CS626-449: Speech, NLP and the Web/Topics in AI

Pushpak Bhattacharyya
CSE Dept.,
IIT Bombay
Lecture-15: Probabilistic parsing; PCFG (contd.)
Example of Sentence labeling: Parsing

[S1[S[S[VP[VBCome][NP[NNPJuly]]]]]]
[,,]
[CC and]
[S [NP [DT the] [JJ UJF] [NN campus]]
[VP [AUX is]]
[ADJP [JJ abuzz]]
[PP[IN with]]
[NP[ADJP [JJ new] [CC and] [VBG returning]]
[NNS students]]]]]]
[. .]]]
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

\[ W^* = \arg\max_W (P(W|SS)) \]

- **Discriminative Model**
  - Compute directly from \( P(W|SS) \)

- **Generative Model**
  - Compute from \( P(W).P(SS|W) \)
Notion of Language Models
Language Models

• N-grams: sequence of n consecutive words/characters

• Probabilistic / Stochastic Context Free Grammars:
  ▪ Simple probabilistic models capable of handling recursion
  ▪ A CFG with probabilities attached to rules
  ▪ Rule probabilities \( \rightarrow \) 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 \( \{w_k\}, k = 1,\ldots,V \)
    \( \{w_k\} = \{ \text{child, teddy, bear, played} \ldots \} \)
  - A set of non-terminals \( \{N_i\}, i = 1,\ldots,n \)
    \( \{N_i\} = \{ \text{NP, VP, DT} \ldots \} \)
  - A designated start symbol \( N_1 \)
  - A set of rules \( \{N^i \rightarrow \zeta^j\} \), where \( \zeta^j \) is a sequence of terminals & non-terminals
    - \( \text{NP} \rightarrow \text{DT NN} \)
  - A corresponding set of rule probabilities
Rule Probabilities

- Rule probabilities are such that

\[ \forall i \sum_i P(N^i \rightarrow \zeta^j) = 1 \]

*E.g.,* \[ P( \text{NP} \rightarrow \text{DT NN}) = 0.2 \]
\[ P( \text{NP} \rightarrow \text{NN}) = 0.5 \]
\[ P( \text{NP} \rightarrow \text{NP PP}) = 0.3 \]

- \[ P( \text{NP} \rightarrow \text{DT NN}) = 0.2 \]
  - Means 20% of the training data parses use the rule \( \text{NP} \rightarrow \text{DT NN} \)
Probability of a sentence

- Notation:
  - $w_{ab}$ - subsequence $w_a \ldots w_b$
  - $N_j$ dominates $w_a \ldots w_b$
  - $\text{yield}(N_j) = w_a \ldots w_b$

- Probability of a sentence $= P(w_{1m})$

$$P(w_{1m}) = \sum_{t} P(w_{1m}, t)$$

- Where $t$ is a parse tree of the sentence

$$= \sum_{t} P(t)P(w_{1m} | t)$$

$$= \sum_{t: \text{yield}(t) = w_{1m}} P(t) \quad \because P(w_{1m} | t) = 1$$

If $t$ is a parse tree for the sentence $w_{1m}$, this will be 1!!