CS6741: Structured Prediction for NLP
Fall 2015

Introduction

Instructor: Yoav Artzi

Slides adapted from Dan Klein and Luke Zettlemoyer
Before We Start

• You are all natural language processors!
• Constantly think of your language (English included)
• Come up with cool examples and share with the class!
What is NLP?

- **Fundamental goal:** deep understand 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?)
Text Categorization

• Input: Document
• Output: Category assignment

Barack Obama → US President
World War II → War
Caloboletus calopus → Mushroom
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

“speech lab”
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!
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?

About 982,000,000 results (0.67 seconds)

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
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 ...
hits:en.wikipedia.org/.../List_of_capitals_in_the_United_Sta...
List of capitals in the United States - Wikipedia, the free encyclopedia. The table below lists the capital and largest city of every state in the United States. State, Capital, Largest city.
Question Answering

• More than search
NL Interaction

place your back against the wall of the t intersection

turn left

go forward along the pink flowered carpet hall two segments to the intersection with the brick hall
NL Understanding

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. Allen
- Anna Maria in The Wind

Information about Joanne Woodward:

| full name | Joanne Gignilliat Trimmier Woodward |
| date of birth | Thursday February 27, 1930 (age: 82 years) |
| place of birth | Thomasville, Georgia, United States |
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 (second half of this class) vs fidelity (later)
A spread of global stocks decline
US stocks opened Monday fell 1,000 points

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.
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

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%.
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!
"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
NLP History: Pre-statistics

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

(Chomsky 1957)
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

• 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.

• 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]
Dark Ambiguities

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

• Three major themes:
  – Sequences
  – Trees
  – Logical forms

• Three aspects:
  – Linguistic formalism
  – Inference
  – Learning
Topics: Sequences

• Part-of-speech (POS) Tagging, Named Entity Recognition (NER), Chuncking
• HMM, MEMM, CRF
Topics: Trees

A ship-shipping ship, shipping shipping-ships.
Topics: Trees

A ship-shipping ship, shipping shipping-ships.
A ship-shipping ship, shipping shipping-ships.

- Constituency Parsing, Dependency Parsing
- CKY Parsing, Transition-based Parsing
- PCFGs, CCG
- TreeBank Parsing
Topics: Logical Forms

Are trees enough?

Move to the chair
Topics: Logical Forms

Are trees enough?

Move to the chair

Need a way to formally represent meaning!
Topics: Logical Forms

\[ \lambda a. \text{move}(a) \land \text{to}(a, \iota(\lambda x. \text{chair}(x))) \]

- Compositional Semantics, Lambda Calculus
- CCG
- Supervised Learning, Weakly-supervised Learning

Move to the chair
Class Goals

• Learn formalisms and algorithms
• Be able to read current research papers
• Understand what it takes to build a system
• Develop a critical view of NLP research
• See where there is room for more research!
Class Requirements

• Assignment:
  – First third of course
  – Multiple parts: annotation, implementing models

• Final project:
  – Option 1: ACL-level research project
  – Option 2: Re-implement a paper
  – Will include multiple deliverables and short presentation
Course Details

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

• Other material on the course website

• Grading: assignment (20%), participation (10%), class project (70%)

• Prerequisites:
  – CS 2110 or equivalent programming experience
  – Machine learning course
  – Master students – with instructor permission
Online

- Course website: http://www.cs.cornell.edu/Courses/cs6741/2015fa/
- Discussion forum: https://piazza.com/cornell/fall2015/cs6741/home
- Contact (but prefer forum): yoav@cs.cornell.edu
Course Policies

- **External code:** Final project – within limits, following original paper. For assignment – no.
- **Late policy:** 2 late days. Not accumulated in groups.
- **Collaboration:** Will be announced with first assignment.
- Electronics in class – OK.
- Please come to class on time!
Questions for Class

• Office hours
  – When? Hangout? Beam?

• Why are you here?
  – Interests?
  – Topics not listed you would like to hear about?