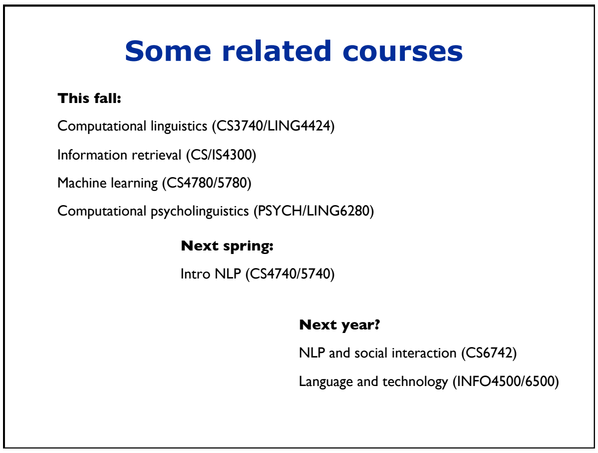
Reference Material

Optional textbook:

Jurafsky and Martin, *Speech and Language Processing*, Prentice-Hall, **2nd edition**.

Other useful references:

- Manning and Schutze. *Foundations of Statistical NLP*, MIT Press, 1999.
- Others listed on course web page…



Some related courses

This fall:

Computational linguistics (CS3740/LING4424)

Information retrieval (CS/IS4300)

Machine learning (CS4780/5780)

Computational psycholinguistics (PSYCH/LING6280)

Next spring:

Intro NLP (CS4740/5740)

Next year?

NLP and social interaction (CS6742)

Language and technology (INFO4500/6500)