CS 188: Artificial Intelligence
Spring 2009

Natural Language Processing

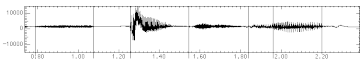
Dan Klein – UC Berkeley

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



"Speech Lab"

- 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 = "Apple Computers"

$\mathbf{f}_i($ $) = [0.3\ 5\ 0\ 0\ \ldots]$

$\mathbf{f}_i($ $) = [0.8\ 4\ 2\ 1\ \ldots]$

Learning to Rank

- Setup

$\mathbf{x}_i$	query
$\mathcal{Y} = \{\mathbf{y}\}$	candidates
$\mathbf{f}_i(\mathbf{y})$	features of candidate $\mathbf{y}$ for query $\mathbf{x}_i$
$\ell_i(\mathbf{y}, \mathbf{y}')$	cost of ranking $\mathbf{y}' \prec \mathbf{y}$ for query $\mathbf{x}_i$

- Optimize, e.g.:

$$\min_{\mathbf{w}} \quad \frac{1}{2}||\mathbf{w}||^2$$
$$\forall i, (\mathbf{y} \prec \mathbf{y}') \quad \mathbf{w}^\top \mathbf{f}_i(\mathbf{y}) \geq \mathbf{w}^\top \mathbf{f}_i(\mathbf{y}') + \ell_i(\mathbf{y}, \mathbf{y}')$$

… 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.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 70% accuracy for multi-sentence temples, 90%+ for single easy fields

Document Understanding?

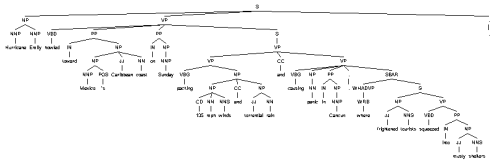
- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the 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

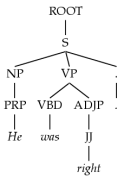
[demo]



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 .

PCFGs

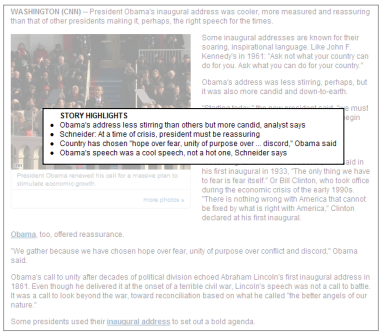
- Natural language grammars are very ambiguous!
- PCFGs are a formal probabilistic model of trees
  - Each "rule" has a conditional probability (like an HMM)
  - Tree's probability is the product of all rules used
- Parsing: Given a sentence, find the best tree – a search problem!



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

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

[demo]

Original Text

新华网石家庄１１月１６日电（记者　张涛）１１月１５日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至１℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 ℃. 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

13

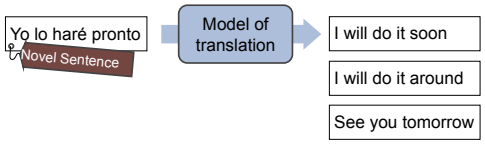
Corpus-Based MT

Modeling correspondences between languages

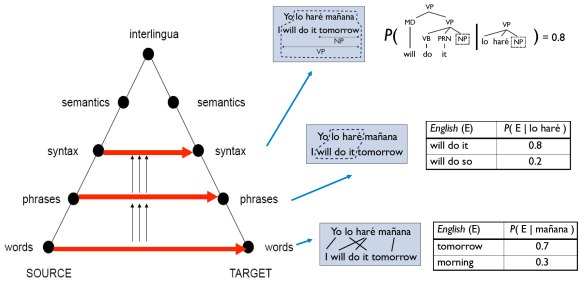
Sentence-aligned parallel corpus:

Yo lo haré mañana I will do it tomorrow	Hasta pronto See you soon	Hasta pronto See you around

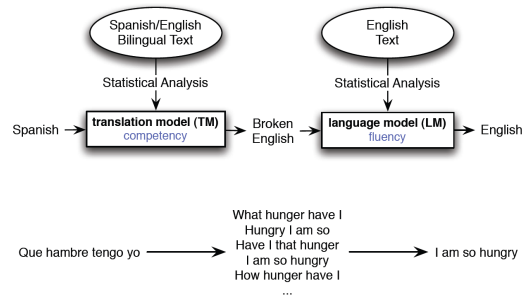
Machine translation system:



Levels of Transfer



MT Overview



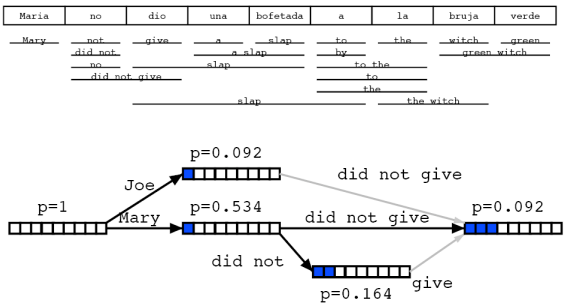
16

A Phrase-Based Decoder

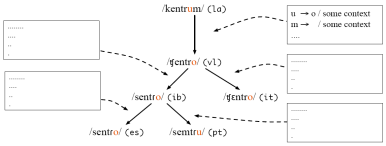
- Probabilities at each step include LM and TM

17

Search for MT



Etc: Historical Change



Gloss	Latin	Italian	Spanish	Portuguese
Word/verb	verbum	verbo	verbo	verbu
Center	centrum	centro	centro	centro

- 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