Course Information

http://www.cs.cmu.edu/~tbergkir/11711fa16/

11-711: Algorithms for NLP, Fall 2016

Instructors: Taylor Berg-Kirkpatrick and Robert Frederking
Lecture: Tuesday and Thursday 1:30pm-2:50pm, GHC 4307
Recitation: Friday 1:30pm-2:20pm, DH 1212
Office Hours: TBA

TAs: Wanli Ma and Kartik Goyal
Office Hours: TBA

Piazza setup soon!
Course Requirements

- **Prerequisites:**
  - Upper division algorithms (dynamic programming)
  - Mastery of basic probability
  - Strong skills in Java or equivalent
  - Deep interest in language

- **Work and Grading:**
  - Four assignments (individual, jars + write-ups)

- **Books:**
  - Primary text: Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
  - Also: Manning and Schuetze, Foundations of Statistical NLP
Other Announcements

- **Course Contacts:**
  - Webpage: materials and announcements
  - Piazza: discussion forum

- **Enrollment:** We’ll try to take everyone who meets the requirements

- **Computing Resources**
  - Experiments can take up to hours, even with efficient code
  - Recommendation: start assignments early

- **Questions?**
Goal: Deep Understanding
- Requires context, linguistic structure, meanings...

Reality: Shallow Matching
- Requires robustness and scale
- Amazing successes, but fundamental limitations
Speech Systems

- **Automatic Speech Recognition (ASR)**
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV

- **Text to Speech (TTS)**
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)
Example: Siri

- Siri contains
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.
... But Language is Complex

An iceberg is a large piece of freshwater ice that has broken off from a snow-formed glacier or ice shelf and is floating in open water.
Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- **SOTA:** ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples
Corpus-Based Methods

- A corpus like a treebank gives us three important tools:
  - It gives us broad coverage

```
ROOT → S
S → NP VP .
NP → PRP
VP → VBD ADJ
```
Corpus-Based Methods

- It gives us statistical information

### All NPs

- NP PP: 11%
- DT NN: 9%
- PRP: 6%

### NPs under S

- NP PP: 9%
- DT NN: 9%
- PRP: 21%

### NPs under VP

- NP PP: 23%
- DT NN: 7%
- PRP: 4%
Corpus-Based Methods

- It lets us check our answers
Semantic Ambiguity

- NLP is much more than syntax!
- Even correct tree structured syntactic analyses don’t fully nail down the meaning

*I haven’t slept for ten days*

*John’s boss said he was doing better*

- In general, every level of linguistic structure comes with its own ambiguities...
Other Levels of Language

- Tokenization/morphology:
  - What are the words, what is the sub-word structure?
  - Often simple rules work (period after “Mr.” isn’t sentence break)
  - Relatively easy in English, other languages are harder:
    - Segmentation

- Morphology
  
  - sarà  andata
  - be+fut+3sg  go+ppt+fem
  - “she will have gone”

- Discourse: how do sentences relate to each other?
- Pragmatics: what intent is expressed by the literal meaning, how to react to an utterance?
- Phonetics: acoustics and physical production of sounds
- Phonology: how sounds pattern in a language
Question Answering:

- More than search
- Ask general comprehension questions of a document collection
- Can be really easy: “What’s the capital of Wyoming?”
- Can be harder: “How many US states’ capitals are also their largest cities?”
- Can be open ended: “What are the main issues in the global warming debate?”

SOTA: Can do factoids, even when text isn’t a perfect match

capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words.
The noun capital of Wyoming has one meaning: Meaning #1: the capital.
www.answers.com/topic/capital-of-wyoming - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
Cheyenne (ˈʃi-ən, -ən) The capital of Wyoming, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-wyoming - 74k - Cached - Similar pages
Example: Watson

"a camel is a horse designed by a committee"
Summarization

- Condensing documents
- An example of analysis with generation

WASHINGTON (CNN) – President Obama’s inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy’s in 1961: “Ask not what your country can do for you. Ask what you can do for your country.”

Obama’s address was less stirring, perhaps, but it was also more candid and down-to-earth.

“Starting today,” the new president said, “we must begin the process of rebuilding the promise of America. This is the only way forward for America.”

Schneider: At a time of crisis, president must be reassuring

Country has chosen “hope over fear, unity of purpose over … discord,” Obama said.

Obama’s speech was a cool speech, not a hot one, Schneider says

President Obama renewed his call for a massive plan to stimulate economic growth.

STORY HIGHLIGHTS
- Obama’s address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen “hope over fear, unity of purpose over … discord,” Obama said
- Obama’s speech was a cool speech, not a hot one, Schneider says

Obama, too, offered reassurance.

“We gather because we have chosen hope over fear, unity of purpose over conflict and discord,” Obama said.

Obama’s call to unity after decades of political division echoed Abraham Lincoln’s first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln’s speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called “the better angels of our nature.”

Some presidents used their inaugural address to set out a bold agenda.
Lindsay Lohan pleaded not guilty Wednesday to felony grand theft of a $2,500 necklace, a case that could return the troubled starlet to jail rather than the big screen. Saying it appeared that Lohan had violated her probation in a 2007 drunken driving case, the judge set bail at $40,000 and warned that if Lohan was accused of breaking the law while free he would have her held without bail. The Mean Girls star is due back in court on Feb. 23, an important hearing in which Lohan could opt to end the case early.
Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (next class) vs fidelity (later)
"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

Video Anniversary of the Tibetan rebellion: China on guard

Agreement on the VAT: "Sarkozy wins case at the worst possible time"
The European finance ministers reached on Tuesday to a compromise allowing the reduction of VAT rates in some sectors, including catering.

Record Mixed reactions after the European agreement on reduction in VAT
| SOURCE | Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante. |
| HUMAN | That would be an interim solution which would make it possible to work towards a binding charter in the long term. |
| 1x DATA | [this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.] |
| 10x DATA | [it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.] |
| 100x DATA | [this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.] |
| 1000x DATA | [that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.] |
Business

Latest News

- **The exchange of financial stocks fell slightly prominent lower**

  12 stocks in Tokyo, ahead of sell orders from the backlash of higher yesterday, with slightly lower values. Nikkei ... ... (11:13) [Full article]

- **Negotiation and integration of Japan Sompo Japan to aggregate in three large camps**

  Sompo Japan Insurance and it's five to start the negotiations for the merger of NIPPONKOA Insurance Co., Ltd. No. 12, 2007, minutes ... ... ... (10:33) [Full
Q: Who signed the Serve America Act?
A: Barack Obama

Los Angeles Times

President Barack Obama received the Serve America Act after congress’ vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.
President Barack Obama received the Serve America Act after congress’ vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.
Example Errors

Input
America Online announced on Monday that the company plans to update its instant messaging service.

Correct
America Online the company its instant messaging service

Guess
America Online the company its instant messaging service
America Online is an American global Internet services and media company operated by Time Warner. It is headquartered at 770 Broadway in Midtown Manhattan, New York City. Founded in 1983 as Quantum Computer Services, it has franchised its services to companies in several nations around the world or set up international versions of its services.
Grounded Language
Grounding with Natural Data

... on the beige loveseat.
What is Nearby NLP?

- **Computational Linguistics**
  - Using computational methods to learn more about how language works
  - We end up doing this and using it

- **Cognitive Science**
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!

- **Speech Processing**
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP
What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We’ll focus on what makes the problems hard, and what works in practice...
Class Requirements and Goals

Class requirements
- Uses a variety of skills / knowledge:
  - Probability and statistics, graphical models
  - Basic linguistics background
  - Strong coding skills (Java)
- Most people are probably missing one of the above
- You will often have to work on your own to fill the gaps

Class goals
- Learn the issues and techniques of statistical NLP
- Build realistic NLP tools
- Be able to read current research papers in the field
- See where the holes in the field still are!
Some Early NLP History

- **1950’s:**
  - Foundational work: automata, information theory, etc.
  - First speech systems
  - Machine translation (MT) hugely funded by military
    - Toy models: MT using basically word-substitution
  - Optimism!

- **1960’s and 1970’s: NLP Winter**
  - Bar-Hillel (FAHQT) and ALPAC reports kills MT
  - Work shifts to deeper models, syntax
  - ... but toy domains / grammars (SHRDLU, LUNAR)

- **1980’s and 1990’s: The Empirical Revolution**
  - Expectations get reset
  - Corpus-based methods become central
  - Deep analysis often traded for robust and simple approximations
  - *Evaluate everything*

- **2000+: Richer Statistical Methods**
  - Models increasingly merge linguistically sophisticated representations with statistical methods, confluence and clean-up
  - *Begin to get both breadth and depth*
Problem: Structure

- **Headlines:**
  - Enraged Cow Injures Farmer with Ax
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor’s Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

- Why are these funny?
**Problem: Scale**

- People *did* know that language was ambiguous!
  - ...but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
  - ...they didn’t realize how bad it would be
Classical NLP: Parsing

- Write symbolic or logical rules:
  
  Grammar (CFG) | Lexicon
  --- | ---
  ROOT $\rightarrow$ S | NN $\rightarrow$ interest
  S $\rightarrow$ NP VP | NNS $\rightarrow$ raises
  NP $\rightarrow$ DT NN | VBP $\rightarrow$ interest
  NP $\rightarrow$ NN NNS | VBZ $\rightarrow$ raises
  NP $\rightarrow$ NP PP | VP $\rightarrow$ VBP NP
  VP $\rightarrow$ VBP NP PP | ...  

- Use deduction systems to prove parses from words
  - Minimal grammar on “Fed raises” sentence: 36 parses
  - Simple 10-rule grammar: 592 parses
  - Real-size grammar: many millions of parses

- This scaled very badly, didn’t yield broad coverage tools
Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire
Outline of Topics

- **Words and Sequences**
  - Speech recognition
  - N-gram models
  - Working with a lot of data

- **Structured Classification**

- **Trees**
  - Syntax and semantics
  - Syntactic MT
  - Question answering

- **Machine Translation**

- **Other Topics**
  - Reference resolution
  - Summarization
  - Diachronics
  - ...

- **Diachronics**...
What’s Next?

- Next class: noisy-channel models and language modeling
  - Introduction to machine translation and speech recognition
  - Start with very simple models of language, work our way up
  - Some basic statistics concepts that will keep showing up

- If you don’t know what conditional probabilities and maximum likelihood estimators are, read up!

- Reading on the web
A Puzzle

- You have already seen N words of text, containing a bunch of different word types (some once, some twice...)

- What is the chance that the N+1\textsuperscript{st} word is a new one?