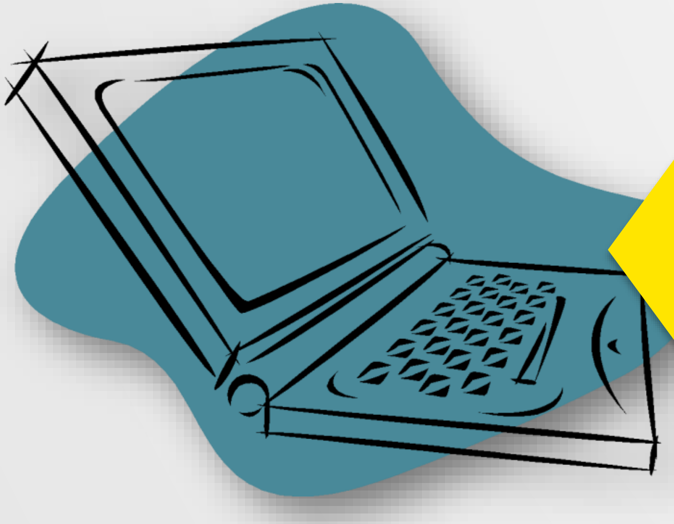
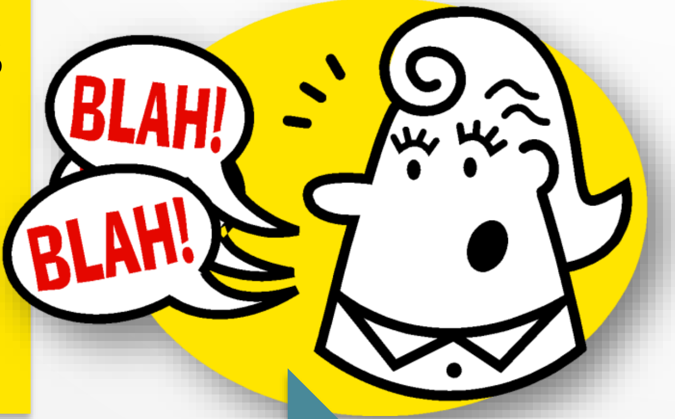
401
ONTARIO
NATURAL
LANGUAGE
COMPUTING

What is natural language computing?



Getting computers to understand everything we say and write.



In this class (and in the field generally), we are interested in learning the ***statistics of language***.

Increasingly, computers give insight into how humans process language, or generate language themselves.

What is Natural Language Computing?

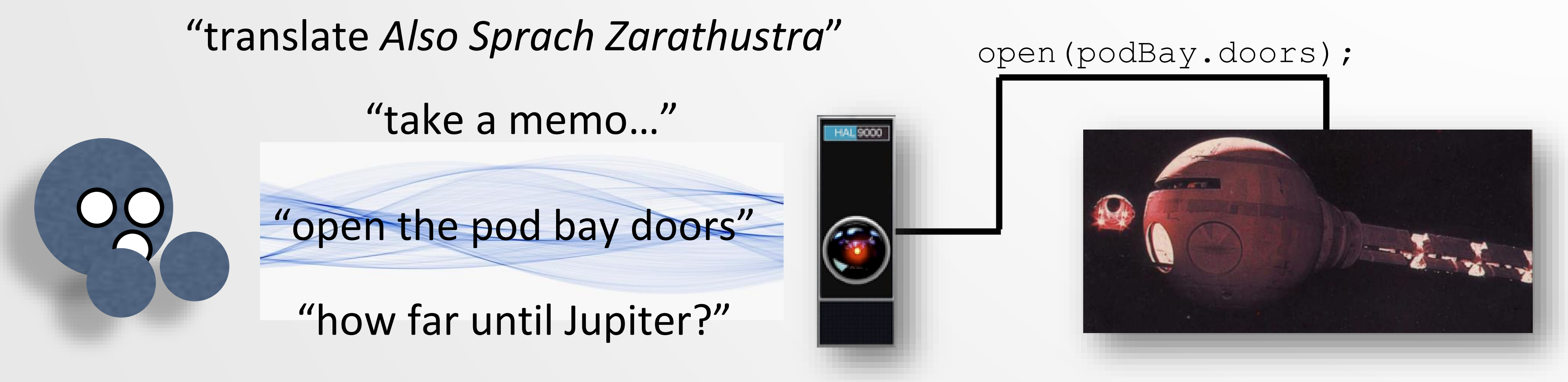
- The computer science (and statistics) behind **natural language processing (NLP)**, also known as **computational linguistics (CL)**.

- Applications
  - Text Classification
  - Automatic translation between languages
  - Automatic speech transcription
  - Spoken language understanding
  - Information Retrieval
  - Text/speech Summarization

Examples

What can natural language do?

A key component of **human-computer interaction**.



“translate *Also Sprach Zarathustra*”

“take a memo…”

“open the pod bay doors”

“how far until Jupiter?”

```
open(podBay.doors);
```

“Can you summarize *2001: A Space Odyssey?*”

We’ve made progress, but why are these things *still* hard to do?

A little deeper

- Language has ***hidden structures***, e.g.,
  - How are **sounds** and **text** related?
    - e.g., why is this: not a '*ghoti*' (*enou**gh***, *w**o**men*, *na**ti**on*)?
  - How are words **combined** to make sentences?
    - e.g., what makes '*colourless green ideas sleep furiously*' **correct** in a way **unlike** '*furiously sleep ideas green colourless*'?
  - How are words and phrases used to produce **meaning**?
    - e.g., if someone asks '*do you know what time it is?*', why is it **inappropriate** to answer '*yes*'?
- We need to organize the way we think about language…

Categories of linguistic knowledge

- **Phonology**: the study of patterns of speech sounds.
  e.g., “read” → */r iy d/*
- **Morphology**: how words can be changed by inflection or derivation.
  e.g., “read”, “reads”, “reader”, “reading”, …
- **Syntax**: the ordering and structure between words and phrases (i.e., grammar).
  e.g., *NounPhrase* → *article adjective noun*
- **Semantics**: the study of how meaning is created by words and phrases.
  e.g., “book” →
- **Pragmatics**: the study of meaning in contexts.
  e.g., explanation span, refutation span

Ambiguity – Phonological

- **Phonology**: the study of patterns of speech sounds.

Problem for *speech synthesis*	"read"	→ */r iy d/*	as in *'I like to **read**'*
	"read"	→ */r eh d/*	as in *'She **read** a book'*
	"object"	→ */aa^1 b jh eh^0 k t /*	as in *'That is an **object**'*
	"object"	→ */ah^0 b jh eh^1 k t /*	as in *'I **object**!'*
Problem for *speech recognition*	"too"	← /t uw/	as in *'**too** much'*
	"two"	← /t uw/	as in *'**two** beers'*

- Ambiguities can often be **resolved** in context, but not always.
  - e.g., */h aw t uw r eh^1 k ah ?? n ay^2 z s (b|p) iy ch/*

    → *'how to recognize speech'*

    → *'how to wreck a nice beach'*

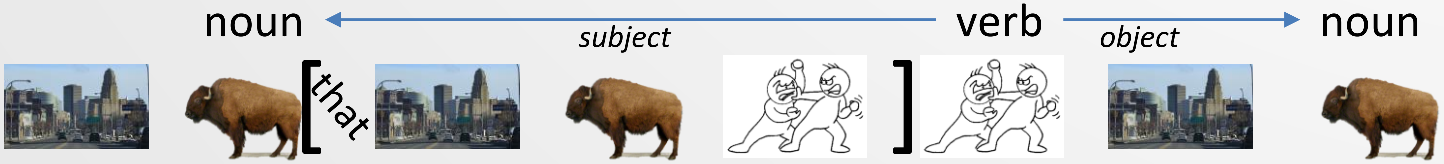
Resolution with syntax

- If you hear the sequence of speech sounds

  /b ah f ae l ow b ah f ae l ow b ah f ae l ow b ah f ae l ow …
  b ah f ae l ow b ah f ae l ow b ah f ae l ow b ah f ae l ow/

  which word sequence is being spoken?

  → “Buff a low buff a lobe a fellow Buff a low buff a lobe a fellow…”
  → “Buffalo buff aloe buff aloe buff aloe buff aloe buff aloe …”
  → “Buff aloe buff all owe Buffalo buffalo buff a lobe …”
  → “Buff aloe buff all owe Buffalo buff aloe buff a lobe …”
  → **“Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo”**



noun ← *subject* — verb — *object* → noun

[that …]

- It’s obvious (to us) that the last option is most likely because we have knowledge of **syntax**, i.e., grammar.

NLP as artificial intelligence

NLP involves **resolving ambiguity** at all levels.

- Reasoning with **linguistic** knowledge
- Reasoning with **world** knowledge
- We sometimes represent the former with *grammars*

We tend to use numerical parameters (probabilities?) to distinguish competing hypotheses.

- E.g., is Google a noun or a verb?
- Examples where Google is a noun (“Google makes Android”) does not mean Google is never a verb (“Google his name”).
- P(noun|Google) > P(verb|Google) > 0.

Aside – Chatbots

- ELIZA (Weizenbaum, 1966): simple pattern matching to imitate a psychiatrist.
- Surprisingly effective despite **unsophisticated methods**.
- e.g.,

  User: Men are all alike.
  ELIZA: In what way?
  User: They're always bugging us about something or other.
  ELIZA: Can you think of a specific example?
  User: My boyfriend made me come here.
  ELIZA: Your boyfriend made you come here.

(Jurafsky and Martin, 2009)

Course outline (approximate)

- Introduction, lexical distributions, language modelling, lexical embeddings (3 lectures)
- Features and classification (2 lecture) *
- Entropy and information theory (2 lectures) *
- Neural language models (2 lectures) *
- Machine translation (3 lectures) **
- Large language models (3 lectures) *
- Acoustics and signal processing (3 lectures) *
- Automatic speech recognition (2 lectures) **
- Speech Synthesis (1 lecture) **
- Information retrieval (1 lecture) **
- Summarization (1 lecture) **
- Ethics for NLP (2 lectures)

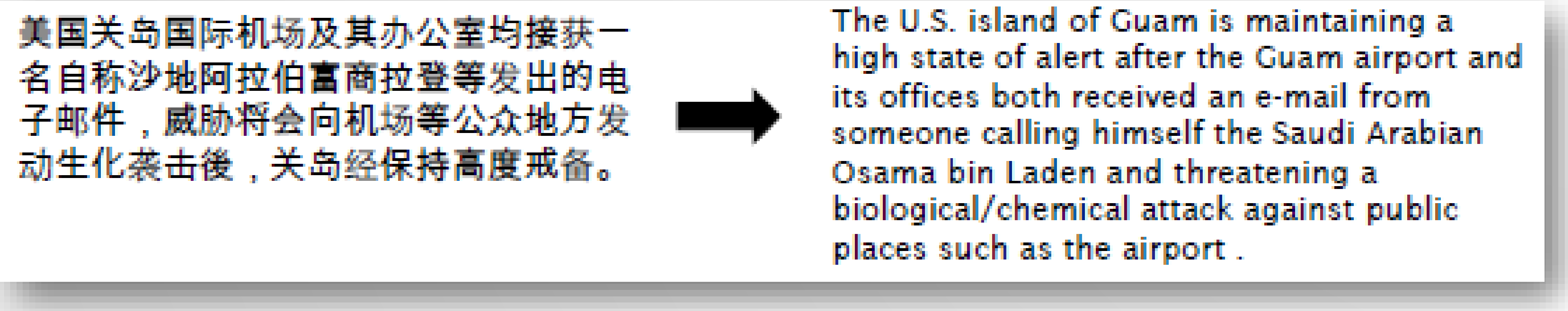
* techniques ** applications

What we will not cover

- Interpretability of language models…*
- Advanced lexical semantics*
- Question answering (including ChatGPT 😭)*
- Information extraction*
- Parsing/generation of natural language*%
- Advanced speech recognition and synthesis¶
- Cognitively based methods§^
- Semantic inference,% semantic change/drift^
- Understanding dialogues and conversations¶
- Advanced ethics for NLP$

* CSC 485 / 2501. % CSC 2517. ¶ CSC 2518. § CSC 2540. ^ CSC 2611. $csc 2528.

Preview: Machine translation

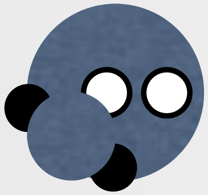


美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

- For years, the holy grail of NLP.
- Requires both **interpretation** and **generation**.
- Over $60B spent annually on human translation in 2022 – projected to reach $96B by 2032
- Machine translation: $1.1B. $3B by 2027.
- 1 in every 4M words of content is translated into at least one other language.

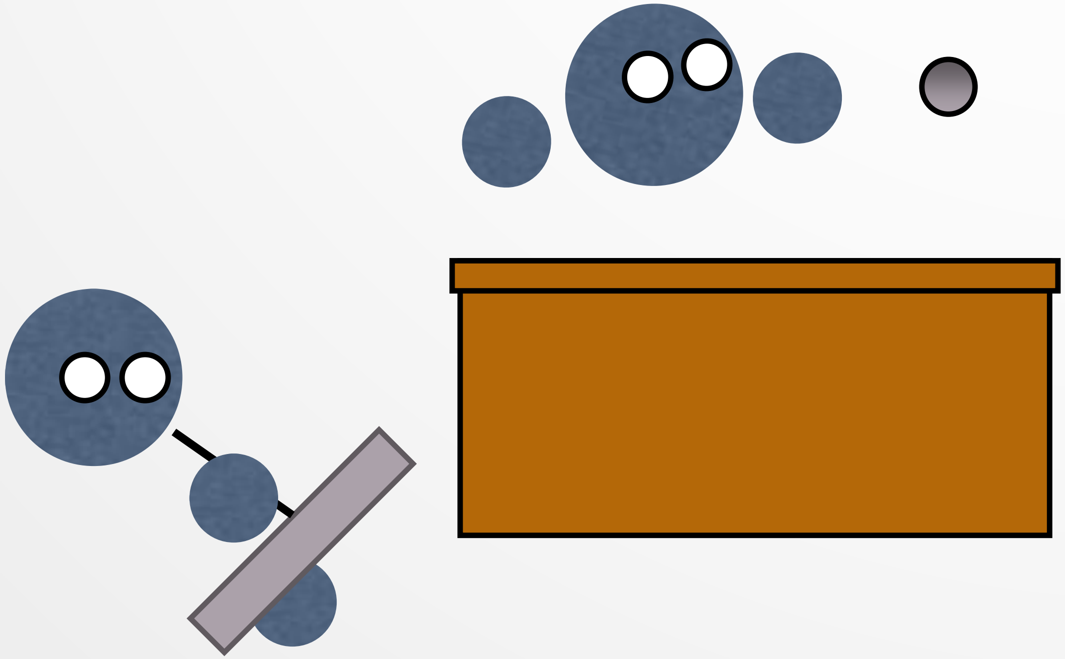
Preview: Speech recognition



Buy ticket...
AC490...
yes

Telephony

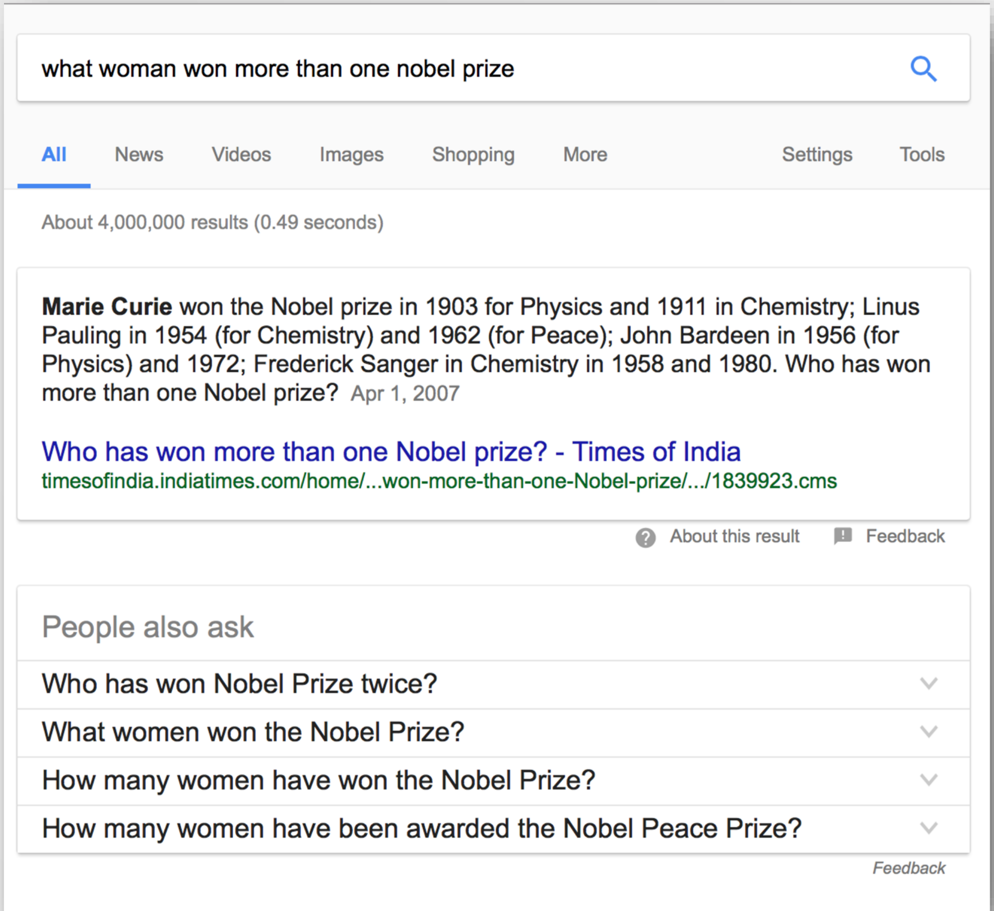
Dictation



Multimodal interaction

Preview: Information retrieval

Google



what woman won more than one nobel prize

About 4,000,000 results (0.49 seconds)

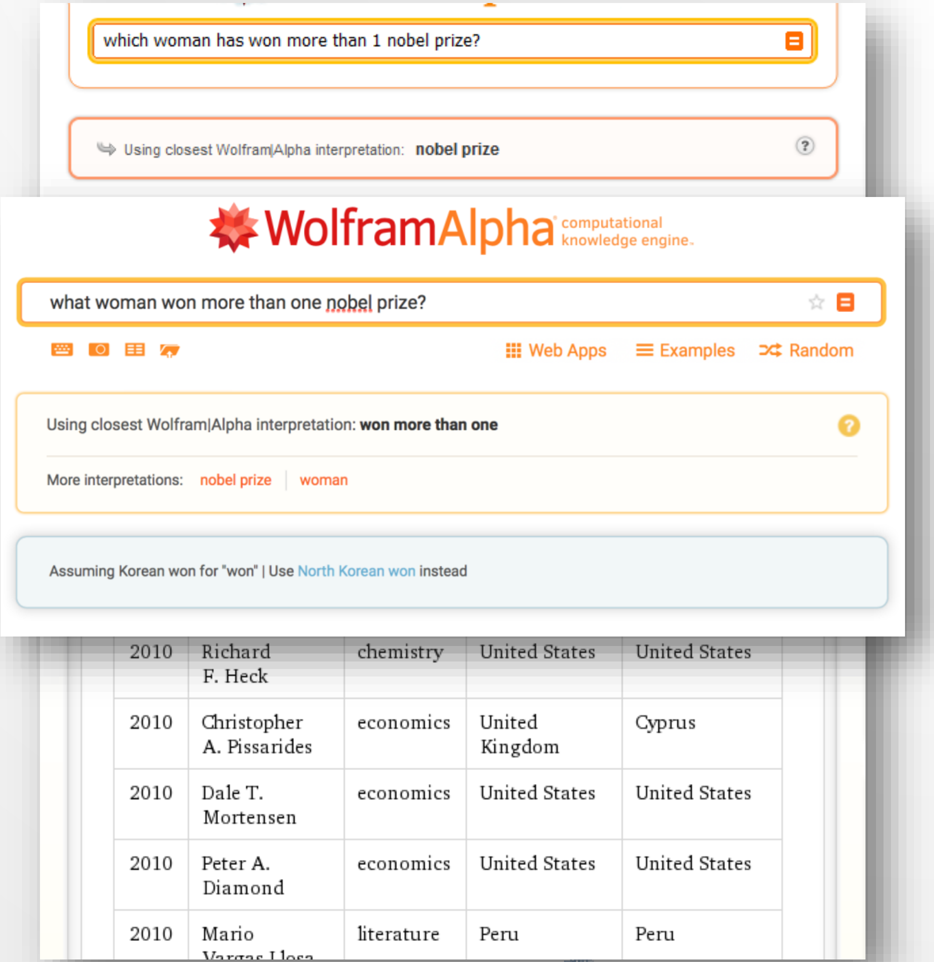
Marie Curie won the Nobel prize in 1903 for Physics and 1911 in Chemistry; Linus Pauling in 1954 (for Chemistry) and 1962 (for Peace); John Bardeen in 1956 (for Physics) and 1972; Frederick Sanger in Chemistry in 1958 and 1980. Who has won more than one Nobel prize? Apr 1, 2007

Who has won more than one Nobel prize? - Times of India
timesofindia.indiatimes.com/home/...won-more-than-one-Nobel-prize/.../1839923.cms

People also ask

- Who has won Nobel Prize twice?
- What women won the Nobel Prize?
- How many women have won the Nobel Prize?
- How many women have been awarded the Nobel Peace Prize?

WolframAlpha computational knowledge engine



which woman has won more than 1 nobel prize?

Using closest Wolfram|Alpha interpretation: **nobel prize**

WolframAlpha computational knowledge engine.

what woman won more than one nobel prize?

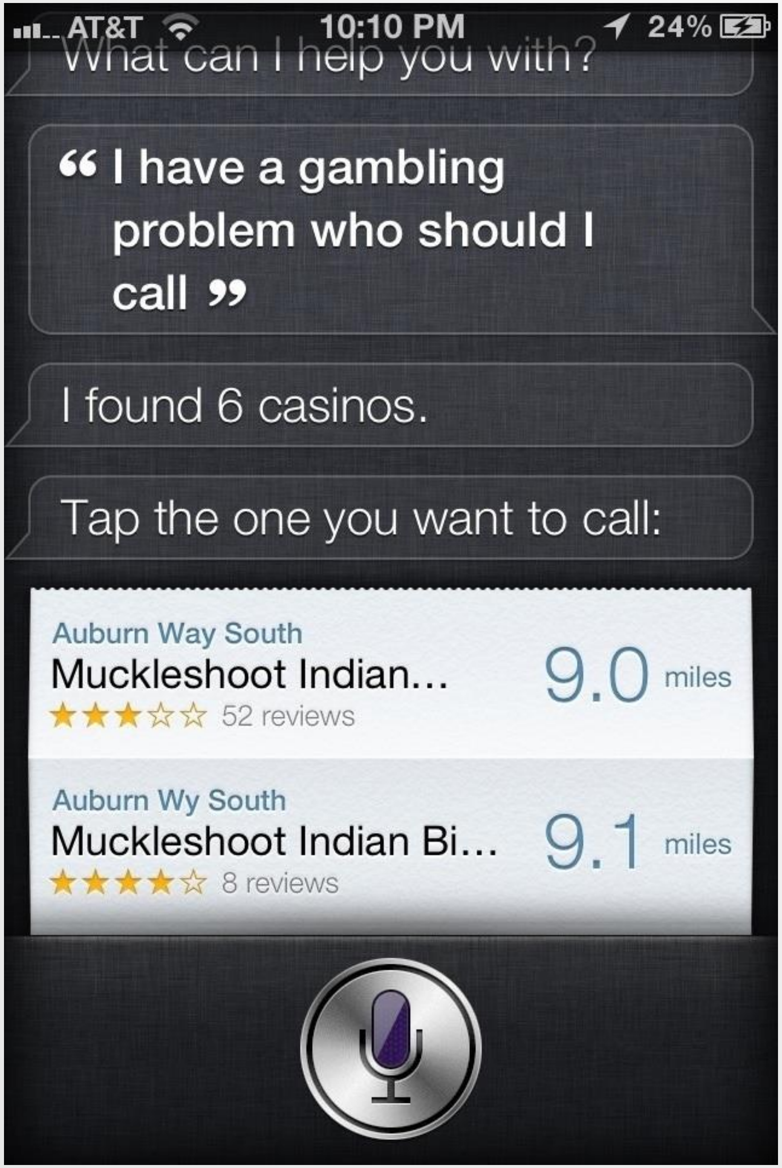
Using closest Wolfram|Alpha interpretation: **won more than one**

More interpretations: nobel prize | woman

Assuming Korean won for "won" | Use North Korean won instead

2010	Richard F. Heck	chemistry	United States	United States
2010	Christopher A. Pissarides	economics	United Kingdom	Cyprus
2010	Dale T. Mortensen	economics	United States	United States
2010	Peter A. Diamond	economics	United States	United States
2010	Mario Vargas Llosa	literature	Peru	Peru

Aside – Spoken Information Retrieval



AT&T 10:10 PM 24%

What can I help you with?

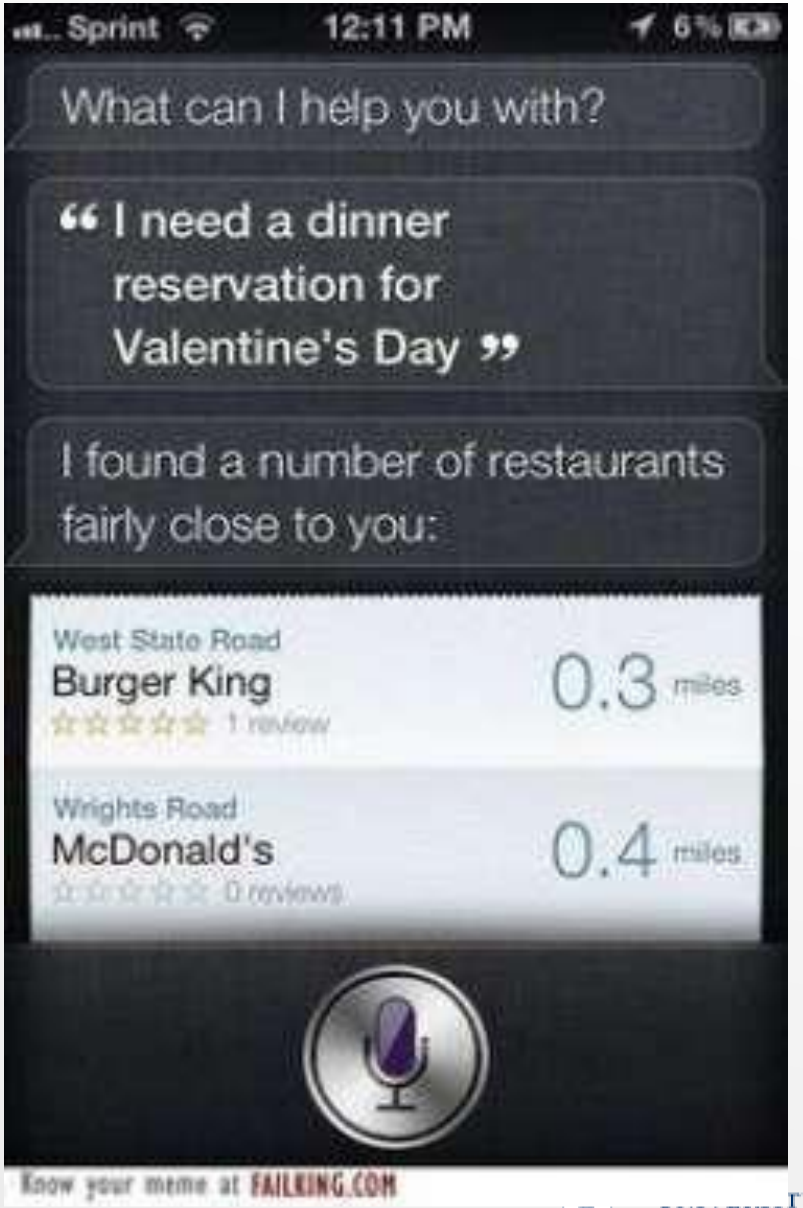
“ I have a gambling problem who should I call ”

I found 6 casinos.

Tap the one you want to call:

Auburn Way South
Muckleshoot Indian… 9.0 miles
★★★☆☆ 52 reviews

Auburn Wy South
Muckleshoot Indian Bi… 9.1 miles
★★★★☆ 8 reviews



Sprint 12:11 PM 6%

What can I help you with?

“ I need a dinner reservation for Valentine's Day ”

I found a number of restaurants fairly close to you:

West State Road
Burger King 0.3 miles
☆☆☆☆☆ 1 review

Wrights Road
McDonald's 0.4 miles
☆☆☆☆☆ 0 reviews

Know your meme at FAILKING.COM

Overview: NLP

- Is natural language processing (the discipline) hard?
  - **Yes**, because **natural language**
    - is highly ambiguous at all levels,
    - is complex and subtle,
    - is fuzzy and probabilistic,
    - involves real-world reasoning.
  - **No**, because **computer science**
    - gives us many powerful statistical techniques,
    - allows us to break the challenges down into more manageable features.
- Is Natural Language Computing (the course) hard?
  - More on this soon…

NLP in Industry

wattpad
Maluuba
A Microsoft company
cymfony
harnessing influence 2.0™
Google
Receptiviti
Nielsen
BuzzMetrics
BETA
hakia
search for meaning
OpenAI
umbria
collective intellect
YAHOO!
Microsoft
SST
SURGICAL SAFETY
TECHNOLOGIES
NUANCE
WINTERLIGHT LABS
IBM
PEARSON
Knowledge
Technologies
J.D.
AND ASSOCIATES
amazon
HUAWEI
THOMSON
REUTERS
at&t
WolframAlpha
computational knowledge engine

Natural language computing

- **Instructor**: Gerald Penn (gpenn@cs, M 4-6 in PT 283A)
- **Meetings**: MW (lecture), F (tutorial) at 10h and 11h
- **Languages**: English, Python.
- **Website**: ~~Quercus~~, www.cs.toronto.edu/~gpenn/csc401/
- **You**: Understand basic **probability**, can **program**,
  or (grads) can pick these up as we go.
- **Syllabus**: Key **theory** and **methods** in statistical natural language computing.
  Focus will be on ***neural models***, ***language models***, and their ***applications***.

Evaluation policies

- **General**: Three assignments : **20%** (each)
  Final exam: **39%**
  Two ethics surveys **: 0.5%** (each)
- **Lateness**: **10%** deduction applied to electronic submissions that are 1 minute late.
  Additional **10%** applied every 24 hours up to 72 hours total, at which point grade is **zero**.
- **Final**: If you **fail** (< 50%) the final exam, then you **fail** the course.
- **Ethics**: Plagiarism and unauthorized collaboration can result in a grade of **zero** on the homework, **failure** of the course, or **suspension** from the University.

Assignments

- Assignment 1: Corpus statistics, sentiment analysis

  task: analyze sentiment of financial reportage

  learn: statistical techniques, features, classification.

- Assignment 2: Neural machine translation

  task: translate between languages

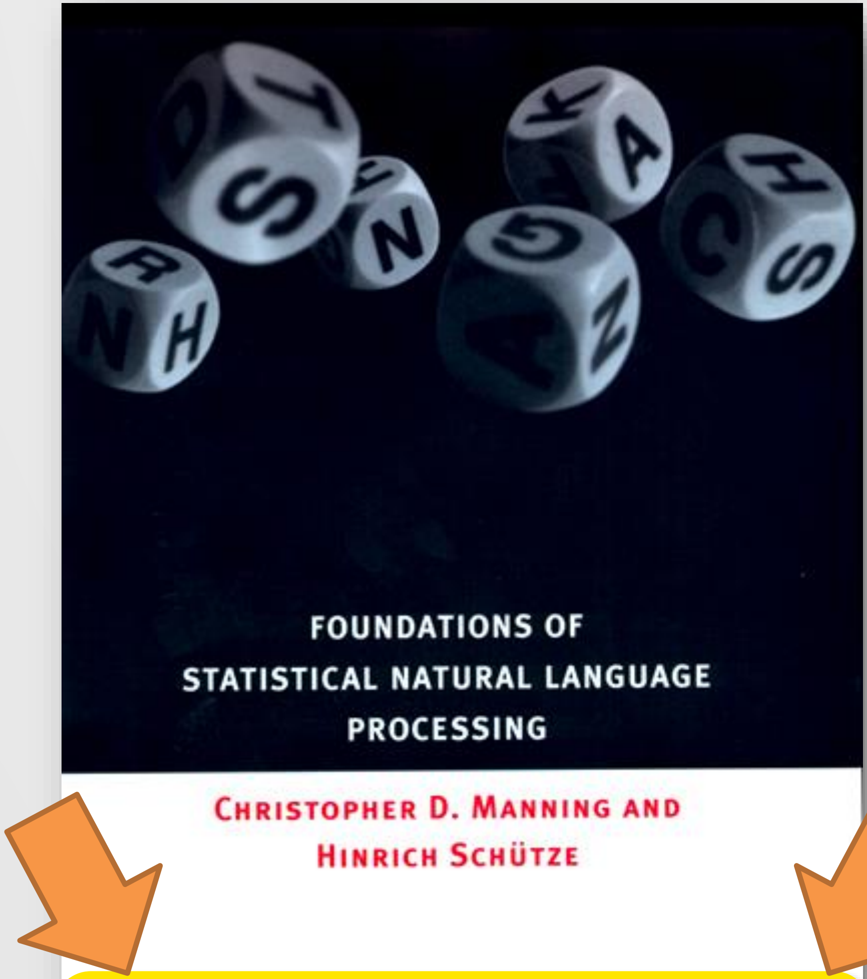
  learn: neural seq2seq and neural anguage models.

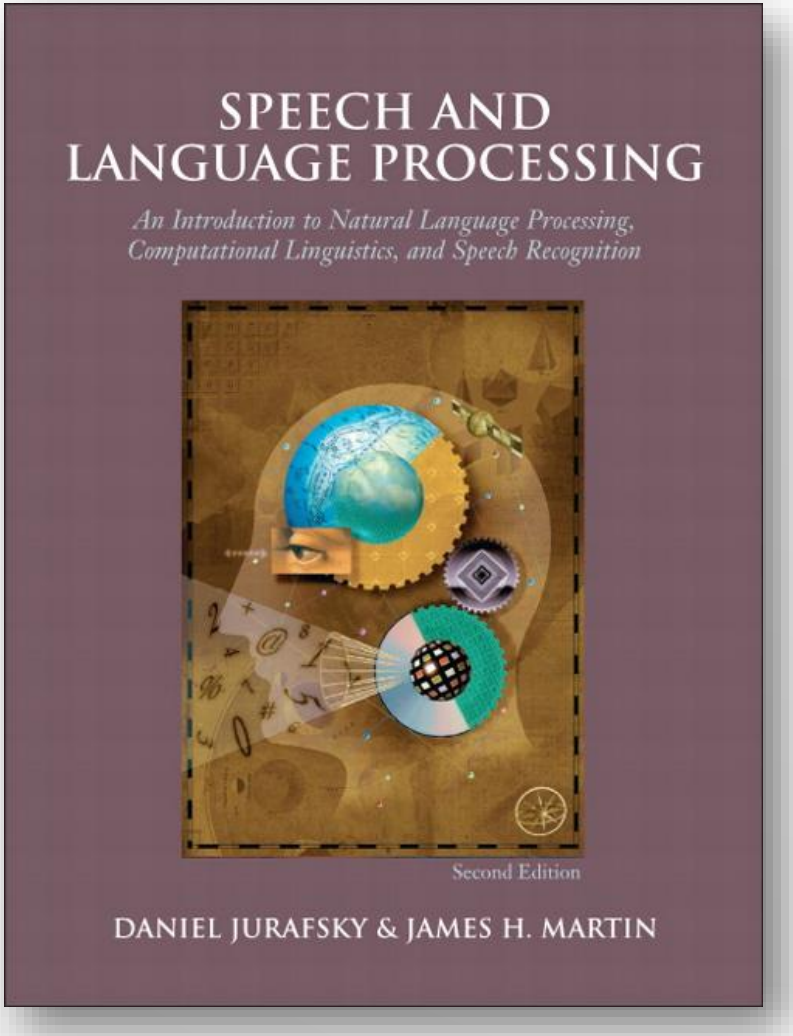
- Assignment 3: Automatic speech recognition

  task: detect lies in speech

  learn: signal processing, phonetics, dynamic algo's.

Reading


FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

http://tinyurl.com/shshhcvm

Assignment 1 – Financial sentiment

- Involves:
  - Working with real news data (e.g., Wall Street Journal),
  - Part-of-speech tagging (more on this later),
  - Large Language Models
  - Classification.
- **Announcements**: Piazza forum, email.
- Start early.

Assignment 1 and reading

- **Assignment 1** available soon (on course webpage)
  - Due 24 September / 8 October
  - TA:

    Winston Wu winstonyt.wu@mail.utoronto.ca
  - First tutorial: this Friday, 6th September

- **Reading**:
  - Manning & Schütze: Sections 1.3—1.4.2,

    Sections 6.0—6.2.1.