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?

Parsing as Search

```
S     S
|     |
NP    NP
|     |
N     N
|     |
Hershey bars N
|     |
protest
```

Hershey bars protest
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!

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

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

<table>
<thead>
<tr>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>顶部</td>
<td>top/roof/</td>
</tr>
<tr>
<td>顶端</td>
<td>summit/peak/top/apex/</td>
</tr>
<tr>
<td>顶头</td>
<td>coming directly towards one/top/end/</td>
</tr>
<tr>
<td>盖</td>
<td>lid/top/cover/canopy/build/Gai/</td>
</tr>
<tr>
<td>盖帽</td>
<td>surpass/top/</td>
</tr>
<tr>
<td>极</td>
<td>extremely/pole/utmost/top/collect/receive/</td>
</tr>
<tr>
<td>尖峰</td>
<td>peak/top/</td>
</tr>
<tr>
<td>面</td>
<td>fade/side/surface/aspect/top/face/flour/</td>
</tr>
<tr>
<td>摘心</td>
<td>/top/topping/</td>
</tr>
</tbody>
</table>

Example from Douglas Hofstadter
A Brief and Biased History

When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode." — Warren Weaver

"Machine Translation" presumably means going by algorithm from machine-readable source text to useful target text... In this context, there has been no machine translation... — John Pierce

Berkeley's first MT grant

MT is the "first" non-numeral compute task

ALPAC report deems MT bad

Statistical MT thrives

Statistical data-driven approach introduced

'47  '58  '66  '90's  '00's

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 haré mañana

I will do it tomorrow

Hasta pronto

See you soon

Hasta pronto

See you around

Machine translation system:

Yo lo haré pronto

Novel sentence

Model of translation

I will do it soon
Learning to Translate

CLASSIC SOUPS

<table>
<thead>
<tr>
<th></th>
<th>Sm.</th>
<th>Lg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.</td>
<td>House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)</td>
<td>1.50</td>
</tr>
<tr>
<td>58.</td>
<td>Chicken Rice Soup</td>
<td>1.85</td>
</tr>
<tr>
<td>59.</td>
<td>Chicken Noodle Soup</td>
<td>1.85</td>
</tr>
<tr>
<td>60.</td>
<td>Cantonese (Wonton) Soup</td>
<td>1.50</td>
</tr>
<tr>
<td>61.</td>
<td>Tomato Clear Egg Drop Soup</td>
<td>1.65</td>
</tr>
<tr>
<td>62.</td>
<td>Regular (Wonton) Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>63.</td>
<td>Hot &amp; Sour Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>64.</td>
<td>Egg Drop Soup</td>
<td>1.10</td>
</tr>
<tr>
<td>65.</td>
<td>Egg Drop (Wonton) Mix</td>
<td>1.10</td>
</tr>
<tr>
<td>66.</td>
<td>Tofu Vegetable Soup</td>
<td>NA</td>
</tr>
<tr>
<td>67.</td>
<td>Chicken Corn Cream Soup</td>
<td>NA</td>
</tr>
<tr>
<td>68.</td>
<td>Crab Meat Corn Cream Soup</td>
<td>NA</td>
</tr>
<tr>
<td>69.</td>
<td>Seafood Soup</td>
<td>NA</td>
</tr>
</tbody>
</table>

Example from Adam Lopez

The HMM Model

E: Thank you, I shall do so gladly.

A: 1 3 7 6 8 8 8 9 0

F: Gracias, lo haré de muy buen grado.

Model Parameters

Emissions: P(F1 = Gracias | E1 = Thank )
Transitions: P(A2 = 3 | A1 = 1)
Levels of Transfer

Machine Translation

Machine translation system:
A Statistical Translation Model

**Synchronous Derivation**

\[
S ightarrow \text{Yo lo haré después}
\]

\[
S ightarrow \text{ADV}
\]

**Synchronous Grammar Rules**

\[ S \rightarrow \langle \text{Yo lo haré ADV ; I will do it ADV} \rangle \]

\[ ADV \rightarrow \langle \text{después ; later} \rangle \]

**A Statistical Model**

Translation model components

factor over applied rules

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?

Example Syntax-Based Translation

foreign:... وزوج قبل الابناء بائى تصرحات قور وسواه الى المقامة...

french:... Il veut à la fois... la plupart du temps...

Example Syntax-Based Translation

[demo: MT]