Lecture 1: Introduction
Taylor Berg-Kirkpatrick – CMU
Slides: Dan Klein – UC Berkeley
Course Information

www.cs.cmu.edu/~tbergkir/11711fa17/

11-711: Algorithms for NLP, Fall 2017

Instructors: Taylor Berg-Kirkpatrick and Robert Frederking
Lecture: Tuesday and Thursday 1:30pm-2:50pm, DH 1212
Recitation: Friday 1:30pm-2:20pm, MM 103
Office Hours: TBA

TAs: Hieu Pham and Maria Ryskina
Office Hours: TBA
Forum: Piazza (link to appear soon)

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)

“Speech Lab”
Example: Siri

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

*Image: Wikipedia*
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
  |   
NP   VP
  |   |
PRP  VBD  ADJP
  |   |
He   was  JJ
      |   |
      right

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
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:
    - Segementation

- **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
Example: Watson

"a camel is a horse designed by"

The Phrase Finder

A camel is a horse designed by committee

Posted by Ruben P. Mendez on April 16, 2004

Does anyone know the origin of this maxim? I heard it way back at the United Nations, which is chockfull of committees. It may have originated there, but I'd like an authoritative explanation. Thanks

- Re: A camel is a horse designed by committee SR 16/April/04
  - Re: A camel is a horse designed by committee Henry 18/April/04
Summarization

- Condensing documents
- An example of analysis with generation
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 the 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 VAT reduction in VAT sectors. The reduction of VAT in the case of Taxis...
<table>
<thead>
<tr>
<th>SOURCE</th>
<th>Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUMAN</td>
<td>That would be an interim solution which would make it possible to work towards a binding charter in the long term.</td>
</tr>
<tr>
<td>1x DATA</td>
<td>[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]</td>
</tr>
<tr>
<td>10x DATA</td>
<td>[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]</td>
</tr>
<tr>
<td>100x DATA</td>
<td>[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]</td>
</tr>
<tr>
<td>1000x DATA</td>
<td>[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]</td>
</tr>
</tbody>
</table>
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興亜 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 article]
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

Guess
America Online, LLC (commonly known as AOL) 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.\(^2\)\(^3\) 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.\(^4\)

America Online

<table>
<thead>
<tr>
<th>Type</th>
<th>Subsidiary of Time Warner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Founded</td>
<td>1983 as Quantum Computer Services</td>
</tr>
</tbody>
</table>
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?**
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 → S | NN → interest
  S → NP VP | NNS → raises
  NP → DT NN | VBP → interest
  NP → NN NNS | VBZ → raises
  VP → 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
  - ...


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
AI: Where Do We Stand?

Hollywood

R2D2

KITT

Wall-E

Reality

Rule based approaches

Early statistical approaches

Modern statistical approaches

Google Translate '08
Almost half of Google News users to read only the titles (poll)

Almost half of the portal users to Google News merely read the headlines without clicking to access the items to which they refer to the websites of various newspapers, according to a poll released Tuesday.

Although Google generates some traffic on the (sites) newspapers, it also diverts much," said Ken Doctor, an analyst at consulting firm Outsell, which conducted the survey, noting that "not less than 44% of Google News visitors looking titles without going on newspaper sites.

This survey seems to bring water to appease critics of Google, such as media mogul Rupert Murdoch, News Corp. boss., Accusing Google and other aggregators do not share advertising revenue with the sites to which they provide links.

Près de la moitié des utilisateurs de Google News ne lisent que les titres (sondage)

Près de la moitié des utilisateurs du portail d'actualités Google News se contentent de lire les titres, sans cliquer dessus pour accéder aux articles auxquels ils renvoient sur les sites des différents journaux, selon un sondage publié mardi.

"Bien que Google génère une certaine fréquentation sur les (sites des) journaux, il en détourné aussi une bonne partie", note Ken Doctor, analyste au cabinet de conseil Outsell qui a réalisé ce sondage, soulignant que "pas moins de 44% des visiteurs de Google News regardent les titres sans aller sur les sites des journaux".

Ce sondage semble apporter de l'eau au moulin des critiques de Google, comme le magnat des médias Rupert Murdoch, patron de News Corp., qui accuse Google et d'autres agrégateurs de ne pas partager les recettes publicitaires avec les sites vers lesquels ils proposent des liens.
Next Week

- We’re going to do an experiment in competitive parsing

- Polls:
  - How many have an EECS research account?
  - How many have a laptop they can bring next week?
  - How many know what a prepositional phrase is?

- Also: Assignment 1 will be out very soon, and assignments will move fast
新华网石家庄11月16日电（记者 张涛）11月15日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至1℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 ℃. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.
Discourse Ambiguity: Reference

“The American Medical Association voted yesterday to install the heir apparent as its president-elect, rejecting a strong, upstart challenge by a District doctor who argued that the nation’s largest physicians’ group needs stronger ethics and new leadership.”

- Can link repeated mentions of an entity both inside a document and across documents, model variation in spelling, transliteration, etc.
- SOTA: varies widely, 70-90%
Much of NLP is annotating text with structure which specifies how it’s assembled.

- Syntax: grammatical structure
- Semantics: “meaning,” either lexical or compositional

John bought a blue car
Why is NLP Hard?

The core problems:

- Ambiguity
- Sparsity
- Scale
- Unmodeled Variables
Syntactic Ambiguities

- Maybe we’re sunk on funny headlines, but normal, boring sentences are unambiguous?

Fed raises interest rates 0.5 % in a measure against inflation
Sometimes life is easy

- Can do text classification pretty well just knowing the set of words used in the document, same for authorship attribution
- Word-sense disambiguation not usually needed for web search because of majority effects or intersection effects (“jaguar habitat” isn’t the car)

Sometimes only certain ambiguities are relevant

*he hoped to record a world record*

Other times, all levels can be relevant (e.g., translation)
Language isn’t Adversarial

- One nice thing: we know NLP can be done!

Language isn’t adversarial:
- It’s produced with the intent of being understood
- With some understanding of language, you can often tell what knowledge sources are relevant

But most variables go unmodeled
- Some knowledge sources aren’t easily available (real-world knowledge, complex models of other people’s plans)
- Some kinds of features are beyond our technical ability to model (especially cross-sentence correlations)
These 7 people include astronauts coming from France and Russia.
**Machine Translation**

**Atlanta, preso il killer del palazzo di Giustizia**

**ATLANTA** - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

**Atlanta, taken the killer of the palace of Justice**

**ATLANTA** - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

- Translation systems encode:
  - Something about fluent language (next class)
  - Something about how two languages correspond (middle of term)
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators
Example Syntax-Based Translation

ورفض البار الأدلة بأي تصريحات فور وصوله إلى المقاطعة.

tac-lang: urfD albaz aladla' baā tSryHat fur uSu l alā almqaT'e.
bckwlr: wrfD AlbAz AlAdIA' bAY tSryHA t fwr wSwl h AIY AlmqA TEp.
Tune.nw.0: al @-@ baz declined to make any statements upon his arrival in the province.
The Dream

- It’d be great if machines could
  - Manage information in our email
  - Translate languages accurately
  - Help us process, summarize, and aggregate text information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us

- But they can’t:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge

- So:
Analysis and Alignment

[Burkett, Blitzer, and Klein 10]
They considered running the ad during the Super Bowl.

running * during: 3k
running it during: 239
considered * during: 2k
considered it during: 112

[Bansal and Klein 11]
Unsupervised Learning

\[ Z_1 \rightarrow W_1 \]
\[ Z_2 \rightarrow W_2 \]
\[ Z_3 \rightarrow W_3 \]

Parameters:

<table>
<thead>
<tr>
<th>0.20</th>
<th>0.45</th>
<th>0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55</td>
<td>0.28</td>
<td>0.17</td>
</tr>
<tr>
<td>0.31</td>
<td>0.51</td>
<td>0.18</td>
</tr>
<tr>
<td>0.06</td>
<td>0.70</td>
<td>0.24</td>
</tr>
<tr>
<td>0.62</td>
<td>0.13</td>
<td>0.15</td>
</tr>
</tbody>
</table>
Coreference Systems

System Accuracy

<table>
<thead>
<tr>
<th>Reference F1</th>
<th>Unsupervised (Haghigi and Klein 07)</th>
<th>Supervised (Bengston &amp; Roth 08)</th>
<th>Unsupervised + Data (Haghigi and Klein 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64.2</td>
<td>76.8</td>
<td>80.8</td>
</tr>
</tbody>
</table>
Cross-Document Identity

Daniel Klein

I am Professor of Economics at George Mason University. I earned my PhD from George Mason University and New York University, where I studied the classical liberal traditions of economics. My research focuses on economic principles, public policy issues, and the tradition of Adam Smith and Friedrich Hayek.

I've published research on policy issues including toll roads, auto insurance, credit reporting, and the Food and Drug Administration. I've also written on spontaneous order, coordination, the demand and supply of assurance, why governments believe in the goodness of bad policy, why people favor government intervention more than they should, and the role of liberty, dignity, and responsibility.

Lately my research dwells on Adam Smith, Russ Roberts and a multi-part audio book club on The Theory of Moral Sentiments.

I participate in the new GMU Econ course sequence and field exam in Smithian Political Economy (SPE). The first SPE field exam will be given in August 2011. (Some sample field questions.) Also, I lead an Adam Smith reading group and an Invisible Hand Seminar for graduate students.

I am the chief editor of Econ Journal Watch, an online journal dedicated to economic criticism from a Smith-Hayek viewpoint. I've contributed several papers on the character heterogeneity of economists. I push the
Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

<table>
<thead>
<tr>
<th>Person</th>
<th>Company</th>
<th>Post</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times newspaper</td>
<td>president and general manager</td>
<td>start</td>
</tr>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times newspaper</td>
<td>executive vice president</td>
<td>end</td>
</tr>
<tr>
<td>Lance R. Primis</td>
<td>New York Times Co.</td>
<td>president and CEO</td>
<td>start</td>
</tr>
</tbody>
</table>

- SOTA: perhaps 80% accuracy for multi-sentence temples, 90%+ for single easy fields
- But remember: information is redundant!
Dark Ambiguities

- **Dark ambiguities**: most structurally permitted analyses are so bad that you can’t get your mind to produce them.

This analysis corresponds to the correct parse of

“This will panic buyers!”

- **Unknown words and new usages**
- **Solution**: We need mechanisms to focus attention on the best ones, probabilistic techniques do this.
Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora

- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what