CS344: Introduction to Artificial Intelligence

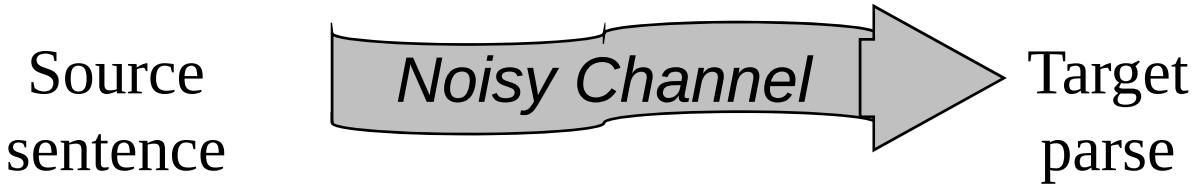
Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay

Lecture 26-27: Probabilistic Parsing

Example of Sentence labeling: Parsing

$[_{S1}[_{S}[_{S}[_{VP}[_{VB}$Come$][_{NP}[_{NNP}$July$]]]]]$
$[_{,}$,$]$
$[_{CC}$ and$]$
$[_{S}$ $[_{NP}$ $[_{DT}$ the$]$ $[_{JJ}$ IIT$]$ $[_{NN}$ campus$]]$
$[_{VP}$ $[_{AUX}$ is$]$
$[_{ADJP}$ $[_{JJ}$ abuzz$]$
$[_{PP}[_{IN}$ with$]$
$[_{NP}[_{ADJP}$ $[_{JJ}$ new$]$ $[_{CC}$ and$]$ $[$ $_{VBG}$ returning$]]$
$[_{NNS}$ students$]]]]]]$
$[_{.}$.$]]]$

Noisy Channel Modeling



$$T^* = \underset{T}{\operatorname{argmax}}\,[P(T|S)]$$

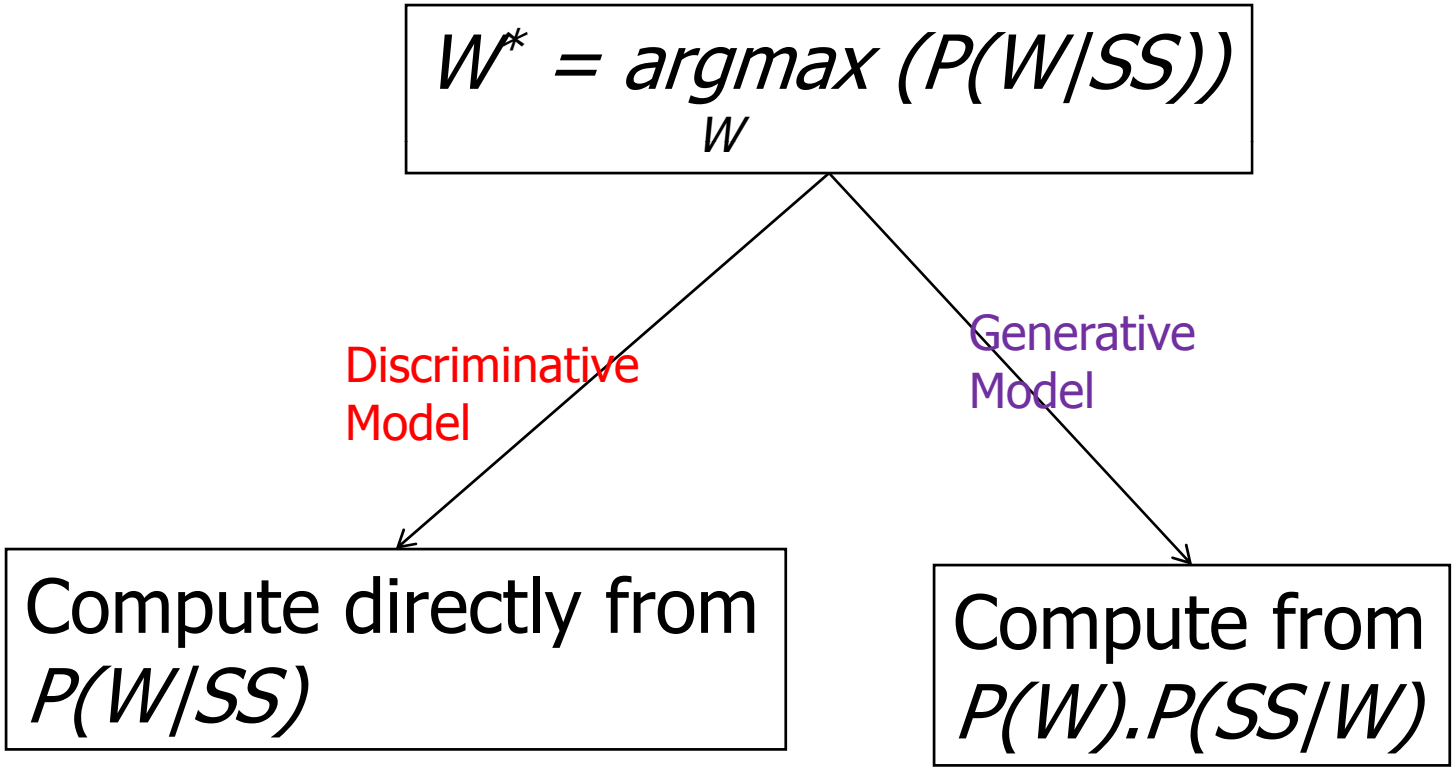
$$= \underset{T}{\operatorname{argmax}}\,[P(T).P(S|T)]$$

$$= \underset{T}{\operatorname{argmax}}\,[P(T)]$$, since given the parse the sentence is completely determined and $P(S|T)=1$

Corpus

- A collection of text called *corpus*, is used for collecting various language data
- With annotation: more information, but manual labor intensive
- Practice: *label automatically; correct manually*
- The famous *Brown Corpus* contains 1 million tagged words.
- **Switchboard:** very famous corpora 2400 conversations, 543 speakers, many US dialects, annotated with orthography and phonetics

Discriminative vs. Generative Model



Notion of Language Models

Language Models

- N-grams: sequence of n consecutive words/chracters

- Probabilistic / Stochastic Context Free Grammars:
  - Simple probabilistic models capable of handling recursion
  - A CFG with probabilities attached to rules
  - Rule probabilities → how likely is it that a particular rewrite rule is used?

PCFGs

- Why PCFGs?
  - Intuitive probabilistic models for tree-structured languages
  - Algorithms are extensions of HMM algorithms
  - Better than the n-gram model for language modeling.

Formal Definition of PCFG

- A PCFG consists of
  - A set of terminals $\{w_k\}$, k = 1,….,V

    $\{w_k\}$ = { child, teddy, bear, played…}
  - A set of non-terminals $\{N^i\}$, i = 1,…,n

    $\{N_i\}$ = { NP, VP, DT…}
  - A designated start symbol N^1
  - A set of rules $\{N^i \rightarrow \zeta^j\}$, where ζ^j is a sequence of terminals & non-terminals

    NP $\rightarrow$ DT NN
  - A corresponding set of rule probabilities

Rule Probabilities

- Rule probabilities are such that

$$\forall i \ \sum_{\mathrm{i}} \mathrm{P}(\mathrm{N}^i \rightarrow \zeta^j) = 1$$

E.g., P(NP $\rightarrow$ DT NN) = 0.2

P(NP $\rightarrow$ NN) = 0.5

P(NP $\rightarrow$ NP PP) = 0.3

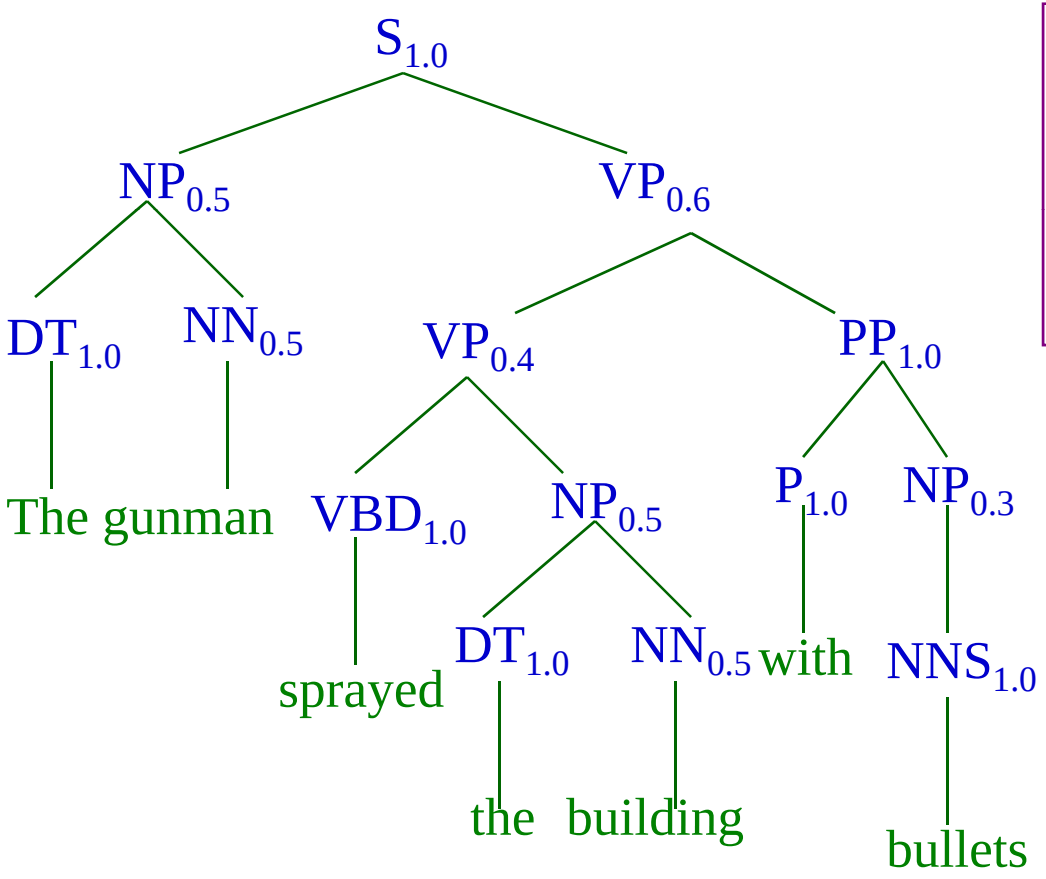
- P(NP $\rightarrow$ DT NN) = 0.2
  - Means 20 % of the training data parses use the rule NP $\rightarrow$ DT NN

Probabilistic Context Free Grammars

- S $\rightarrow$ NP VP 1.0
- NP $\rightarrow$ DT NN 0.5
- NP $\rightarrow$ NNS 0.3
- NP $\rightarrow$ NP PP 0.2
- PP $\rightarrow$ P NP 1.0
- VP $\rightarrow$ VP PP 0.6
- VP $\rightarrow$ VBD NP 0.4
- DT $\rightarrow$ the 1.0
- NN $\rightarrow$ gunman 0.5
- NN $\rightarrow$ building 0.5
- VBD $\rightarrow$ sprayed 1.0
- NNS $\rightarrow$ bullets 1.0

Example Parse t_1`

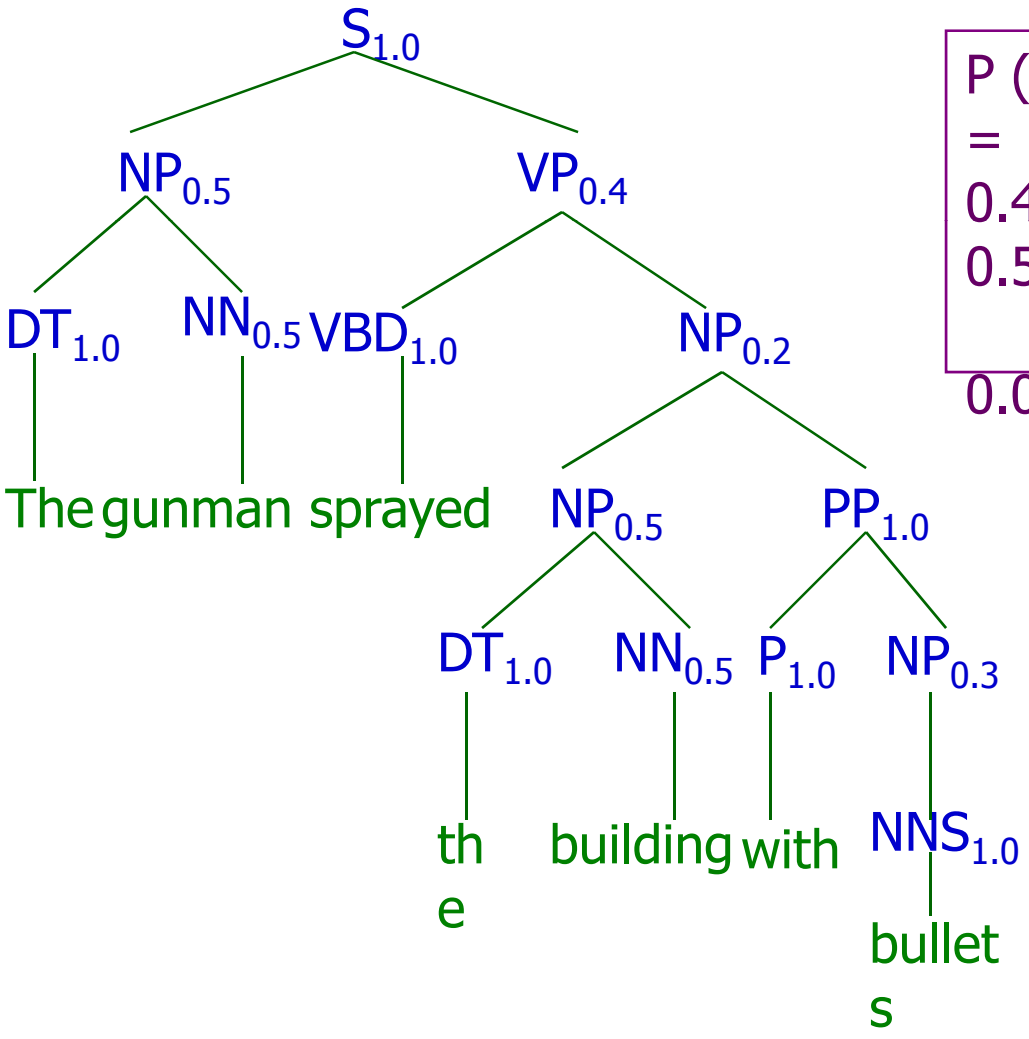
- The gunman sprayed the building with bullets.



$P(t_1) = 1.0 * 0.5 * 1.0 * 0.5 * 0.6 * 0.4 * 1.0 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0 =$
0.00225

Another Parse t_2

- The gunman sprayed the building with bullets.



$P(t_2)$
$= 1.0 * 0.5 * 1.0 * 0.5 * 0.4 * 1.0 * 0.2 * 0.5 * 1.0 * 0.5 * 1.0 * 1.0 * 0.3 * 1.0$
$= 0.0015$

Is NLP Really Needed

Post-1

- POST----5 TITLE: "Wants to invest in IPO? Think again" |

Hereâ€™s a sobering thought for those who believe in investing in IPOs. Listing gains â€” the return on the IPO scrip at the close of listing day over the allotment price â€” have been falling substantially in the past two years. Average listing gains have fallen from 38% in 2005 to as low as 2% in the first half of 2007.Of the 159 book-built initial public offerings (IPOs) in India between 2000 and 2007, two-thirds saw listing gains. However, these gains have eroded sharply in recent years.Experts say this trend can be attributed to the aggressive pricing strategy that investment bankers adopt before an IPO. â€œWhile the drop in average listing gains is not a good sign, it could be due to the fact that IPO issue managers are getting aggressive with pricing of the issues,â€ says Anand Rathi, chief economist, Sujan Hajra.While the listing gain was 38% in 2005 over 34 issues, it fell to 30% in 2006 over 61 issues and to 2% in 2007 till mid-April over 34 issues. The overall listing gain for 159 issues listed since 2000 has been 23%, according to an analysis by Anand Rathi Securities.Aggressive pricing means the scrip has often been priced at the high end of the pricing range, which would restrict the upward movement of the stock, leading to reduced listing gains for the investor. It also tends to suggest investors should not indiscriminately pump in money into IPOs.But some market experts point out that India fares better than other countries. â€œInternationally, there have been periods of negative returns and low positive returns in India should not be considered a bad thing.

Post-2

- POST----7TITLE: "[IIM-Jobs] ***** Bank: International Projects Group - Manager"|
Please send your CV & cover letter to anup.abraham@*****bank.com ***** Bank, through its International Banking Group (IBG), is expanding beyond the Indian market with an intent to become a significant player in the global marketplace. The exciting growth in the overseas markets is driven not only by India linked opportunities, but also by opportunities of impact that we see as a local player in these overseas markets and / or as a bank with global footprint. IBG comprises of Retail banking, Corporate banking & Treasury in 17 overseas markets we are present in. Technology is seen as key part of the business strategy, and critical to business innovation & capability scale up. The International Projects Group in IBG takes ownership of defining & delivering business critical IT projects, and directly impact business growth. Role: Manager Â– International Projects Group Purpose of the role: Define IT initiatives and manage IT projects to achieve business goals. The project domain will be retail, corporate & treasury. The incumbent will work with teams across functions (including internal technology teams & IT vendors for development/implementation) and locations to deliver significant & measurable impact to the business. Location: Mumbai (Short travel to overseas locations may be needed) Key Deliverables: Conceptualize IT initiatives, define business requirements

Sentiment Classification

- Positive, negative, neutral – 3 class
- Sports, economics, literature - multi class
- Create a representation for the document
- Classify the representation

The most popular way of representing a document is feature vector (indicator sequence).

Established Techniques

- Naïve Bayes Classifier (NBC)
- Support Vector Machines (SVM)
- Neural Networks
- K nearest neighbor classifier
- Latent Semantic Indexing
- Decision Tree ID3
- Concept based indexing

Successful Approaches

The following are successful approaches as reported in literature.

- NBC – simple to understand and implement
- SVM – complex, requires foundations of perceptions

Mathematical Setting

We have training set

A: Positive Sentiment Docs

B: Negative Sentiment Docs

Indicator/feature vectors to be formed

Let the class of positive and negative documents be C_+ and C_- , respectively.

Given a new document **D** label it positive if

$$P(C_+|D) > P(C_-|D)$$

Priori Probability

Document	Vector	Classification
D1	V1	+
D2	V2	-
D3	V3	+
..	..	..
D_{4000}	V_{4000}	-

Let T = Total no of documents
And let |+| = M
So,|-| = T-M

P(D being positive)=M/T

Priori probability is calculated without considering any features of the new document.

Apply Bayes Theorem

Steps followed for the NBC algorithm:

- Calculate Prior Probability of the classes. $P(C_+)$ and $P(C_-)$
- Calculate feature probabilities of new document. $P(D|C_+)$ and $P(D|C_-)$
- Probability of a document **D** belonging to a class **C** can be calculated by Baye's Theorem as follows:

$$P(C|D) = \frac{P(C) * P(D|C)}{P(D)}$$

- Document belongs to C_+ , if

$$P(C_+) * P(D|C_+) > P(C_-) * P(D|C_-)$$

Calculating $P(D|C_+)$

- Identify a set of features/indicators to represent a document and generate a feature vector (V_D). $V_D = \langle x_1 , x_2 , x_3 \ldots x_n \rangle$
- Hence, $P(D|C_+) = P(V_D|C_+)$

$$= P(\langle x_1 , x_2 , x_3 \ldots x_n \rangle \mid C_+)$$

$$= \frac{|\langle x1,x2,x3\ldots..xn\rangle, C_+|}{|C_+|}$$

- Based on the assumption that all features are Independently Identically Distributed (IID)

$$= P(\langle x_1 , x_2 , x_3 \ldots x_n \rangle \mid C_+)$$

$$= P(x_1 |C_+) * P(x_2 |C_+) * P(x_3 |C_+) *\ldots. P(x_n |C_+)$$

$$= \prod_{i=1}^{n} P(x_i |C_+)$$

Baseline Accuracy

- Just on Tokens as features, **80%** accuracy
- 20% probability of a document being misclassified
- On large sets this is significant

To improve accuracy…

- Clean corpora
- POS tag
- Concentrate on critical POS tags (e.g. *adjective*)
- Remove ‘objective’ sentences ('of' ones)
- Do aggregation

Use minimal to sophisticated NLP