Speech and Language

Speech and Language

Speech technologies
- Automatic speech recognition (ASR)
- Text-to-speech synthesis (TTS)
- Dialog systems

Language processing technologies
- Machine translation
- Information extraction
- Web search, question answering
- Text classification, spam filtering, etc...

Digitizing Speech

Speech input is an acoustic wave form

```
<table>
<thead>
<tr>
<th>s</th>
<th>p</th>
<th>ee</th>
<th>ch</th>
<th>l</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Speech in an Hour

```
Adding 100 Hz + 1000 Hz Waves
```

Spectral Analysis

```
<table>
<thead>
<tr>
<th>Frequency gives pitch; amplitude gives volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)</td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Fourier transform of wave displayed as a spectrogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>darkness indicates energy at each frequency</td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Amplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
```

```
| 0 | 0 | 0 | 0 |
```

```
<table>
<thead>
<tr>
<th>db</th>
<th>0.05s</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```
Spectrum

Frequency components (100 and 1000 Hz) on x-axis

**Part of [ae] from “lab”**

- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

**Back to Spectra**

- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.

- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

**Acoustic Feature Sequence**

- Time slices are translated into acoustic feature vectors (~39 real numbers per slice)

- These are the observations, now we need the hidden states X

**State Space**

- P(E|X) encodes which acoustic vectors are appropriate for each phoneme (each kind of sound)
- P(X|X') encodes how sounds can be strung together
- We will have one state for each sound in each word
- From some state x, can only:
  - Stay in the same state (e.g. speaking slowly)
  - Move to the next position in the word
  - At the end of the word, move to the start of the next word
- We build a little state graph for each word and chain them together to form our state space X

**HMMs for Speech**

- Word Model
- Observation Sequence (spectral feature vectors)
Decoding

- While there are some practical issues, finding the words given the acoustics is an HMM inference problem.

- We want to know which state sequence $x_{1:T}$ is most likely given the evidence $e_{1:T}$:

  $$x_{1:T}^* = \arg \max_{x_{1:T}} P(x_{1:T} | e_{1:T})$$

  $$= \arg \max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

- From the sequence $x$, we can simply read off the words.

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...

Problem: Ambiguities

- Headlines:
  - Enraged Cow Injures Farmer With Ax
  - 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?

Grammar: 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 – search!

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow NP \ VP$</td>
<td>375/420</td>
</tr>
<tr>
<td>$S \rightarrow PRP$</td>
<td>320/392</td>
</tr>
<tr>
<td>$NP \rightarrow PRP$</td>
<td>127/539</td>
</tr>
<tr>
<td>$VP \rightarrow VBD ADJP$</td>
<td>32/401</td>
</tr>
</tbody>
</table>

Transitions with Bigrams

While there are some practical issues, finding the words given the acoustics is an HMM inference problem.

We want to know which state sequence $x_{1:T}$ is most likely given the evidence $e_{1:T}$:

$$x_{1:T}^* = \arg \max_{x_{1:T}} P(x_{1:T} | e_{1:T})$$

$$= \arg \max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

From the sequence $x$, we can simply read off the words.

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...

Problem: Ambiguities

- Headlines:
  - Enraged Cow Injures Farmer With Ax
  - 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?

Grammar: 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 – search!

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow NP \ VP$</td>
<td>375/420</td>
</tr>
<tr>
<td>$S \rightarrow PRP$</td>
<td>320/392</td>
</tr>
<tr>
<td>$NP \rightarrow PRP$</td>
<td>127/539</td>
</tr>
<tr>
<td>$VP \rightarrow VBD ADJP$</td>
<td>32/401</td>
</tr>
</tbody>
</table>
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.

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]

The Problem with Dictionary Look-ups

- /top/roof/
- /summit/peak/top/apex/
- /coming directly towards one/top/end/
- /tidal/top/carry/canopy/build/Gal/
- /surpass/top/
- /extremely/pole/utmost/top/collect/receive/
- /peak/top/
- /level/side/surface/aspect/top/flower/
- /shoaring /top/patching/

A Brief and Biased History

MT is the “first” non-numeral compute task
ALPAC report deemed MT bad
Statistical data-driven approach introduced

Data-Driven Machine Translation

Target language corpus:
I will get to it soon
See you later
He will do it

Sentence-aligned parallel corpus:
Yo lo hare mañana
I will do it tomorrow
Hasta pronto
See you soon
Hasta pronto
See you around

Machine translation system:
Model of translation
Yo lo hare pronto
New Sentence
I will do it soon

Learning to Translate

CLASSIC SOUPS

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>57.</td>
<td>House Chicken Soup (Chicken, Celery, Potatoes, Onions, Carrots)</td>
<td>1.50</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>58.</td>
<td>Chicken Rice Soup</td>
<td>1.85</td>
<td>3.25</td>
<td></td>
</tr>
<tr>
<td>59.</td>
<td>Chicken Noodle Soup</td>
<td>1.85</td>
<td>3.25</td>
<td></td>
</tr>
<tr>
<td>60.</td>
<td>Cotton(ene) Chicken Soup</td>
<td>1.85</td>
<td>3.25</td>
<td></td>
</tr>
<tr>
<td>61.</td>
<td>Tomato, Corn, Egg Drop Soup</td>
<td>1.65</td>
<td>2.95</td>
<td></td>
</tr>
<tr>
<td>62.</td>
<td>Blackened Chicken Soup</td>
<td>1.10</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>63.</td>
<td>Hot &amp; Sour Soup</td>
<td>1.10</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>64.</td>
<td>Egg Drop Soup</td>
<td>1.10</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>65.</td>
<td>Egg Drop &amp; Vegetable Mix</td>
<td>1.10</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>66.</td>
<td>turtle Vegetable Soup</td>
<td>NA</td>
<td>3.50</td>
<td></td>
</tr>
<tr>
<td>67.</td>
<td>Chicken Corn Cream Soup</td>
<td>NA</td>
<td>3.50</td>
<td></td>
</tr>
<tr>
<td>68.</td>
<td>Crab Meat Corn Cream Soup</td>
<td>NA</td>
<td>3.50</td>
<td></td>
</tr>
<tr>
<td>69.</td>
<td>Seafood Soup</td>
<td>NA</td>
<td>3.50</td>
<td></td>
</tr>
</tbody>
</table>

Example from Adam Lopez
The HMM Model

E: Thank you, I shall do so gladly.
A: Gracias, lo haré de muy buen grado.

Model Parameters
Emissions: P(e_i | a_j) = Gracias | Thank you
Transitions: P(a_i | a_j) = 3 | A_i = 0

Levels of Transfer

Machine Translation

Machine translation system:
Yo lo haré de muy buen grado
Después de veras

Model of translation
Yo lo haré después
I will do it later

A Statistical Translation Model

Synchronous Derivation

S
Yo lo haré después
I will do it later

A Statistical Model

Translation model components
factor over applied rules

How well are these rules supported by the data?

Language model factors over n-grams

How well is this output sentence supported by the data?

Synchronous Grammar Rules
S → (Yo lo haré ADV; I will do it ADV)
ADV → (después; later)

Example Syntax-Based Translation

[Demo: MT]