

CS 188: Artificial Intelligence

Language

Pieter Abbeel – UC Berkeley
 Slides from Dan Klein

What is NLP?

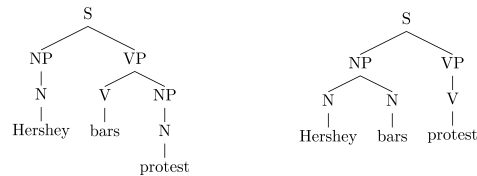


- Fundamental goal: analyze and process human language, broadly, robustly, accurately...
- End systems that we want to build:
 - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering...
 - Modest: spelling correction, text categorization...

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 Defiant
 - Stolen Painting Found by Tree
 - Kids Make Nutritious Snacks
- Why are these funny?

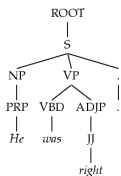
Parsing as Search



Hershey bars protest

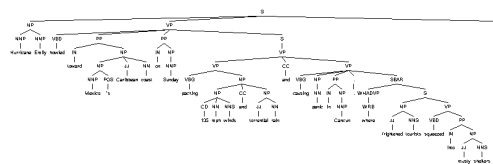
Grammar: PCFGs

- 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

"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Brune Bruze, correspondant de "World" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "libèrent ses airs Tibétain".

Les nets Le Dalai Lama dénonce l' "indes" imposé en Tibet depuis sa fuite, en 1959.

Voici, Annuaire de la rébellion

"It is impossible for journalists to enter Tibetan areas"

Brune Bruze, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "leave not Tibet".

From: The Dalai Lama denounces the "indes" imposed since he fled Tibet in 1959.

Voici: Anniversary of the Tibetan rebellion: China vs. grant.


- 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

顶部	/top/roof/
顶端	/summit/peak/top/apex/
顶头	/coming directly towards one/top/end/
盖	/lid/top/cover/canopy/build/Gai/
盖帽	/surpass/top/
极	/extremely/pole/utmost/top/collect/receive/
尖峰	/peak/top/
面	/fade/side/surface/aspect/top/face/flour/
摘心	/top/topping/


Example from Douglas Hofstadter

A Brief and Biased History



Warren Weaver

"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.'"



John Pierce

"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...

Berkeley's first MT grant

MT is the "first" non-numeral compute task

ALPAC report deems MT bad

Statistical data-driven approach introduced

Statistical MT thrives

'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

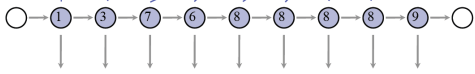
		Sm.	Lg.
鸡 葱 麵 汤	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50 2.75
雞 飯 湯	58.	Chicken Rice Soup	1.85 3.25
雞 麵 湯	59.	Chicken Noodle Soup	1.85 3.25
廣東 (卷) 吞	60.	Cantonese (Wonton) Soup	1.50 2.75
薑 蒜 湯	61.	Tomato Clear Egg Drop Soup	1.65 2.95
(卷) 吞 湯	62.	Regular (Wonton) Soup	1.10 2.10
酸 辣 湯	63.	Hot & Sour Soup	1.10 2.10
雲 花 湯	64.	Egg Drop Soup	1.10 2.10
(卷) 吞 湯	65.	Egg Drop (Wonton) Mix	1.10 2.10
豆 腐 菜 湯	66.	Tofu Vegetable Soup	NA 3.50
雞 玉 米 湯	67.	Chicken Corn Cream Soup	NA 3.50
蟹 肉 玉 米 湯	68.	Crab Meat Corn Cream Soup	NA 3.50
海 鮮 湯	69.	Seafood Soup	NA 3.50

Example from Adam Lopez

The HMM Model

E: Thank you , I shall do so gladly .

1 2 3 4 5 6 7 8 9

A: 

F: Gracias , lo haré de muy buen grado .

Model Parameters

Emissions: $P(F_1 = \text{Gracias} \mid E_{A_1} = \text{Thank})$ Transitions: $P(A_2 = 3 \mid A_1 = 1)$

