Inside-outside probabilities

Pawan Goyal

CSE, IIT Kharagpur

Week 5: Lecture 5

How to get the rule probabilities

Parsed Training Data

You can count!

$$\hat{P}(N^j \rightarrow \delta) = \frac{C(N^j \rightarrow \delta)}{\sum_\gamma C(N^j \rightarrow \gamma)}$$

How to get the rule probabilities

Parsed Training Data

You can count!

$$\hat{P}(N^j \rightarrow \delta) = \frac{C(N^j \rightarrow \delta)}{\sum_\gamma C(N^j \rightarrow \gamma)}$$

But what if the training data is not available?

i.e. gold standard parse is not known.

How to get the rule probabilities

Parsed Training Data

You can count!

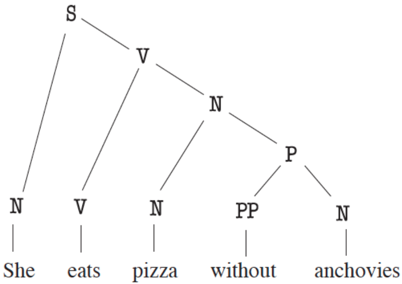
$$\hat{P}(N^j \rightarrow \delta) = \frac{C(N^j \rightarrow \delta)}{\sum_\gamma C(N^j \rightarrow \gamma)}$$

But what if the training data is not available?

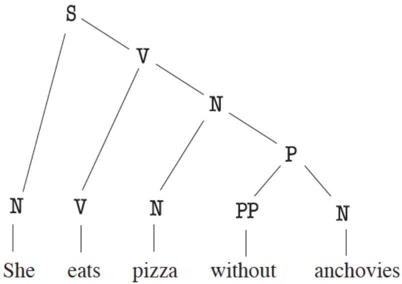
i.e. gold standard parse is not known.

- Underlying CFG is known and we are given a set of sentences
- For each sentence, we can find out all the possible parses
- *Maximize the likelihood of the sentences in the data under the PCFG constraints*

Example data



Example data



Rules of the form $A \rightarrow BC$

S → N V

V → V N

N → N P

P → PP N .

Rules of the form $A \rightarrow w$

N → She

V → eats

N → pizza

PP → without

N → anchovies.

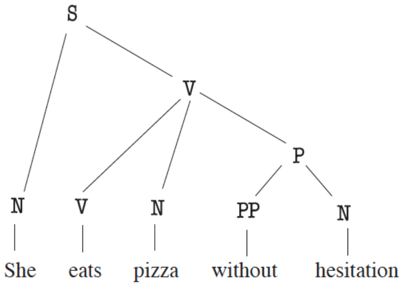
Example data

Is any other parse possible for *She eats pizza without anchovies* syntactically?

Example data

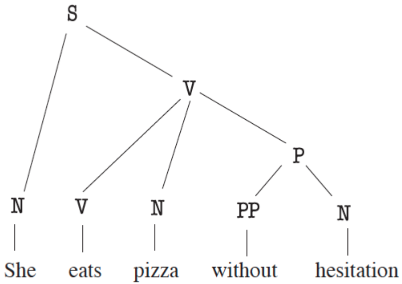
Is any other parse possible for *She eats pizza without anchovies* syntactically?

Consider *She eats pizza without hesitation*



Example data

Is any other parse possible for *She eats pizza without anchovies* syntactically?
Consider *She eats pizza without hesitation*



New Context-free rules:

$$V \rightarrow V\ N\ P$$

$$N \rightarrow \text{hesitation .}$$

Estimating the model parameters

We need to find probabilities such as

- $\phi(S \rightarrow N\ V)$
- $\phi(N \rightarrow pizza)$

Estimating the model parameters

We need to find probabilities such as

- $\phi(S \rightarrow N\ V)$
- $\phi(N \rightarrow pizza)$

Requirements

For each non-terminal A, the derivation probabilities sum up to 1

$$\sum_{\alpha} \phi(A \rightarrow \alpha) = 1$$

Estimating the model parameters

We need to find probabilities such as

- $\phi(S \rightarrow N\ V)$
- $\phi(N \rightarrow pizza)$

Requirements

For each non-terminal A, the derivation probabilities sum up to 1

$$\sum_{\alpha} \phi(A \rightarrow \alpha) = 1$$

For the example grammar:

$$\begin{aligned}
\phi(\mathtt{N} \rightarrow \mathtt{N\ P}) + \phi(\mathtt{N} \rightarrow \text{pizza}) + \phi(\mathtt{N} \rightarrow \text{anchovies}) &\quad + \\
+\ \phi(\mathtt{N} \rightarrow \text{hesitation}) + \phi(\mathtt{N} \rightarrow \text{She}) &= 1 \\
\phi(\mathtt{V} \rightarrow \mathtt{V\ N}) + \phi(\mathtt{V} \rightarrow \mathtt{V\ N\ P}) + \phi(\mathtt{V} \rightarrow \text{eats}) &= 1
\end{aligned}$$

$$\begin{aligned}
\phi(\mathtt{S} \rightarrow \mathtt{N\ V}) &= 1 \\
\phi(\mathtt{P} \rightarrow \mathtt{PP\ N}) &= 1 \\
\phi(\mathtt{PP} \rightarrow \text{without}) &= 1
\end{aligned}$$

Likelihood computation

W_1 = "She eats pizza without anchovies"

W_2 = "She eats pizza without hesitation".

Likelihood computation

$W_1 =$ "She eats pizza without anchovies"

$W_2 =$ "She eats pizza without hesitation".

$$
\begin{aligned}
P_\phi(W_1, T_1) \quad &= \quad \phi(\mathtt{S} \rightarrow \mathtt{N\ V})\, \phi(\mathtt{V} \rightarrow \mathtt{V\ N})\, \phi(\mathtt{N} \rightarrow \mathtt{N\ P}) \ \times \\
&\times \quad \phi(\mathtt{P} \rightarrow \mathtt{PP\ N})\, \phi(\mathtt{N} \rightarrow \text{She})\, \phi(\mathtt{V} \rightarrow \text{eats}) \ \times \\
&\times \quad \phi(\mathtt{N} \rightarrow \text{pizza})\, \phi(\mathtt{PP} \rightarrow \text{without})\, \phi(\mathtt{N} \rightarrow \text{anchovies})
\end{aligned}
$$

$$
\begin{aligned}
P_\phi(W_2, T_1) \quad &= \quad \phi(\mathtt{S} \rightarrow \mathtt{N\ V})\, \phi(\mathtt{V} \rightarrow \mathtt{V\ N\ P})\, \phi(\mathtt{P} \rightarrow \mathtt{P\ PP}) \ \times \\
&\times \quad \phi(\mathtt{N} \rightarrow \text{She})\, \phi(\mathtt{V} \rightarrow \text{eats})\, \phi(\mathtt{N} \rightarrow \text{pizza}) \times \\
&\times \quad \phi(\mathtt{PP} \rightarrow \text{without})\, \phi(\mathtt{N} \rightarrow \text{hesitation})
\end{aligned}
$$

Likelihood computation

$$\begin{aligned} P_\phi(W_1, T_2) &= \phi(\mathtt{S} \to \mathtt{N}\ \mathtt{V})\, \phi(\mathtt{V} \to \mathtt{V}\ \mathtt{N}\ \mathtt{P})\, \phi(\mathtt{P} \to \mathtt{P}\ \mathtt{PP}) \times \\ &\times \phi(\mathtt{N} \to \text{She})\, \phi(\mathtt{V} \to \text{eats})\, \phi(\mathtt{N} \to \text{pizza}) \times \\ &\times \phi(\mathtt{PP} \to \text{without})\, \phi(\mathtt{N} \to \text{anchovies}) \end{aligned}$$

$$\begin{aligned} P_\phi(W_2, T_1) &= \phi(\mathtt{S} \to \mathtt{N}\ \mathtt{V})\, \phi(\mathtt{V} \to \mathtt{V}\ \mathtt{N})\, \phi(\mathtt{N} \to \mathtt{N}\ \mathtt{P}) \times \\ &\times \phi(\mathtt{P} \to \mathtt{PP}\ \mathtt{N})\, \phi(\mathtt{N} \to \text{She})\, \phi(\mathtt{V} \to \text{eats}) \times \\ &\times \phi(\mathtt{N} \to \text{pizza})\, \phi(\mathtt{PP} \to \text{without})\, \phi(\mathtt{N} \to \text{hesitation}) \end{aligned}$$

Likelihood computation

$$\begin{aligned} P_\phi(W_1, T_2) &= \phi(\mathtt{S} \to \mathtt{N\ V})\,\phi(\mathtt{V} \to \mathtt{V\ N\ P})\,\phi(\mathtt{P} \to \mathtt{P\ PP}) \times \\ &\times \phi(\mathtt{N} \to \text{She})\,\phi(\mathtt{V} \to \text{eats})\,\phi(\mathtt{N} \to \text{pizza}) \times \\ &\times \phi(\mathtt{PP} \to \text{without})\,\phi(\mathtt{N} \to \text{anchovies}) \end{aligned}$$

$$\begin{aligned} P_\phi(W_2, T_1) &= \phi(\mathtt{S} \to \mathtt{N\ V})\,\phi(\mathtt{V} \to \mathtt{V\ N})\,\phi(\mathtt{N} \to \mathtt{N\ P}) \times \\ &\times \phi(\mathtt{P} \to \mathtt{PP\ N})\,\phi(\mathtt{N} \to \text{She})\,\phi(\mathtt{V} \to \text{eats}) \times \\ &\times \phi(\mathtt{N} \to \text{pizza})\,\phi(\mathtt{PP} \to \text{without})\,\phi(\mathtt{N} \to \text{hesitation}) \end{aligned}$$

Likelihood of the corpus

Probability of a sentence W : $P_\phi(W) = \sum_T P_\phi(W,T)$

Likelihood computation

$$\begin{aligned} P_\phi(W_1, T_2) &= \phi(\texttt{S} \to \texttt{N V})\,\phi(\texttt{V} \to \texttt{V N P})\,\phi(\texttt{P} \to \texttt{P PP}) \times \\ &\times \ \phi(\texttt{N} \to \text{She})\,\phi(\texttt{V} \to \text{eats})\,\phi(\texttt{N} \to \text{pizza}) \times \\ &\times \ \phi(\texttt{PP} \to \text{without})\,\phi(\texttt{N} \to \text{anchovies}) \end{aligned}$$

$$\begin{aligned} P_\phi(W_2, T_1) &= \phi(\texttt{S} \to \texttt{N V})\,\phi(\texttt{V} \to \texttt{V N})\,\phi(\texttt{N} \to \texttt{N P}) \times \\ &\times \ \phi(\texttt{P} \to \texttt{PP N})\,\phi(\texttt{N} \to \text{She})\,\phi(\texttt{V} \to \text{eats}) \times \\ &\times \ \phi(\texttt{N} \to \text{pizza})\,\phi(\texttt{PP} \to \text{without})\,\phi(\texttt{N} \to \text{hesitation}) \end{aligned}$$

Likelihood of the corpus

Probability of a sentence W : $P_\phi(W) = \sum_T P_\phi(W, T)$

If the training data comprises of sentences $W_1, W_2, \ldots, W_N$, then the likelihood is

$$L(\phi) = P_\phi(W_1) P_\phi(W_2) \cdots P_\phi(W_N)$$

Likelihood maximization

Approach

Starting at some initial parameters ϕ, re-estimate to obtain new parameters ϕ' for which $L(\phi') \geq L(\phi)$. *Repeat until convergence*

Parameter Estimation

Given some rule probabilities ϕ and training corpus $W_1, W_2 \ldots W_n$, the new parameters are obtained as:

$$\phi'(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C}) = \frac{count(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C})}{\sum_\alpha count(\mathtt{A} \rightarrow \alpha)}$$

$$\phi'(\mathtt{A} \rightarrow \mathrm{w}) = \frac{count(\mathtt{A} \rightarrow \mathrm{w})}{\sum_\alpha count(\mathtt{A} \rightarrow \alpha)}$$

What is $count(.)$?

Parameter Estimation

Given some rule probabilities ϕ and training corpus $W_1, W_2 \ldots W_n$, the new parameters are obtained as:

$$\phi'(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C}) = \frac{count(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C})}{\sum_\alpha count(\mathtt{A} \rightarrow \alpha)}$$

$$\phi'(\mathtt{A} \rightarrow \mathrm{w}) = \frac{count(\mathtt{A} \rightarrow \mathrm{w})}{\sum_\alpha count(\mathtt{A} \rightarrow \alpha)}$$

What is $count(.)$?

$$count(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C}) = \sum_{i=1}^{N} c_\phi(\mathtt{A} \rightarrow \mathtt{B}\ \mathtt{C}, W_i)$$

$$count(\mathtt{A} \rightarrow \mathrm{w}) = \sum_{i=1}^{N} c_\phi(\mathtt{A} \rightarrow \mathrm{w}, W_i)$$

Parameter Estimation

Given some rule probabilities ϕ and training corpus $W_1, W_2 \dots W_n$, the new parameters are obtained as:

$$\phi'(\mathtt{A} \to \mathtt{B}\ \mathtt{C}) = \frac{count(\mathtt{A} \to \mathtt{B}\ \mathtt{C})}{\sum_\alpha count(\mathtt{A} \to \alpha)}$$

$$\phi'(\mathtt{A} \to \mathrm{w}) = \frac{count(\mathtt{A} \to \mathrm{w})}{\sum_\alpha count(\mathtt{A} \to \alpha)}$$

What is $count(.)$?

$$count(\mathtt{A} \to \mathtt{B}\ \mathtt{C}) = \sum_{i=1}^{N} c_\phi(\mathtt{A} \to \mathtt{B}\ \mathtt{C}, W_i)$$

$$count(\mathtt{A} \to \mathrm{w}) = \sum_{i=1}^{N} c_\phi(\mathtt{A} \to \mathrm{w}, W_i)$$

$c_\phi(A \to \alpha, W_i)$ is the expected number of times $(A \to \alpha)$ is used in generating the sentence W_i, when the rule probabilities are given by ϕ

Computing Expected counts

Inside probabilities

The nonterminal A derives the string of words $w_i, \ldots w_j$ in the sentence :
$\beta_{ij}(A) = P_\phi(A \Rightarrow^* w_i \ldots w_j)$

Computing Expected counts

Inside probabilities

The nonterminal A derives the string of words $w_i, \ldots w_j$ in the sentence :
$\beta_{ij}(A) = P_\phi(A \Rightarrow^* w_i \ldots w_j)$

Outside probabilities

Beginning with the start symbol S we can derive the string
$w_1 \ldots w_{i-1} A w_{j+1} \ldots w_n$: $\alpha_{ij}(A) = P_\phi(S \Rightarrow^* w_1 \ldots w_{i-1} A w_{j+1} \ldots w_n)$

Computing Expected counts

Inside probabilities

The nonterminal A derives the string of words $w_i, \dots w_j$ in the sentence :
$\beta_{ij}(A) = P_\phi(A \Rightarrow^* w_i \dots w_j)$

Outside probabilities

Beginning with the start symbol S we can derive the string
$w_1 \dots w_{i-1} A w_{j+1} \dots w_n$: $\alpha_{ij}(A) = P_\phi(S \Rightarrow^* w_1 \dots w_{i-1} A w_{j+1} \dots w_n)$

Expected count

$$c_\phi(A \rightarrow BC, W) = \frac{\phi(A \rightarrow BC)}{P_\phi(W)} \sum_{1 \le i \le j \le k \le n} \alpha_{ik}(A) \beta_{ij}(B) \beta_{j+1,k}(C)$$

$$c_\phi(A \rightarrow w, W) = \frac{\phi(A \rightarrow w)}{P_\phi(W)} \sum_{1 \le i \le n} \alpha_{ii}(A)$$

And how to compute inside-outside probabilities

Inductively, as discussed earlier

$$\beta_{ii}(A) = \phi(A \rightarrow w_i)$$

$$\alpha_{1n}(S) = 1$$