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

Feature-Based Ranking

$q = \text{"Apple Computers"}$

$f_i(x) = [0.3 \ 5 \ 0 \ 0 \ ...]$

Learning to Rank

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

- Optimize, e.g.:
  \[
  \min_w \frac{1}{2} \|w\|^2 \quad \forall i \ (y < y') \quad w^T f_i(y) \geq w^T f_i(y') + \ell_i(y, y')
  \]
  ... lots of variants are possible!
Information Extraction

- Unstructured text to database entries
- SOTA: perhaps 70% accuracy for multi-sentence templates, 90%+ for single easy fields

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

Problem: Ambiguities

- 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?

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                         375/420
S -> NP VP                       320/392
NP -> PRP                       127/539
VP -> VBD ADJP                  32/401
```

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

Syntactic Analysis [demo]

- 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.

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

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:

<table>
<thead>
<tr>
<th>Source Sentence</th>
<th>Target Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yo lo haré mañana</td>
<td>Hasta pronto</td>
</tr>
<tr>
<td>Hasta pronto</td>
<td>I will do it tomorrow</td>
</tr>
<tr>
<td>See you soon</td>
<td>I will do it around</td>
</tr>
<tr>
<td>See you around</td>
<td>See you tomorrow</td>
</tr>
</tbody>
</table>

Levels of Transfer

A Phrase-Based Decoder

<table>
<thead>
<tr>
<th>Matrix</th>
<th>so</th>
<th>do</th>
<th>the</th>
<th>today</th>
<th>I will</th>
</tr>
</thead>
<tbody>
<tr>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
</tr>
<tr>
<td>do</td>
<td>did</td>
<td>did</td>
<td>did</td>
<td>did</td>
<td>did</td>
</tr>
<tr>
<td>target</td>
<td>source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Probabilities at each step include LM and TM

MT Overview

Statistical Analysis

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation model (TM)</th>
<th>Broken English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>competency</td>
<td></td>
</tr>
</tbody>
</table>

Search for MT

<table>
<thead>
<tr>
<th>Matrix</th>
<th>so</th>
<th>do</th>
<th>the</th>
<th>today</th>
<th>I will</th>
</tr>
</thead>
<tbody>
<tr>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
<td>not</td>
</tr>
<tr>
<td>do</td>
<td>did</td>
<td>did</td>
<td>did</td>
<td>did</td>
<td>did</td>
</tr>
<tr>
<td>target</td>
<td>source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P=0.092 did not give

Mary p=0.534 did not give

P=0.092 give
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