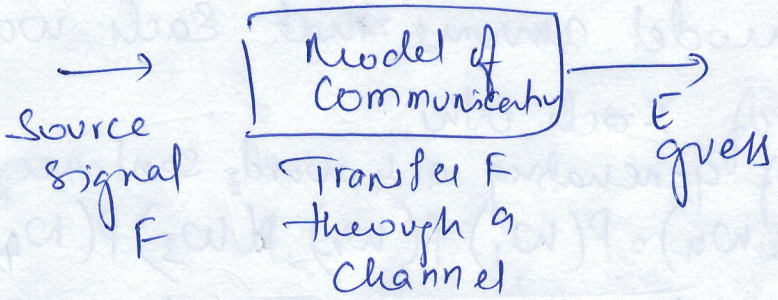
Statistical Machine Translation (SMT)

Given a foreign sentence F, we seek the English sentence E that maximizes $P(E|F)$. The most likely translation is written as

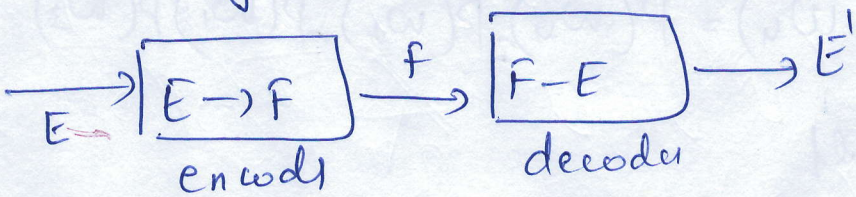
$$\underset{E}{\text{argmax}}\, P(E|F)$$

The $\underset{E}{\text{argmax}}$ can be read as "The English sentence E, out of all such sentences which yields the highest value for $P(E|F)$."

This problem is modelled using noisy channel model



Given F guess E



This model produces E' not E which is an approximation for E. The idea is to produce E' as close as we can be to E.

$$E' = \underset{E}{\text{argmax}}\, P(E|F) \approx \underset{E}{\text{argmax}}\, \underbrace{(F|E)}_{\text{lexical translation model}} \overset{\uparrow}{P(E)} \quad \text{— (1)}$$

Language model

The language model must predict the probability of generating the sentence $P(E)$.

A sentence generation can be modelled as a sequence of words using a chain rule

$$P(w) = P(w_1)\, P(w_2|w_1)\, P(w_3|w_2 w_1) \cdots P(w_n|w_{n-1}, w_{n-2})$$

↑ This is too lengthy.

To get rid of this lengthy subsequence we apply markov assuption

$$P(w_i|w_{i-1}, w_{i-2} \cdots w_1) = P(w_i|w_{i-1}^{i-n-1})$$

i.e Unigram model assumes that each word is independent of each other

The probability of generating a 4 word sentence is:

$$P(w_1 w_2 w_3 w_4) = P(w_1)\, P(w_2)\, P(w_3)\, P(w)$$

drawback of this is word order does not matter.

$$P(w_1 \cdots w_4) = P(w_4)\, P(w_2)\, P(w_1)\, P(w_3)$$

Bigram model

$$P(w_1 \cdots w_4) = P(w_1|\langle \underset{\text{sentence}}{\text{beginning of}} \rangle)\, P(w_2|w_1)\, P(w_3|w_2)\, P(w_4|w_3)$$

The probability of each word is conditioned on the previous word only.

Trigram model

$$P(w_1 - - w_4) = P(w_1 | \langle \text{begin of sentence} \rangle) P(w_2 | w_1) P(w_3 | w_2 w_1) P(w_4 | w_3 w_2)$$

The longer the better syntatically correct Sentences will be generated. But too longer subsequences can be generated with nearly zero probability.

How to compute $P(w_3 | w_2 w_1)$?

$$P(w_3 | w_2 w_1) = \frac{\text{Count}(w_1 w_2 w_3)}{\text{Count}(w_1 w_2)}$$

All paraters can be computed in this form for language models.

why $\text{argmax}_E P(E|F)$ has to be modeled as a product of $\text{argmax}_E P(F|E) P(E)$?

The idea is to estimate $P(E)$ & $P(F|E)$ and take the product a then rank all candidate english Sentences based on this product and pick the one that is highest.

Consider translating a foreign sentence F TO E.

O/p Sentences	P(E)	P(F|E)	P(F|E) P(E)
This rose is	low	high	low
is this rose	low	high	low
is rose this	low	high	low
This Pink is	low	low	low
Pink is good	low	low	low
This is rose	high	high	high

What factors influence P(F|E)?

1. Alignment factor a

How to assign values for P(F|E)?

$$P(F|E) = \frac{Count(F, E)}{Count(E)} \leftarrow \text{Sentence level}$$

Not possible to get true translation of sentences since the no of such exact sentences is very low. Hence we model this using a hidden variable a, that represent alignment between the individual words in a sentence. as marginilization over all alignment variables.

$$P(F|E) = \sum_{a}^{If} P(f, a|e) \leftarrow \text{word level}$$

or

$$\sum^{If} P(a, f|e)$$

According to Marginilization role

For Continious random variables $P(x) = \int_y P(x,y)\,dy$

For discrete random variables

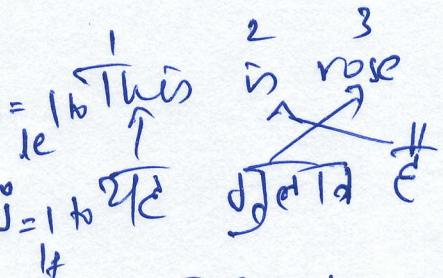
$$P(x) = \sum_y P(x,y)$$

Since our alignment variables are discrete $P(F|E)$ is marginalized using alignment variables as $\sum_a^{\text{all}} P(a,f|e)$ at word level.

E: vector of English words

F: vector of Foreign "

a: vector of alignment indices



$a[1] \to 1$

$a[2] \to 3$

$a[3] \to 2$

$a = \langle 1\ 3\ 2 \rangle$

$P(a,f|e)$ is called alignment Probability.

Interpretation of a:

$P(1\ 3\ 2 \mid \text{"This is rose"}) > P(1\ 2\ 3 \mid \text{"This is rose"})$

How Probable are the alignment a and the translation of f given e?

2) length factor of $F(l_f)$

$$P(F, a \mid E) = \sum_{l_f} P(F, a, l_f \mid E)$$

Again we introduce the marginalization over the lengths of foreign sentences. l_f.

$$P(F, a, l_f \mid E) = P(l_f \mid E) P(F, a \mid E, l_f)$$

Apply chain rule by which joint probalities are converted into conditional probalities.

Now if a 3 words sentences are given in E & F

$$P(F, a, 3 \mid E) = \underbrace{P(3 \mid E)}_{\substack{\text{Probality of} \\ \text{generating 3 worded} \\ \text{foreign sentence}}} P(f_1 a_1, f_2 a_2, f_3 a_3 \mid e, 3)$$

which can be generalized to a length l_f in foreign sentence as

$$P(l_f \mid E) \prod_{j=1}^{l_f} P(f_j, a_j \mid f_1^{j-1}, a_1^{j-1}, e, m)$$

Assumption is that there is one-to-one mapping between the words in E and F

If the english sentence has l_e words

$$E = E_1^{l_e} = e_1, e_2, \ldots e_{l_e}$$

Foreign stence has l_f words

$$F = F_1^{l_f} = f_1, f_2 \ldots f_{l_f}$$

then the alignment a, can be represented by a series $a_1^{l_f} = a_1, a_2 \ldots a_{l_f}$ of l_f values each between 0 to l_e such that if the word in position j of the foreign sentence is connected to the word in position i of the English sentence then $a_j = i$ & if it is not connected to any English word then $a_j = 0$.

$$P(F, a|E) = P(l_f|E) \prod_{j=1}^{l_f} P(a_j | a_1^{j-1}, f_1^{j-1}, l_f, E) \longrightarrow (2)$$

$$\prod_{j=1}^{l_f} P(f_j | a_1^{j}, f_1^{j-1}, l_f, E)$$

where $P(l_f|E)$ is called as length probability

$\prod_{j=1}^{l_f} (a_j | a_1^{j-1}, f_1^{j-1}, l_f, E)$ is an alignment probability

$\prod_{j=1}^{l_f} P(f_j | a_1^{j}, f_1^{j-1}, l_f, E)$ is called translation probability

IBM-Model I

In the equation (2) the conditional probabilities on the RHS cannot be taken as independent parameters. Hence IBM Model-1 makes assumptions

(1) $Pr(l_f | E)$ is independent of E & is fixed as a constant parameter ϵ.

(2) $P(a_j | a_1^{j-1}, f_1^{j-1}, l_f, E)$ depends on l_e, hence must be $\frac{1}{(l_e+1)}$

(3) $P(f_j | a_1^{j}, f_1^{j-1}, l_f, E)$ depends on f_j and e_{a_j}

∴ Estimation of translation probabilities for Model-I is the joint likelihood of a foreign string & an alignment given an English string is

$$P(F, a | E) = \frac{\epsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} t(f_j | e_{a_j})$$

$$P(F|E) = \frac{\epsilon}{(l_e+1)^{l_f}} \sum_{a_1=0}^{l_e} \sum_{a_2=0}^{l_e} \cdots \sum_{a_{l_f}=0}^{l_e} \prod_{j=1}^{l_f} t(f_j | e_{a_j})$$

We wish to adjust the translation probabilities so as to maximize $P(F|E)$ subject to the constraints that for each e

$$\sum_f t(f|e) = 1$$

There are many ways to define $P(F|E)$.

IBM Model-1 ignores lexical translations (word to word) i.e Each word in Foreign sentence is a translate of exactly zero/more of English sentence.

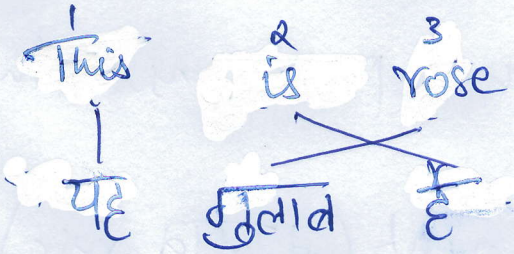
$$f = \langle f_1 \; - \; - \; f_{l_f} \rangle \quad e = \langle e_1 \; - \; - \; e_{l_e} \rangle$$

Assumption: Each Foreign word aligned to exactly One English word, we can represent an alignment of the French words by an array a of length l_f $\langle a_1 \; - \; - \; a_{l_f} \rangle$

Each a_i takes a value 0 to l_e denoting the index of English word to which the Foreign word f_i is aligned to.

If $a_i = 0$ means the foreign word f_i is aligned to NULL.

Consider a sentence pair $\langle F, E \rangle$



$a = \langle 1, 3, 2 \rangle \quad l_e = 3 \quad l_f = 3$

A length l_f is chosen according to a distribution $P(l_f|l_e)$ in this example $P(3|3)$

Then each foreign word position aligns to an English word (or null) according to uniform distribution

$$P(a_i = j \mid l_e) = \frac{1}{l_e + 1} \quad \text{in this case } \frac{1}{3+1}$$

Finally each foreign word f_i is translated according to a ~~uniform~~ distribution ~~$P(a_i$~~ conditioned on the aligned English word $P(f_i \mid e_{a_i})$

Here for this alignment we multiply

$P($ये$\mid$the$)$ $P($जुलाब $\mid$rose$)$ $P($है $\mid$is$)$

How to estimate $P(f \mid e)$

Suppose you have ever seen a English word rose aligned with जुलाब & जुलाब
↓ 7 time ↓ 3 time

$P($जुलाब$\mid$rose$) = 0.7$ $P($जुलाब$\mid$rose$) = 0.3$

Since these correspond to the proportions in the data you observed.

$$\therefore P(\text{जुलाब} \mid \text{rose}) = \frac{\text{Count}(\text{जुलाब}, \text{rose})}{\text{Count}(\text{जुलाब}, \text{rose}) + \text{Count}(\text{जुलाब}, \text{rose})}$$

$$= \frac{7}{10} = 0.7$$

$$P(\text{जुलाब} \mid \text{rose}) = \frac{\text{count}(\text{जुलाब}, \text{rose})}{}$$

$$P(F|E) = \frac{\epsilon}{(l_e+1)^{l_f}} \prod_{j=1}^{l_f} \sum_{i=0}^{l_e} t(f_j|e_{a_i})$$

IBM Model-1 & EM

Need the expected no of times word e connects to word f in the translation (f|e)

this is the (expected) count of f given e for (f|e)

$$c(f|e,F,E) = \sum_a P(a|E,F) \sum_{j=1}^{l_f} \delta(f,f_j)\,\delta(e,e_{a_j})$$

denotes the no of times e connects to f in a

∴ we need to compute $P(a|E,F)$

$$P(a|E,F) = \frac{P(F,a|E)}{P(F|E)}$$ → we have derive this for IBM Model-1

$$= \frac{\epsilon/(l_e+1)^{l_f} \prod_{j=1}^{l_f} t(f_j|e_{a(j)})}{\epsilon/(l_e+1)^{l_f} \prod_{j=1}^{l_f} \sum_{i=0}^{l_e} t(f_j|e_i)}$$

$$= \prod_{j=1}^{l_f} \frac{t(f_j|e_{a(j)})}{\sum_{i=0}^{l_e} t(f_j|e_i)}$$

Maximization Step

Now we to collect the counts

Evidence from a sentence pair E, F that a word f is a translation of word e.

$$c(f|e;E,F) = \sum_a P(a|E,F) \sum_{j=1}^{l_f} \delta(f,f_j)\,\delta(e,e_{a(j)})$$

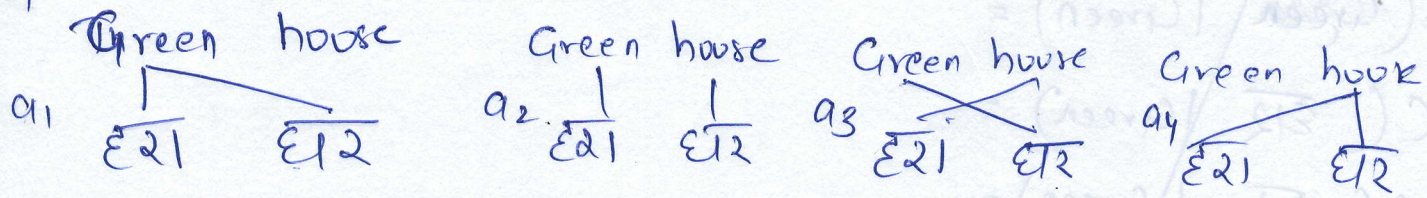
$$= \frac{t(f|e)}{\sum_{i=0}^{l_e} t(f|e_i)} \sum_{j=1}^{l_f} \delta(f,f_j)\,\delta(e,e_i)$$

After collecting the counts over the corpus, we can estimate the model as

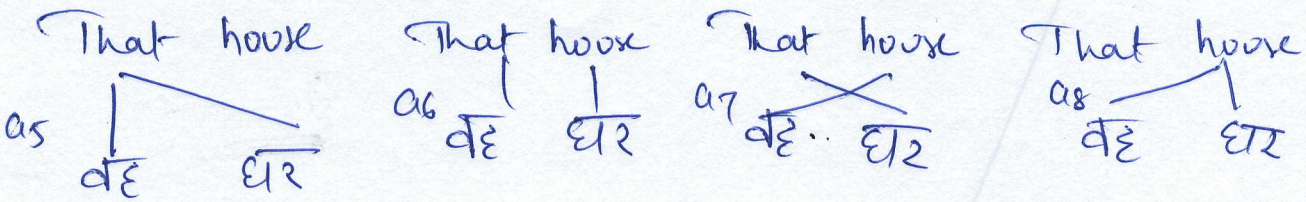
$$t(f|e,E,F) = \frac{\sum_{(E,F)} c(f|e;E,F)}{\sum_f \sum_{(E,F)} c(f|e;E,F)}$$

Learning parameters : IBM Model-I

S_1



S_2



Initial all t(f|e) uniformly

Given → / t ↓	हरा	घर	वह
Green	1/3	1/3	1/3
house	1/3	1/3	1/3
That	1/3	1/3	1/3

Compute posterior for each alignment

$P(a_1, f|e) = P(\text{हरा}|\text{Green})\, P(\text{घर}|\text{Green}) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

$P(a_2, f|e) = P(\text{हरा}|\text{Green})\, P(\text{घर}|\text{house}) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

$P(a_5, f|e) = P(\text{वह}|\text{That})\, P(\text{घर}|\text{That}) = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

$P(a_8) = \frac{1}{9}$

Normalize alignments

i.e Compute $P(a \mid F, E) = \frac{P(a, F \mid E)}{P(F \mid E)}$

$P(F \mid E) = a_1 + a_2 + a_3 + a_4 = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{4}{9}$

$\frac{P(a_1, f \mid e)}{P(F \mid E)} = \frac{\frac{1}{9}}{\frac{4}{9}} = \frac{1}{4}$

$P(a_2) = \frac{1}{4}$

$P(a_4) = \frac{1}{4}$

$P(a_5, f \mid e) = \frac{1}{4}$

$P(a_8, f \mid e) = \frac{1}{4}$

Compute fractional counts

$C(\overline{\text{इस}}, \text{Green}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ from (a_1, a_2)

$C(\overline{\text{इस}}, \text{Green}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_1, a_3)

$C(\overline{\text{वह}}, \text{Green}) = 0$ (These words never occured in any alignments)

$C(\overline{\text{इस}}, \text{house}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_3, a_4)

$C(\overline{\text{घर}}, \text{house}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_2, a_4)

$C(\overline{\text{वह}}, \text{house}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_7, a_8)

$C(\overline{\text{इस}}, \text{That}) = 0$

$C(\overline{\text{घर}}, \text{That}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_5, a_7)

$C(\overline{\text{वह}}, \text{That}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ (a_5, a_6)

$$\text{totalCount}(\text{Green}) = \frac{1}{2} + \frac{1}{2} + 0 = 1$$

$$(\text{house}) = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$$

$$(\text{that}) = 0 + \frac{1}{2} + \frac{1}{2} = 1$$

Maximization step

$$t(\text{हरा} \mid \text{Green}) = \frac{1}{2}$$

$$t(\text{घर} \mid \text{Green}) = \frac{1}{2} / \frac{3}{2} = \frac{1}{3}$$

$$t(\text{वह} \mid \text{Green}) = 0$$

$$t(\text{हरा} \mid \text{house}) = \frac{1}{2} / \frac{3}{2} = \frac{1}{3}$$

$$t(\text{घर} \mid \text{house}) = \frac{1}{2} / \frac{3}{2} = \frac{1}{3}$$

$$t(\text{वह} \mid \text{house}) = 1/2 / \frac{3}{2} = \frac{1}{3}$$

$$t(\text{हरा} \mid \text{That}) = 0$$

$$t(\text{घर} \mid \text{That}) = 1/2 / 1 = \frac{1}{2}$$

$$t(\text{वह} \mid \text{That}) = \frac{1}{2} / 1 = \frac{1}{2}$$

Keep iterating over Expaction & Maximization Steps until Convergence of parameters $t(f_j \mid e_{a_j})$

IBM Model-2

In IBM model-1 all alignments are Equally likely which is wrong. Since the languages having SOV order like most Indian languages the verbs predominately move to the end of the Sentence. Once the position of the noun is fixed adjectives qualify there nouns. Hence it is realistic to assume a probability distribution over position of word as

$$a(a_j \mid j, l_e, l_f)$$

(a_j: English, j: Foreign)

Where j is the position of the foreign word and a_j is the position of English word. The function $a(a_j|j, l_e, l_f)$ models the probability of word in the position a_j in source sentence being reordered to j^{th} position in the target sentence. For Example

P = 1 That 2 house 3 is 4 beautiful

j = वह 1 घर 2 सुंदर 3 है 4

$a = \langle 1, 2, 4, 3 \rangle$

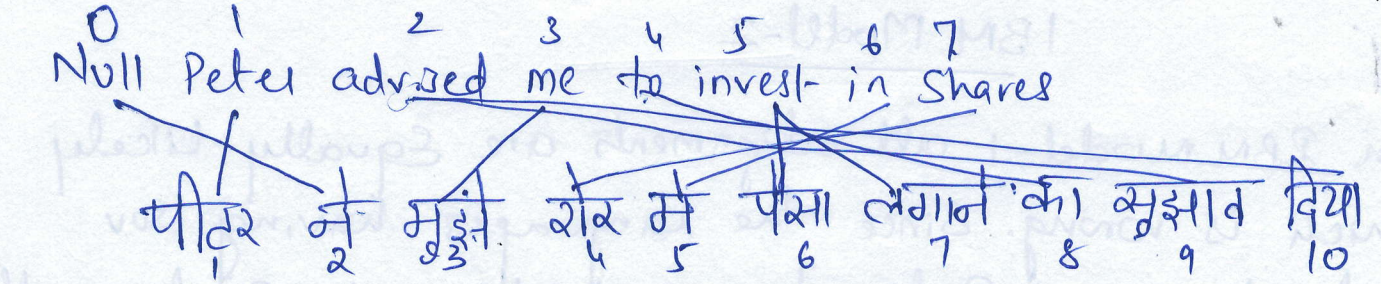
$t(\text{वह} \mid \text{That})\ a(1|1,4,4) \times$

$t(\text{घर} \mid \text{house})\ a(2|2,4,4) \times$

$t(\text{सुंदर} \mid \text{beautiful})\ a(4|3,4,4) \times$

$t(\text{है} \mid \text{is})\ a(3|4,4,4)$

Another



$a = \langle 1\ 0\ 3\ 7\ 6\ 5\ 5\ 4\ 2\ 2 \rangle$

$t(\text{पीटर} | \text{Peter})\ a(1|1, 7, 10)$

$t(\text{ने} | \text{Null})\ a(0|2, 7, 10)$

$t(\text{मुझे} | \text{me})\ a(3|3, 7, 10)$

$t(\text{शेयर} | \text{shares})\ a(7|4, 7, 10)$

$t(\text{में} | \text{in})\ a(6|5, 7, 10)$

$t(\text{पैसा} | \text{invest})\ a(5|6, 7, 10)$

$t(\text{लगाने} | \text{invest})\ a(5|7, 7, 10)$

$t(\text{का} | \text{to})\ a(4|8, 7, 10)$

$t(\text{सुझाव} | \text{invest})\ a(2|9, 7, 10)$

$t(\text{दिया} | \text{invest})\ a(2|10, 7, 10)$

$$P(f_1, f_2 \ldots f_{l_f}, a_1, a_2 \ldots a_{l_f} | e_1, e_2 \ldots e_{l_e})$$

$$\prod_{j=1}^{l_f} a(a_j | j, l_e, l_f)\ t(f_j | e_{a_j})$$

To Estimate $a(a_j|j, l_e, l_f)$ we use EM as we did in Model-1 for translation parameters.

$$a(a_j|j, l_e, l_f) = \frac{c(a_j|j, l_e, l_f)}{c(j, l_e, l_f)}$$

where $c(a_j|j, l_e, l_f)$ is the no of times we see an English sentence of length l_e & foreign sentence length l_f were word j in the foreign sentence is aligned to word a_j in the English sentence.

$c(j, l_e, l_f)$ the no of times we see English & foreign sentences with length l_e & l_f respectively

IBM Model-3

The assumption in IBM Models 1 & 2 is that Each word in English generates one or many foreign words. but their distribution is Uniform whether it generates one word or many words.

This factor is handeld as a fertility factor in IBM Model3 for Each English word e in English Sentence E, we choose a fertility ϕ. The choice of ϕ depends only on e which is modeled as a probability distribution parameter $n(\phi|e)$. These parameters for Each English word e is computed from the corpus. The parameter $n(\phi|e)$ is used for non Null words only.

The fertility of a Null token depends on the context of English words & foreign sentence lengths. Hence the fertility of a Null word is treated differently. $n(1|NULL)$, $n(2|NULL)$ - - - $n(25|NULL)$ ... needs to Computed which is infeasible.

∴ we represent a parameter P_0 which is the probability of having No Null token inserted after Each English word e. P_1 is the probability of generating one Null token inserted after Each English word e.

The number of Foreign words generated from a Null token is represented as ϕ_0.

$$\text{No of o/p words} = \sum_{i=1}^{l_e} \phi_i = l_f - \phi_0$$

The estimation of ϕ_0 is treated as a binomial distribution since each English word e may generate a Null (P_1) or not (P_0)

$P(\phi_0)$ is the probability of generating ϕ_0 foreign words from a null token is estimated as

$$P(\phi_0) = \binom{l_f - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_f - 2\phi_0} \quad \left(\text{using binomial distribution}\right)$$

IBM Model-3 Parameters:

1. word / lexical translations from IBM Model 1: $t(f_j | e_a$
2. Fertility of Non Null words from model 3: $n(\phi | e_i)$
3. Fertility of Null words: $P(\phi_0)$
4. Distortion probability distribution which will be in the reverse direction of alignment probability shown in IBM Model 2. $d(j | a(j), l_e, l_f)$

$$\therefore P(F|E) = \sum_{a} P(F, a|E)$$

$$= \sum_{a_1=0}^{l_e} \cdots \sum_{a_{l_f}=0}^{l_e} P(F, a|E)$$

$$= \sum_{a(i)=0}^{l_e} \cdots \sum_{a(l_f)=0}^{l_e} \prod_{j=1}^{l_f} \binom{l_f - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_f - 2\phi} \times$$

$$\prod_{i=1}^{l_e} \phi_i \, n(\phi_i|e_i) \times$$

$$\prod_{j=1}^{l_f} t(f_j|e_{aj}) \, d(j|a(j), l_e, l_f)$$