CS5740: Natural Language Processing

Introduction

Instructor: Yoav Artzi
TA: Max Grusky

Slides adapted from Dan Klein, Luke Zettlemoyer, and Yejin Choi
Technicalities

• People:
  – Instructor: Yoav Artzi
    • Office hours: Monday 5pm, Baron
  – TA: Max Grusky:
    • Office hours: Thursday, 1pm, by Skype (coordinate)

• Webpage (everything is there):
  – http://www.cs.cornell.edu/courses/cs5740/2017sp/

• Discussion group on Piazza
• Chat on Slack
• Assignments on CMS
  – Repositories on Github Classroom
Technicalities

• Grading:
  – 40% assignments, 25% exam, and 30% class review quizzes, 5% participation
  – Participation = class + Piazza + Slack

• Enrollment and prerequisites:
  – At least B in CS 5785 (Applied ML) or equivalent Cornell Course
  – Or: instructor permission
  – Audit? Talk to me after class
Technicalities

• Quizzes:
  – First five minutes of every class, no extensions
  – Each quiz: 1.5% of the grade, up to 30%, only top 20 quizzes count
  – It is not possible to re-take a missed quiz
  – A missed quiz gets zero
  – Just like an exam: no copying, chatting, and not taking the quiz remotely → all AI violations

• Quiz practice
  – Phones and laptops
  – http://socrative.com
  – Use NetID to identify
  – Today’s room: NLP5
Technicalities

• Collaboration:
  – All assignments must be done in pairs

• Use of external code/tools – specified in each assignment
  – If have doubt – ask!

• Late submissions:
  – 10% off for every 12 hours, rounded up
    • E.g., 25 hours late → grade starts at 70
  – No late submission for final exam

• All assignments should be implemented in Python
Technicalities

• Books (recommended, not required):
  – D. Jurafsky & James H. Martin, Speech and Language Processing
  – C.D. Manning & H. Schuetze, Foundations of Statistical Natural Language Processing

• Other material on the course website
Technicalities

• Come on time
  – Late? Enter quietly and sit at the back
  – Quiz starts on time

• No laptops or phones in class
  – Except during the quiz
WHY ARE YOU HERE?
What is this class?

• Depth-first technical NLP course
• Learn the language of natural language processing
• What this class is not?
  – It is not a tutorial to NLTK, TensorFlow, etc.
  – Stack Overflow already does this well
Class Goals

• Learn about the issues and techniques of modern NLP
• Be able to read current research papers
• Build realistic NLP tools
• Understand the limitation of current techniques
Main Themes

• Linguistic Issues
  – What are the range of language phenomena?
  – What are the knowledge sources that let us make decisions?
  – What representations are appropriate?
• 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 …
Main Models

• Generative Models
• Discriminative Models
  – Neural Networks
• Graphical Models
What is NLP?

- **Fundamental goal:** deep understanding of broad language
  - Not just string processing or keyword matching!
- **End systems that we want to build:**
  - Simple:
  - Complex:
What is NLP?

- **Fundamental goal:** deep understanding of broad language
  - Not just string processing or keyword matching!
- **End systems that we want to build:**
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)
Today

• Prominent applications
  – Try to imagine approaches
  – What’s behind current limitations?
• Some history
• Key problems
Machine Translation

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? How to combine? [learning to translate]
  - How to make efficient? [fast translation search]
La Bourse de Shanghai dégringolait de plus de 6 % mardi 25 août à l’ouverture, après s’être déjà effondrée de presque 8,5 % la veille, dans un marché affolé par l’affaiblissement persistant de l’économie chinoise et miné par des inquiétudes sur la conjoncture mondiale.

Dans les premiers échanges, l’indice composite chutait de 6,41 % soit 205,78 points, à 3 004,13 points. La Bourse de Shenzhen plongeait quant à elle de 6,97 %.

The Shanghai Stock Exchange tumbled more than 6% Tuesday, August 25 at the opening, having already collapsed by almost 8.5% yesterday, in a panicked market the persistent weakening of the Chinese economy and undermined by concerns about the global economy.

In early trade, the composite index fell by 6.41% or 205.78 points to 3 004.13 points. The Shenzhen Stock Exchange dived for its 6.97% to 1 751.28 points. The Hong Kong Stock Exchange, meanwhile, opened down 0.67%.
A spread of global stocks decline
US stocks opened Monday fell 1,000 points
NATHANIEL POPPER, NEIL GOUGH 09:54

Monday, A-share market fell 8.5 percent, taking all the gains this year. Investors worried about the economic downturn runaway Chinese stock market "Black Monday" spread to the US and European and Asian markets, the Dow opened down over a thousand points within minutes.
Kiek Lietuvoje kainuoja užsienyje vogtas dviratis? Kokiais keliais jie čia patenka ir kodėl policija pro pirštus žiūri į klastinčią prekybą vogtais daiktais? Atsakymų į šiuos bei kitus klausimus ieškovo Lietuvoje viešėjusi Danijos valstybinės televizijos „DR“...

As far as Lithuania free bike stolen abroad? In what ways are placed here and why the police connive at a thriving trade in stolen items? Answers to these and other questions put Lithuania who visited the Danish public television DR ...
Summarization

• Condensing documents
  – Single or multiple docs
  – Extractive or abstractive
• Very context-dependent!
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</td>
<td>president and general manager</td>
<td>start</td>
</tr>
<tr>
<td></td>
<td>newspaper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell T. Lewis</td>
<td>New York Times</td>
<td>executive vice president</td>
<td>end</td>
</tr>
<tr>
<td></td>
<td>newspaper</td>
<td></td>
<td></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: good performance on simple templates (e.g., person-role)
• Harder without defining template
Tagging: Back to Text

Court Ruling in Kansas Makes Senate Race a True Tossup
by Nate Cohn
Question Answering

The Knowledge Graph
Learn more about one of the key breakthroughs behind the future of search.
Question Answering

- More than search

What’s the capital of Wyoming?

About 984,000 results (0.54 seconds)

Wyoming / Capital

Cheyenne
Question Answering

• More than search

How many US states’ capitals are also their largest cities?

State Capitals and Largest Cities - Infoplease
www.infoplease.com › United States › States ▾
State Capitals and Largest Cities. The following table lists the capital and largest city of every state in the United States. State, Capital, Largest city.

State Capitals and Largest Cities - Fact Monster
www.factmonster.com › United States › States ▾ Fact Monster ▾
State Capitals and Largest Cities. The following table lists the capital and largest city of every state in the United States. State, Capital, Largest city.

List of capitals in the United States - Wikipedia, the free ...
https://en.wikipedia.org/.../List_of_capitals_in_the_United_Sta... ▾ Wikipedia ▾
Austin is the largest state capital that is not also the state's largest city. .... The Confederate States of America had two capitals during its existence. The first ... In many cases, former capital cities of states are outside the current state borders.
State capitals - Insular area capitals - Former national capitals
Question Answering

• More than search

What are the main issues in the global warming debate?

Global warming controversy - Wikipedia, the free encyclopedia
Jump to Mainstream scientific position, and challenges to it - [edit]. Main article:
Scientific opinion on climate change. Summary of opinions from climate ...

Climate Change ProCon.org
climatechange.procon.org
The pro side argues rising levels of atmospheric greenhouse gases are a direct result of human activities such as burning fossil fuels, and that these increases are causing significant and increasingly severe climate changes including global warming, loss of sea ice, sea level rise, stronger storms, and more droughts.
Is human activity a substantial - Footnotes & Sources - Carbon Dioxide (CO2)

Climate Change and Global Warming — Global Issues
www.globalissues.org/issue/178/climate-change-and-global-warming
Some of the major conferences in recent years are also discussed. 32 articles on "Climate Change and Global Warming" and 1 related issue: ...
Question Answering

WolframAlpha

Oscar for best actress 1958

Assuming year of award ceremony | Use year of film release instead

Input Interpretation:

Academy Awards
actress in a leading role
1958 (year of award ceremony)

Result:

Joanne Woodward in The Three Faces of Eve

Other nominees:

Lana Turner in Peyton Place | Elizabeth Taylor in Raintree County | Deborah Kerr in Heaven Knows, Mr. Allison | Anna Magnani in The Wind

Information about Joanne Woodward:

<table>
<thead>
<tr>
<th>full name</th>
<th>Joanne Gigailliat Trimmier Woodward</th>
</tr>
</thead>
<tbody>
<tr>
<td>date of birth</td>
<td>Thursday February 27, 1930 (age: 82 years)</td>
</tr>
<tr>
<td>place of birth</td>
<td>Thomasville, Georgia, United States</td>
</tr>
</tbody>
</table>

"What's the best movie to see this weekend"

That would probably start an argument. But here’s a list of highly-regarded movies:
Natural Language Instruction

• What makes this possible?
• Limitations?
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.
"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake" because he was
(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery
Bang, bang, his silver hammer came down upon her head
Speech Systems

• Automatic Speech Recognition (ASR)
  – Audio in, text out
  – SOTA: 16% PER, Google claims 8% WER

• Text to Speech (TTS)
  – Text in, audio out
  – SOTA: mechanical and monotone
“Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers”

-- David Stork (HAL’s Legacy, 2001) on A. Rosenfeld’s vision
Image Captioning: The Good

a bunch of bananas sitting on top of a wooden table
logprob: -8.52

a man is playing tennis on a tennis court
logprob: -6.77

a train traveling down tracks next to a lush green field
logprob: -7.65

a pizza with toppings on a white plate
logprob: -7.40

http://cs.stanford.edu/people/karpathy/deepimagesent/generationdemo/
Image Captioning: The Bad

- A young boy is eating a piece of cake  
  logprob: -7.75

- A pair of scissors with a pair of scissors  
  logprob: -9.07

- A man is holding a cell phone in his hand  
  logprob: -8.90

- A large jetliner flying through a blue sky  
  logprob: -5.79
Image Captioning: The Sitting

- A cat is sitting on a toilet seat
  logprob: -7.95

- A pizza sitting on top of a white plate
  logprob: -6.15

- A bunch of luggage sitting on top of a hard wood floor
  logprob: -10.50

- A large airplane sitting on top of an airport runway
  logprob: -6.70

- A laptop computer sitting on top of a wooden desk
  logprob: -6.38

- A group of people sitting around a table with a cake
  logprob: -8.83
NLP History: Pre-statistics

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless
NLP History: Pre-statistics

(1) Colorless green ideas sleep furiously.
(2) Furiously sleep ideas green colorless

It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not.” (Chomsky 1957)
NLP History: Pre-statistics

• 70s and 80s: more linguistic focus
  – Emphasis on deeper models, syntax and semantics
  – Toy domains / manually engineered systems
  – Weak empirical evaluation
NLP History: ML and Empiricism

“Whenever I fire a linguist our system performance improves.” – Jelinek, 1988

• 1990s: Empirical Revolution
  – Corpus-based methods produce the first widely used tools
  – Deep linguistic analysis often traded for robust approximations
  – Empirical evaluation is essential
NLP History: ML and Empiricism

“Whenever I fire a linguist our system performance improves.” – Jelinek, 1988

- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential

“Of course, we must not go overboard and mistakenly conclude that the successes of statistical NLP render linguistics irrelevant (rash statements to this effect have been made in the past, e.g., the notorious remark, “Every time I fire a linguist, my performance goes up”). The information and insight that linguists, psychologists, and others have gathered about language is invaluable in creating high-performance broad-domain language understanding systems; for instance, in the speech recognition setting described above, a better understanding of language structure can lead to better language models.”

NLP History: ML and Empiricism

“Whenever I fire a linguist our system performance improves.” – Jelinek, 1988

• 1990s: Empirical Revolution
  – Corpus-based methods produce the first widely used tools
  – Deep linguistic analysis often traded for robust approximations
  – Empirical evaluation is essential

• 2000s: Richer linguistic representations used in statistical approaches, scale to more data!

• 2010s: you decide!
Related Fields

• 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?
  – 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
Key Problems

We can understand programming languages. Why is NLP not solved?
Key Problems

We can understand programming languages. Why is NLP not solved?

• Ambiguity
• Scale
• Sparsity
Key Problem: Ambiguity

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

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.
Semantic Ambiguity

At last, a computer that understands you like your mother.

[Example from Lillian Lee]
Semantic Ambiguity

At last, a computer that understands you like your mother.

- **Direct Meanings:**
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
  - It understands you like (it understands) your mother
- **But there are other possibilities, e.g.** mother could mean:
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
- **Context matters, e.g. what if** previous sentence was:
  - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients. 🍃

[Example from Lillian Lee]
Key Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be “good” ones (or ruled out pragmatically)
Key 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
Key Problem: Sparsity

• 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
Key Problem: Sparsity

• However: sparsity is always a problem
  – New unigram (word), bigram (word pair)
The NLP Community

- Conferences: **ACL**, **NAACL**, **EMNLP**, EACL, CoNNL, COLING, *SEM, LREC, CICLing, …
- Journals: CL, **TACL**, …
- Also in AI and ML conferences: AAAI, IJCAI, ICML, NIPS