Natural Language Processing
CSE 325/425

Sihong Xie

Lecture 24:

- Training alignment models
- Phase-based translation

learn $P(f_j | e_{a_j})$ from data

$\{(E_s, F_s)\}_{s=1}^{|S|}$

Training alignment models

- We have describe two alignment models
  - IBM Model-1
  - HMM $P(F, A|E) = P(J|I) \times \prod_{j=1}^{J} P(a_j|a_{j-1}, I) P(f_j|e_{a_j})$

translation model parameters to be learned.

a_j comes from Alignment matrix A

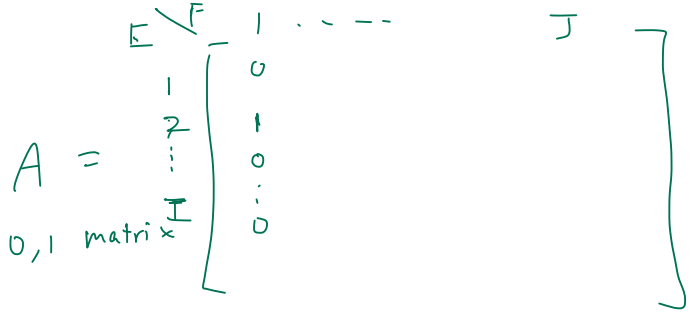
and is unknown

- How to find he probability $P(f_j|e_{a_j})$ $\triangleq$ prob of translating e_{a_j} into f_j where a_j = index of word in E translated from j-th word in F
  - Given an alignment A of two sentences (E, F), this probability can be found using MLE.
  - Example: $P(verde|green)$ is estimated by the number of times *verde* is aligned with *green*, divided by number of *green*.

(E_1, F_1) ... $(E_{|S|}, F_{|S|})$

A_1 ... $A_{|S|}$

EM training



- However, we don't have a large corpus of aligned source and target sentences and need to estimate both $P(f_j|e_{a_j})$ and A.
  - Global model parameter: translation probability $P(f_j|e_{a_j})$
  - Local hidden (latent) variables A
  - Similar to the EM training for HMM with model parameters and latent tags.
- E-step: given translation probabilities, estimate probability of each possible alignment of each (E, F).
- M-step: given the distribution of alignments, estimate translation probabilities.

EM training

neither the Model-2, nor the HMM

- Assume a simplified model $P(F, A|E) = \prod_{j=1}^{J} P(f_j|e_{a_j})$ → critical for phase-based Translation,
  - a fixed Alignment
  - This is the common portion of both IBM Model-1 and HMM.
- Use the example training corpus of two pairs of (E, F)

E_1	green house	E_2	the house
	(A) ?		(A) ?
F_1	casa verde	F_2	la casa

- Parameters: translation probabilities initialized uniformly

Max Entropy

P(casa|green)=1/3	P(verde|green)=1/3	P(la|green)=1/3
P(casa|house)=1/3	P(verde|house)=1/3	P(la|house)=1/3
P(casa|the)=1/3	P(verde|the)=1/3	P(la|the)=1/3

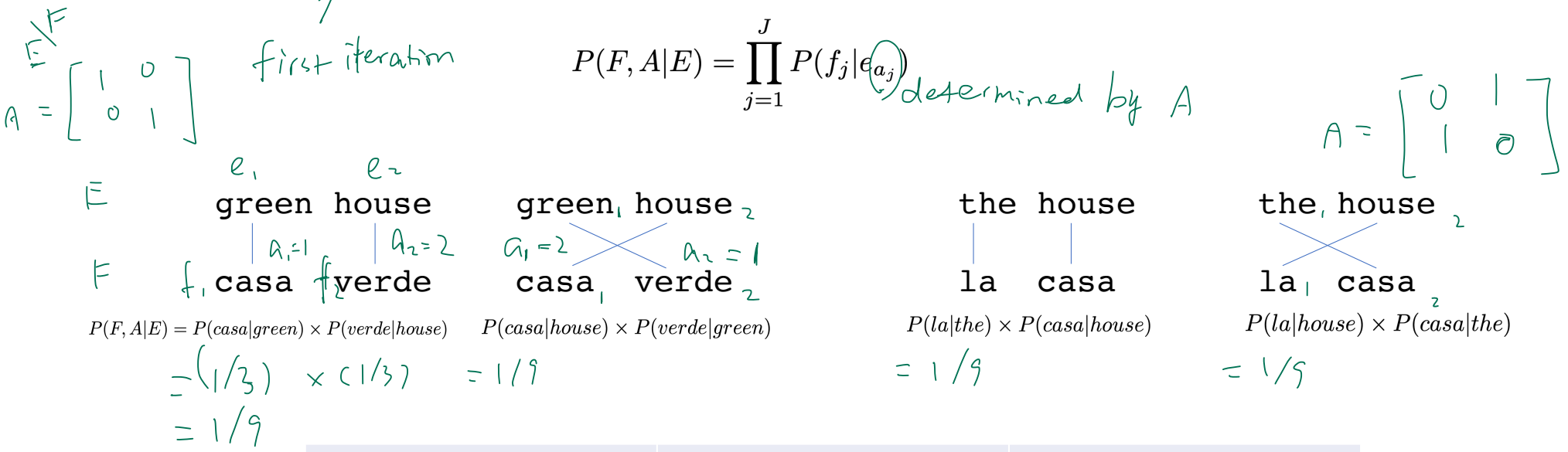
English.

Spanish

EM training

- E-step 1a: for each pair and each possible alignment A, find

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



P(casa|green)=1/3	P(verde|green)=1/3	P(la|green)=1/3
P(casa|house)=1/3	P(verde|house)=1/3	P(la|house)=1/3
P(casa|the)=1/3	P(verde|the)=1/3	P(la|the)=1/3

EM training

- E-step 1b: for each pair and each possible alignment A, find

$$P(A|E,F) = P(A|E) = \frac{P(F,A|E)}{\sum_A P(F,A|E)} = \frac{P(F,A|E)}{P(F|E)}$$

(in HMM, $P(Q|O)$)

green house	green house	the house	the house
casa verde	casa verde	la casa	la casa
A_1, Prob $\frac{1}{2}$	A_2, Prob $\frac{1}{2}$	A_1, $\frac{1}{2}$	A_2, $\frac{1}{2}$
$P(F,A|E) = P(casa|green) \times P(verde|house)$ $= 1/9$	$P(casa|house) \times P(verde|green)$ $= 1/9$	$P(la|the) \times P(casa|house)$	$P(la|house) \times P(casa|the)$

$P(A_1|E,F) = \frac{(1/9)}{(2/9)} = \frac{1}{2}$

- E-step 1c: find expected counts of word translation.

Expected Count(casa | green) = Σ Prob (when "casa" is aligned with "green")

count(casa|green)=1/2	count(verde|green)=1/2	count(la|green)=0	Total(green)=1
count(casa|house)=1/2+1/2	count(verde|house)=1/2	count(la|house)=1/2	Total(house)=2
count(casa|the)=1/2	count(verde|the)=0	count(la|the)=1/2	Total(the)=1

EM training

- M-step 1: re-estimate the translation probabilities by row normalization

count(casa|green)=1/2	count(verde|green)=1/2	count(la|green)=0	Total(green)=1
count(casa|house)=1/2+1/2	count(verde|house)=1/2	count(la|house)=1/2	Total(house)=2
count(casa|the)=1/2	count(verde|the)=0	count(la|the)=1/2	Total(the)=1

$P(f_j \mid e_{a_j})$

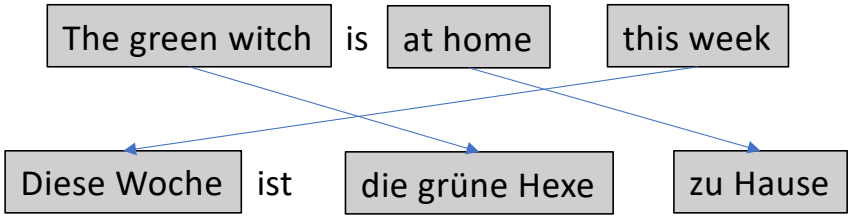
P(casa|green)=1/2	P(verde|green)=1/2	P(la|green)=0
P(casa|house)=1/2	P(verde|house)=1/4	P(la|house)=1/4
P(casa|the)=1/2	P(verde|the)=0	P(la|the)=1/2

- Observation: the unlikely translation has a smaller probability.
- Then go to E-step 2, M-step 2, E-step 3, and so on …

HW 6 ?

Phase-based translation

- We have been working with word-based translation models.
- Phase-based translation has some more advantages:
  - phases are constituent unit that can somehow be move freely in a sentence.



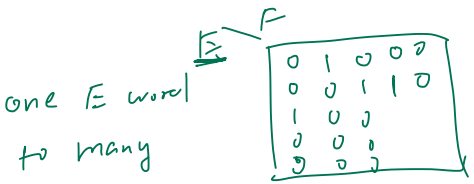
  - phases have more contexts for faithful and robust translations.

Mary didn't slap the green witch

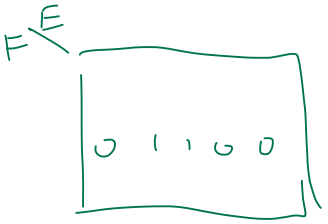
Maria no dió una bofetada a la bruja verde

Mary not gave a slap to the witch green

word – word translation (one-to-one) ⇒ phase – phase translation (many-to-many)



symmetrization



Phase-based translation

shallow parsing (stay at the phase level.)

- Segment *F* into phases $F = (\bar{f}_1, \ldots, \bar{f}_I)$
- Generate $E = (\bar{e}_1, \ldots, \bar{e}_J)$, with faithfulness $P(F|E) = \prod_{i=1}^{I} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1})$
  - Phase-translation table records $\phi(\bar{f}_i|\bar{e}_i)$
  - Distribution of "distortion" $d(a_i - b_{i-1})$
    - a_i : the position of the first word of the *i*-th phase in *E* (translation of the *i*-th)
    - b_{i-1} : the position of the first word of the (*i-1*)-th phase in *E* (translation of the (*i-1*)-th)

preserve "locality": two adjacent English phrases (i, i-1) should be translated into Spanish phrases "close to" each other. ?

d(distance) ↓ if distance ↑

	e_0	e_1	e_2	e_3	e_4	e_5
E		Mary$_1$	didn't$_2$	slap$_7$	the$_4$	green witch$_6$
F		*Maria*$_1$	*no*$_2$	*dió una bofetada*$_5$	*a*$_6$ *la*$_7$	*bruja verde*$_8$ *verde*$_9$
	b_0	a_1	a_2		b_4	a_5
	=	=	$= b_2$	a_3	=	=
	0	b_1			7	8
		=	$= 2$	$= 3$		
		1				

$i=2$, $i=3$, $i=4$, $i=5$

$d(a_i - b_{i-1}) = 1$ if $i=3$

$Pr(F, A|E) = Pr(\text{Maria}|\text{Mary}) \times d(a_1 - b_0) \times Pr(\text{no}|\text{didn't}) \times d(a_2 - b_1)$
$\times \cdots \times Pr(\text{bruja verde}|\text{green, house}) \times d(a_5 - b_4)$