Inside outside Algorithm

Jinman Zhao

Jun 2021

1 Introduction

We can use MLE on an annotated corpus to get the probability of PCFG rules. Then use inside outside algorithm algo to improve the likelihood of a different, unannotated corpus.

2 Inside and Outside

2.1 Inside

The inside probability $\alpha(A, i, j)$ is the total probability of generating words $x_{i:j}$ given the root nonterminal A.

- Initialize: if $A \to x_i \in R$, then $\alpha(A, i, i) = p(A \to x_i)$, otherwise 0.

- From bottom to top:

$$\alpha(A, i, j) = \sum_{A \to BC} \sum_{k=i}^{j-1} p(A \to BC) \cdot \alpha(B, i, k) \cdot \alpha(C, k+1, j)$$

2.2 Outside

The outside probability $\beta(A, i, j)$ is the total probability of generating words $x_{1:i-1} A x_{j+1:n}$.

- Initialize: $\beta(S, 1, n) = 1$, otherwise 0.

- From top to bottom:

$$\beta(A, i, j) = \sum_{B \to AC} \sum_{k=j+1}^{n} p(B \to AC) \cdot \beta(B, i, j) \cdot \alpha(C, j+1, k)$$
$$+ \sum_{B \to CA} \sum_{k=1}^{i-1} p(B \to CA) \cdot \beta(B, k, j) \cdot \alpha(C, k, i-1)$$

2.3 Some meaning

- Total probability of all trees:

$$Z = \alpha(S, 1, n)$$

- Total probability of trees contain $(A \to BC, i, j, k)$:

$$\mu(A \to BC, i, j, k) = p(A \to BC)\beta(A, i, j)\alpha(B, i, k)\alpha(C, k+1, j)$$

- Total probability of containing A from $x_{i:j}$:

$$\mu(A, i, j) = \alpha(A, i, j)\beta(A, i, j)$$

- Total probability of A is the preterminal of x_i:

$$\mu(A, i) = \mu(A, i, i)$$

2.4 Adjust the Parameters(EM)

$$count(A \to BC) = \frac{\sum_{i,k,j} \mu(A \to BC, i, k, j)}{Z}$$

$$p^{new}(A \to BC) = \frac{count(A \to BC)}{\sum_{X,Y} count(A \to XY)} = (easy\,to\,prove)\frac{\sum_{i,k,j} \mu(A \to BC, i, k, j)}{\sum_{i,j} \mu(A, i, j)}$$

3 Pseudo-code

Algorithm 1: Inside outside algorithm

Result: Update probability of rules
1   p = initial probability from annotated corpus;
2   S = sentences from new unannotated corpus;
3   **while** $|p - p_{new}| < threshold$ **do**
4   **foreach** $sent \in S$ **do**
5       $inside\_p$ = inside probability of sent;
6       $outside\_p$ = outside probability of sent;
7   compute p_{new}
8   **end**
9   **end**