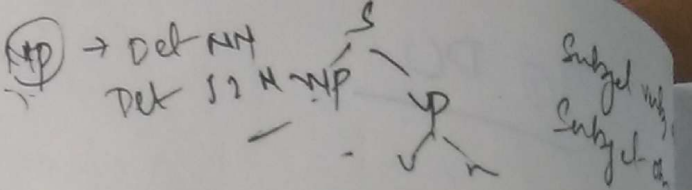
Code Mixed N2G



(1) LM from Monolingual (M) and from Code mixed (C) for a specific N-Gram. (N)

(2) Import N-Grams not in M but in C to M. Call these NGrams (C-N)

(3) A POS-Sequence of N is called Fragment (F)

(4) For generation we use a HMM to co-generate Fragment Sequence (F_S) and word sequence (S) so

~~$P(F_S|S)$~~ $P(S|F_S)$

(5) The HMM chooses a Fragment f_0 based on $\pi(f_0)$ (initial prob) scaled by a random variable ρ_0

$$f_0 \leftarrow \pi(f_0)\,\rho_0$$

(6) ~~Word Sequence~~ An HMM is defined as

$$p(F_S|S) = P(S|F_S)\,P(F_S)$$

$= \prod P(W_{seq_i} | f_i) \times \prod P(f_i | f_{i-1})$

Right

$\# P(W_{seq} | f_i)$

$P(W_{seq_i} | f_i) = P(w_{p_i} w_{q_i} w_{r_i} | t_{p_i} t_{q_i} t_{r_i})$

$= P(w_{p_i} | t_{p_i} t_{q_i} t_{r_i}) \times \delta_{p_i}$

$\times P(w_{q_i} | t_{p_i} t_{q_i} t_{r_i}) \times \delta_{q_i}$

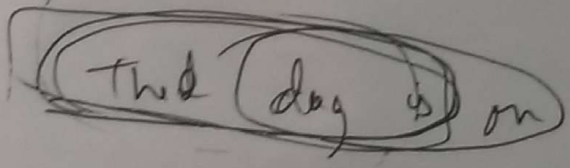
$\times P(w_{r_i} | t_{p_i} t_{q_i} t_{r_i}) \times \delta_{r_i}$

These are count based and derived from corpus.

$\# P(f_i | f_{i-1})$

$= P(f_i | f_{i-1}) \times \delta_{f_i}$

→ ke



The dog is → x

dog is on → y

is on the

on the skul.

$P(t_i | t_{i-1})$ → transition.