Natural Language Processing
CSE 325/425

Sihong Xie

Lecture 23:

- Statistical machine translation
- Alignment models (IBM model-1 and HMM)

Statistical MT framework

- Goals of MT
  - Fluency and faithfulness

English French

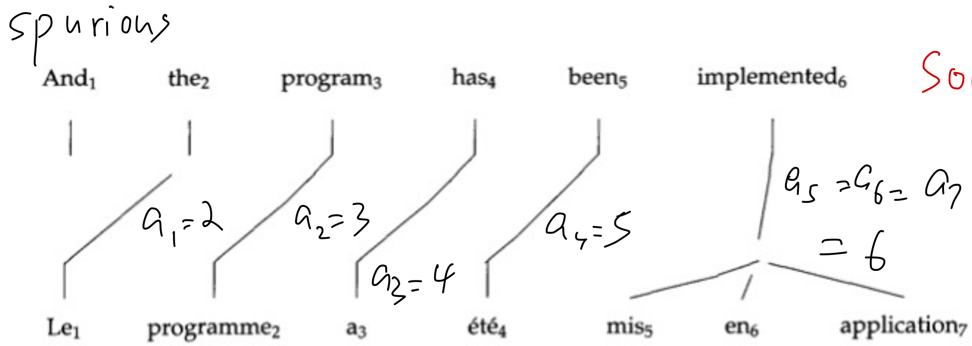
$$\text{best-translation } \hat{T} = \text{argmax}_T \underbrace{\mathrm{P(T)}}_{\text{LM, n-gram}} \underbrace{P(S|T)}_{\text{faithfulness}}$$

  - More formally,
    - $E=(e_1, \ldots, e_I)$: English (the target language).
    - $F=(f_1, \ldots, f_J)$: a foreign language (Spanish/French/…).
    - Language model $P(E)$
    - Translation model: $P(F|E)$
    - Using the Bayes theorem, *decoding* is defined as

$$E^* \quad \cancel{F}^* = \arg\max_{\cancel{F}\ E} P(E|F) = \arg\max_{\cancel{F}\ E} \frac{P(F|E)P(E)}{P(F)} = \arg\max_{\cancel{F}\ E} P(F|E)P(E)$$

Alignment

- Word alignment: mapping words in *E* to words in *F*.
  - Multiple target words can be mapped to one source word.
  - Example:



$E^* = \underset{E}{\arg\max}\; P(E) \times P(F|E)$

Source E

⇓ Translate

Target F

In ML: this is a so-called "structured prediction"

Applications: CV, NLP

- Two representation of an alignment *A*
  - Recording the mapping *A* directly: a_1=2, a_2=3, …, a_7=6.
  - Use an alignment matrix

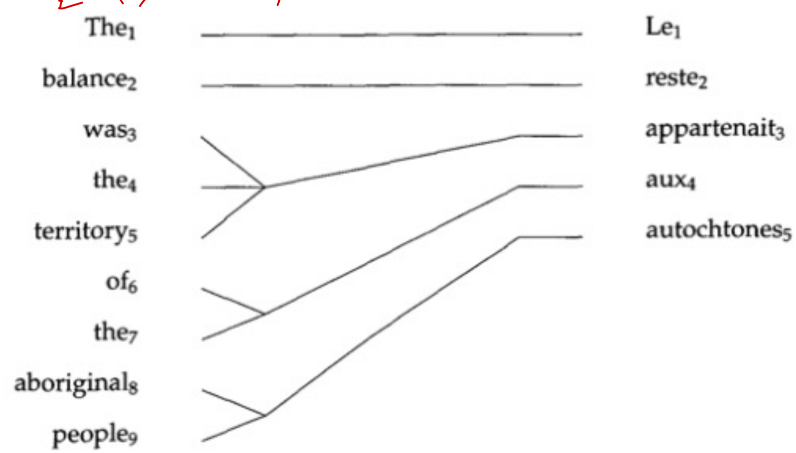
E \ F	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0
6	0	0	0	0	1	1	1

Alignment

Alignment matrix 9 × 5

- One ~~source~~ target word mapped to multiple ~~target~~ source words.

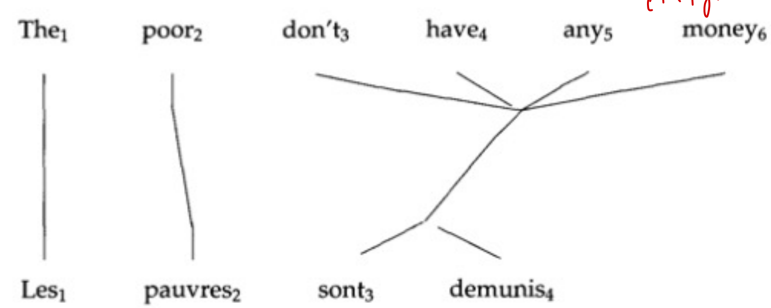
E (Source) F (Target)



- Many-to-many mapping (Phase-based Alignment)

E

F



- Allow spurious word NULL for words that can't be mapped.

Alignment (IBM model 1)

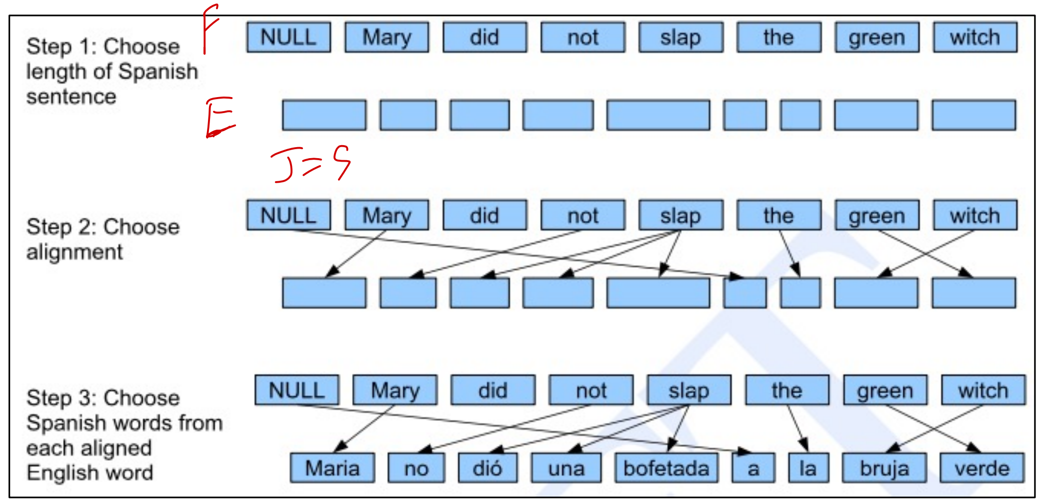
Attention model in Neural network ML

A : random matrix

$A \in \mathcal{A}$: Sample space of all possible alignments.

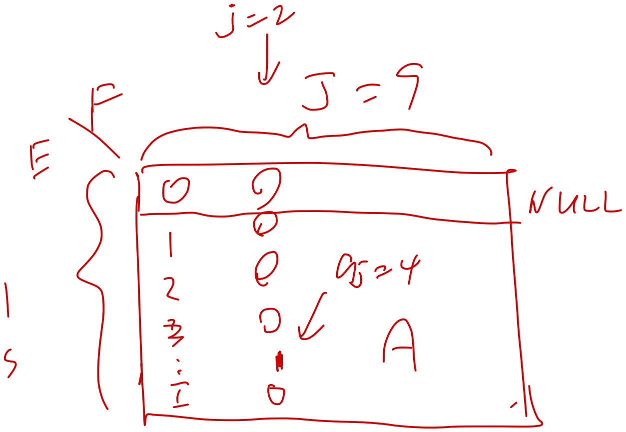
Pr (A)

- Published in *Brown, etc.1993*.
- Translation model $P(F|E) = \sum_{A} P(F,A|E)$
- Generative story
  1. generate length of the source/foreign language *J;*
  2. generate an alignment *A*;
  3. generate words $E=(e_1, \ldots, e_I)$ in the target language.



P(F, A|E)?

The Mathematics of Statistical Machine Translation. Brown, etc. 1993 Computational Linguistics



$J+1 = I+1$ rows

Alignment (IBM model 1)

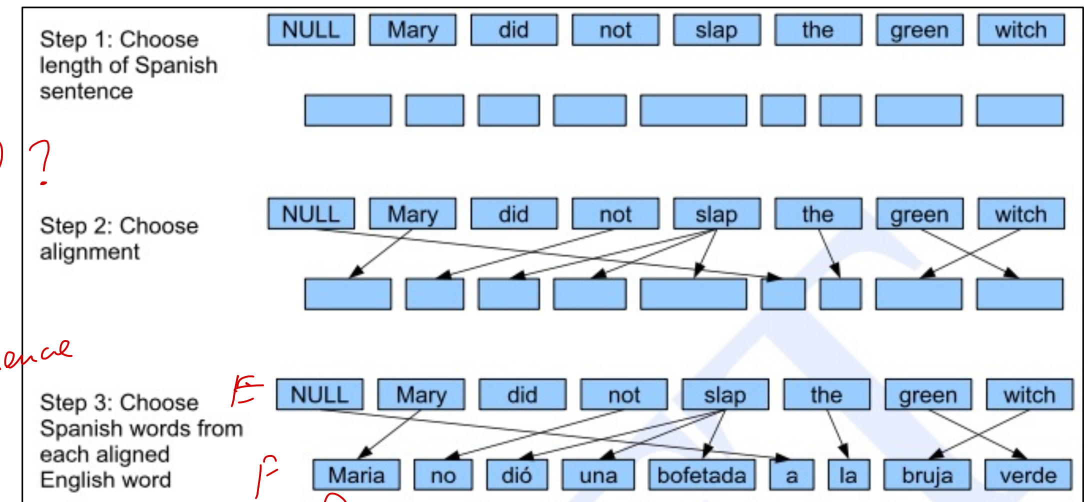
- Probability of generating a length *J*: a small positive number ϵ
- Probability of an alignment between *I* and *J* words: $P(A(I,J)) = \frac{1}{(I+1)^J}$
- Probability of the foreign word f_j given the aligned target word e_{a_j}

$j \in \{1, \cdots, J\}$

$a_j \in \{0, \cdots, I\}$

$$P(f_j|e_{a_j})$$

$P(no|not)$?



- Probability of the foreign sentence and an alignment *A*:

conditional Independence

$$P(F, A|E) = \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} P(f_j|e_{a_j})$$

A

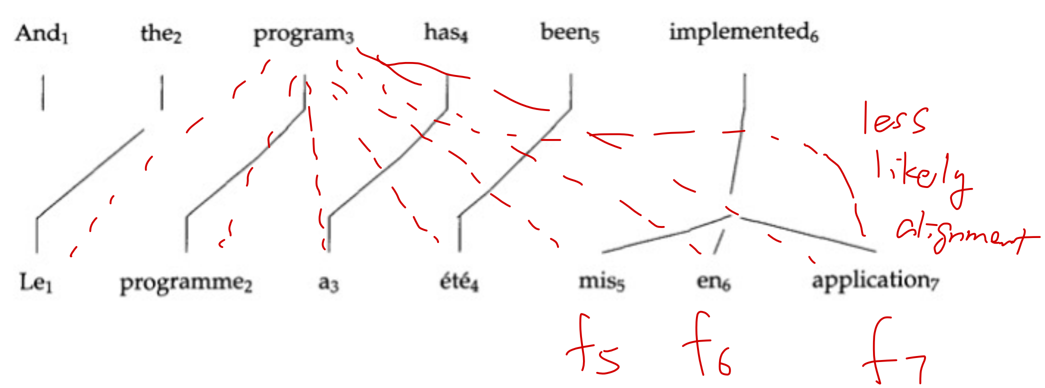
- Probability of *F* given *E*:

$$P(F|E) = \sum_A P(F, A|E) = \sum_A \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} P(f_j|e_{a_j})$$

$j = 1 \cdots 9$

Alignment (HMM)

- IBM Model 1 makes several strong assumptions
  - two foreign words are mapped independently;
  - all alignments have the same probability.
- Both are not true
  - the last three French words are mapped jointly to the last English word.
    - need to use some joint probability to model such linguistic phenomena;
  - locality: two consecutive French words are likely mapped to consecutive English words.
    - some alignments are unlikely.
    - can you give one?



$$Pr(f_5, f_6, f_7 \mid A, E) \neq Pr(f_5 \mid A, E) \cdots Pr(f_7 \mid A, E)$$

Alignment (HMM)

- HMM alignment model
  - generates an alignment and the observed foreign sentence
    - $a_j \in \{0, \dots I\}$
    - alignment ⇔ POS-tags.
    - a_j $\quad$ $a_1 \dots a_{j-1}$
    - transition probabilities: align the next foreign word, given previous alignments.
    - foreign sentence ⇔ observed words in a sentence.
    - emission probabilities: emit a word given the up-to-date alignments. $a_1 \dots a_j$

f_j

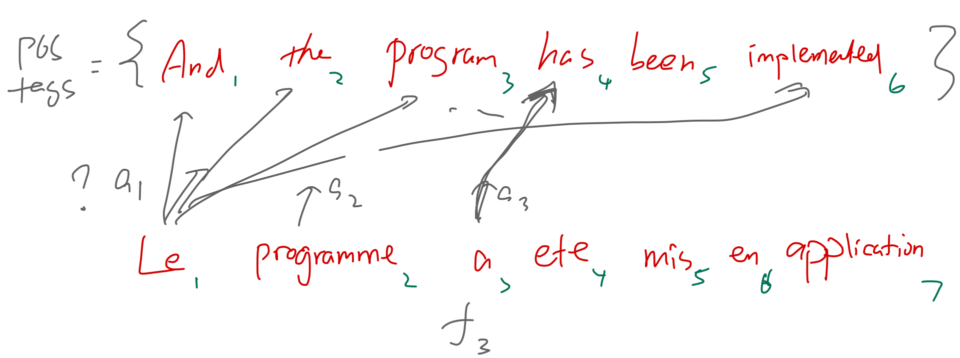
$$P(f_1^J, a_1^J | e_1^I) = P(J|e_1^I) \times \prod_{j=1}^{J} P(f_j, a_j | f_1^{j-1}, a_1^{j-1}, e_1^I)$$

$$= P(J|e_1^I) \times \prod_{j=1}^{J} P(a_j | f_1^{j-1}, a_1^{j-1}, e_1^I) \times P(f_j | f_1^{j-1}, a_1^{j}, e_1^I)$$

a_3 $\quad$ a_1^2

HMM

$$\begin{array}{ccccccccc} a_1 & \rightarrow & a_2 & \rightarrow & a_3 & \rightarrow & \cdots & \rightarrow & a_J \\ \downarrow & & \downarrow & & \downarrow & & & & \downarrow \\ f_1 & & f_2 & & f_3 & & & & f_J \end{array}$$



Alignment (HMM)

- Make some Markov assumptions to simplify the above joint probability.

disgard $(a_1 \cdots a_{j-2})$

$$P(a_j|f_1^{j-1},a_1^{j-1},e_1^I) = P(a_j|a_{j-1},I)$$
$$P(f_j|f_1^{j-1},a_1^{j},e_1^I) = P(f_j|e_{a_j})$$

$Pr(a \mid has)$

f_3 $e_{a_3} = e_4$ (b/c $a_3 = 4$)

- The final joint distribution

$$P(F,A|E) = P(J|I) \times \prod_{j=1}^{J} P(a_j|a_{j-1},I)P(f_j|e_{a_j})$$

- The translation model

$$P(F|E) = P(J|I) \times \sum_{A} \prod_{j=1}^{J} P(a_j|a_{j-1},I)P(f_j|e_{a_j})$$

$$= \sum_{A} Pr(F, A \mid E)$$

sum - product
$\Downarrow$
product - sum
(forward Alg.)

Alignment (HMM)

- The transition probability should encourage alignment locality.
  - "*the English words that generate neighboring Spanish words are likely to be nearby*" -- SLP
  - $P(a_j|a_{j-1}, I)$ should be large if a_j is close to a_{j-1}
  - Locality is a relative concept and absolute positions are not relevant.
    - P(7|6, 15) = P(9|8, 15). place – invariance
  - Model "jumps" of the alignment pointers
    - $P(a_j|a_{j-1}, I)$ is a decreasing function of the jump $|a_j - a_{j-1}|$

E

F

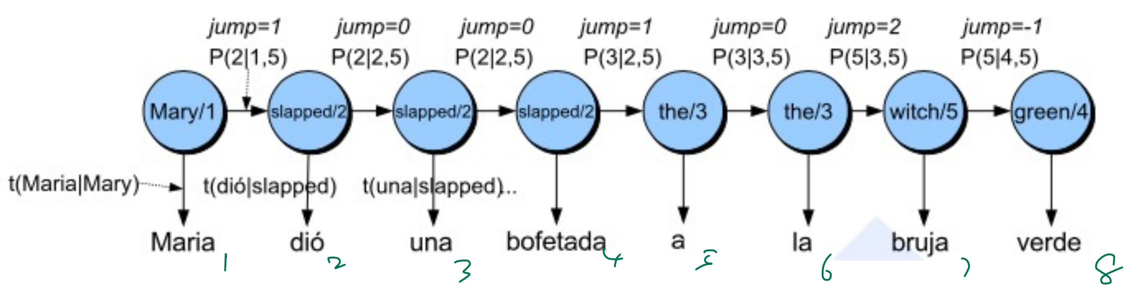
$a_2 = 1$

$a_3 = 10$

$j-1=2$ $\quad$ $j=3$

$I = 8$

$J = 8$



Mary_1 slapped_2 the_3

green_4 witch_5

$$Pr(F, A \mid E) = P(J \mid I) \times Pr(2 \mid 1, 5) \times Pr(\text{Maria} \mid \text{Mary})$$

$$\times\, Pr(2 \mid 2, 5) \times Pr(\underset{f_2}{\text{dió}} \mid \underset{e_{a_2}}{\text{slapped}}) \times \ldots$$

f_1 $\quad$ e_{a_1}