What is NLP?

- Fundamental goal: analyze and process human language, broadly, robustly, accurately...
- End systems that we want to build:
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering...
  - Modest: spelling correction, text categorization...
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)

---

Information Retrieval

- **General problem:**
  - Given information needs, produce information
  - Includes, e.g. web search, question answering, and classic IR

- **Common case: web search**

$q = “Apple Computers”$
Feature-Based Ranking

\[ q = \text{“Apple Computers”} \]

\[ f_i(\text{query}) = [0.3 \ 5 \ 0 \ 0 \ \ldots] \]

\[ f_i(\text{candidate}) = [0.8 \ 4 \ 2 \ 1 \ \ldots] \]

Learning to Rank

- **Setup**
  - \( x_i \): query
  - \( \mathcal{Y} = \{y\} \): candidates
  - \( f_i(y) \): features of candidate \( y \) for query \( x_i \)
  - \( \ell_i(y, y') \): cost of ranking \( y' \prec y \) for query \( x_i \)

- **Optimize**, e.g.:

  \[
  \min_w \frac{1}{2} \|w\|^2 \\
  \forall i, (y \prec y') \quad w^T f_i(y) \geq w^T f_i(y') + \ell_i(y, y')
  \]

  ... lots of variants are possible!
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 70% accuracy for multi-sentence templates, 90%+ for single easy fields

Document Understanding?

- **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
Problem: Ambiguities

- **Headlines:**
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor’s Desk
  - Iraqi Head Seeks Arms
  - Local HS Dropouts Cut in Half
  - Juvenile Court to Try Shooting Defendant
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks

- Why are these funny?

Syntactic Analysis

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.
PCFGs

- Natural language grammars are very ambiguous!
- PCFGs are a formal probabilistic model of trees
  - Each “rule” has a conditional probability (like an HMM)
  - Tree’s probability is the product of all rules used
- Parsing: Given a sentence, find the best tree – a search problem!

```
ROOT
   ↓
  S
   ↓
NP VP
   ↓
PRP VBD ADJP
   ↓
He walks right
```

```
ROOT → S       375/420
S → NP VP .     320/392
NP → PRP        127/539
VBP ADJP        32/401
```

Summarization

- Condensing documents
  - Single or multiple
  - Extractive or synthetic
  - Aggregative or representative
  - Even just shortening sentences
- Very context-dependent!
- An example of analysis with generation
Machine Translation

Original Text

November 16, 16 (Xinhua News Agency) November 15 in the city of Chengdu, Sichuan Province "heating day," local windy, rainy days, the minimum temperature dropped to 1 C. 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.

- SOTA: much better than nothing, but more an understanding aid than a replacement for human translators
- New, better methods

Translated Text

Shijiazhuang, November 16 (Xinhua News Agency) November 15 in the city of Chengdu, Sichuan Province "heating day," local windy, rainy days, the minimum temperature dropped to 1 C. 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.

- SOTA: much better than nothing, but more an understanding aid than a replacement for human translators
- New, better methods

Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

Yo lo haré mañana
I will do it tomorrow

Hasta pronto
See you soon

Hasta pronto
See you around

Machine translation system:

Yo lo haré pronto
I will do it sooner

Model of translation

Yo lo haré pronto
I will do it sooner

I will do it soon

I will do it around

See you tomorrow
Levels of Transfer

MT Overview
A Phrase-Based Decoder

- Probabilities at each step include LM and TM

Search for MT

- Probabilities at each step include LM and TM
Etc: Historical Change

- Change in form over time, reconstruct ancient forms, phylogenies
- … just an example of the many other kinds of models we can build

Want to Know More?

- Check out the Berkeley NLP Group:

  nlp.cs.berkeley.edu