Natural Language Processing

Taylor Berg-Kirkpatrick – CMU
Slides: Dan Klein – UC Berkeley
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.
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
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
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
```

- $\text{ROOT} \rightarrow \text{S}$
- $\text{S} \rightarrow \text{NP VP}$
- $\text{NP} \rightarrow \text{PRP}$
- $\text{VP} \rightarrow \text{VBD ADJP}$
Corpus-Based Methods

- It gives us statistical information

<table>
<thead>
<tr>
<th></th>
<th>All NPs</th>
<th>NPs under S</th>
<th>NPs under VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP PP</td>
<td>11%</td>
<td>9%</td>
<td>23%</td>
</tr>
<tr>
<td>DT NN</td>
<td>9%</td>
<td>9%</td>
<td>7%</td>
</tr>
<tr>
<td>PRP</td>
<td>6%</td>
<td>21%</td>
<td>4%</td>
</tr>
</tbody>
</table>
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

\[ I \text{ haven’t slept for ten days } \]

\[ \text{John’s boss said he was doing better} \]

- 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**
  
  哲学家維特根斯坦出生于維也納

  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 (chē-ēn’ , -ēn’) 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
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 anew the work of remaking America."

"This is not the work of one person, or one party, but all of us as a people," Obama said.

"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)
"Il est impossible aux journalistes de rentrer dans les régions tibétaines."

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'ilégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes
Portfolio | Reportage | Vidéo

Accord sur la TVA : "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

Agreement on the VAT: "Sarkozy wins case at the worst possible time"
Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.

That would be an interim solution which would make it possible to work towards a binding charter in the long term.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.</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>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1x DATA</th>
<th>10x DATA</th>
<th>100x DATA</th>
<th>1000x DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[this] [constituerait] [assistance] [transitoire] [who] [permettrait]</td>
<td>[it] [would] [a solution] [transitional] [which] [would] [of] [lead]</td>
<td>[this] [would be] [a transitional solution] [which would] [lead to] [a</td>
<td>[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]</td>
</tr>
<tr>
<td>[licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante]</td>
<td>[to] [term] [to a] [charter] [to] [value] [binding] [.]</td>
<td>[a charter] [legally binding] [.]</td>
<td></td>
</tr>
<tr>
<td>[.].</td>
<td>[.].</td>
<td>[.].</td>
<td>[.].</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...
Data and Knowledge

- Classic knowledge representation worry: How will a machine ever know that...
  - Ice is frozen water?
  - Beige looks like this: 
  - Chairs are solid?

- Answers:
  - 1980: write it all down
  - 2000: get by without it
  - 2020: learn it from data
Q: Who signed the Serve America Act?
A: Barack Obama

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, 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**

**Type**

Subsidiary of Time Warner

**Founded**

1983 as Quantum Computer Services
Grounded Language
… 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
Example: Language change, reconstructing ancient forms, phylogenies
... just one example of the kinds of linguistic models we can build
What is NLP research?

- Three aspects we often investigate:
  - **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)
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 → 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
The (Effective) NLP Cycle

- Pick a problem (usually some disambiguation)
- Get a lot of data (usually a labeled corpus)
- Build the simplest thing that could possibly work

**Repeat:**
- Examine the most common errors are
- Figure out what information a human might use to avoid them
- Modify the system to exploit that information
  - Feature engineering
  - Representation redesign
  - Different machine learning methods

- We’re do this over and over again