More Word Alignment

David Kauchak
CS159 – Fall 2020

Some slides adapted from

Philipp Koehn
School of Informatics
University of Edinburgh

Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

Dan Klein
Computer Science Department
UC Berkeley

1

Admin

Assignment 5a solutions posted

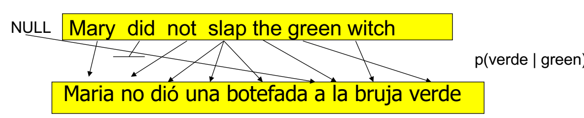
Assignment 5b due Monday

2

Language translation

3

Word models: IBM Model 1



Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

4

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

$$p(f_i \mid e_{a_i}) = \frac{count(f \ aligned\text{-}to \ e)}{count(e)}$$

p(el | the) = 0.5
p(Los | the) = 0.5

Any problems concerns?

5

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

Getting data like this is expensive!

Even if we had it, what happens when we switch to a new domain/corpus

6

Training without alignments

a b

x y

How should these be aligned?

c b

z x

There is some information!
(Think of the alien translation task last time)

7

Training without alignments

a b

x y

IBM model 1: Each foreign word (bottom) is aligned to one English word (ignore NULL for now)

What are the possible alignments?

8

Training without alignments

IBM model 1: Each foreign word is aligned to one English word

9

Training without alignments

0.01 0.9 0.08 0.01

IBM model 1: Each foreign word is aligned to 1 English word

If I told you how likely each of these were, does that help us with calculating p(f | e)?

10

Training without alignments

0.01 0.9 0.08 0.01

IBM model 1: Each foreign word is aligned to 1 English word

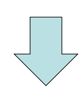
$$p(f_i | e_{a_i}) = \frac{count(f\ aligned\text{-}to\ e)}{count(e)}$$

Use partial counts and sum:
- count(y | a) 0.9+0.01
- count(x | a) 0.01+0.01

11

One the one hand

0.01 0.9 0.08 0.01



If you had the likelihood of each alignment, you could calculate p(f|e)

$$p(f_i | e_{a_i}) = \frac{count(f\ aligned\text{-}to\ e)}{count(e)}$$

12

One the other hand

$$p(F, a_1 a_2 \ldots a_{|F|} \mid E) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

If you had p(f|e) could you calculate the probability of the alignments?

$$p(f_i \mid e_{a_i})$$

13

One the other hand

We want to calculate the probability of the alignment, e.g.

$$p(alignment1|F,E) = p(A_1|F,E)$$

We can calculate $p(A_1, F \mid E)$ using the word probabilities.

$$p(F, a_1 a_2 \ldots a_{|F|} \mid E) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

14

One the other hand

We want to calculate the probability of the alignment, e.g.

$$p(alignment1|F,E) = p(A_1|F,E)$$

We can calculate $p(A_1, F \mid E)$ using the word probabilities.

$$p(A, F|E) \ ? \ p(A|F,E)$$

How are these two probabilities related?

15

Our friend the chain rule

$$p(A_1, F|E) = p(A_1|F,E) * p(F|E)$$

$$p(A_1|F,E) = \frac{p(A_1,F|E)}{P(F|E)}$$

What is P(F|E)?

Hint: how do we go from $p(A_1,F|E)$ to P(F|E)?

16

Our friend the chain rule

$p(A_1, F|E) = p(A_1|F, E) * p(F|E)$

$p(A_1|F, E) = \frac{p(A_1,F|E)}{P(F|E)}$

$p(A_1|F, E) = \frac{p(A_1,F|E)}{\sum_A P(A,F|E)}$

sum over the variable!

How likely is this alignment, compared to all other alignments under the model

17

One the other hand

Alignment 1	Alignment 2	Alignment 3	Alignment 4
a b / x y	a b / x y	a b / x y	a b / x y
$p(x\|a)*p(y\|b)$	$p(x\|b)*p(y\|a)$	$p(x\|b)*p(y\|b)$	$p(x\|a)*p(y\|a)$
$p(F,a_1\|E)$	$p(F,a_2\|E)$	$p(F,a_3\|E)$	$p(F,a_4\|E)$

$$p(F, a_1 a_2 ... a_{|F|} | E) = \prod_{i=1}^{|F|} p(f_i | e_{a_i})$$

18

One the other hand

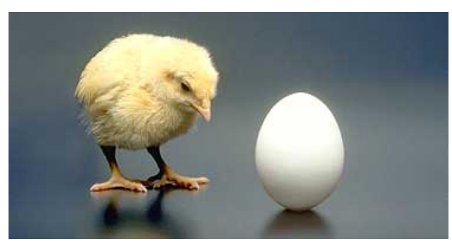
Alignment 1	Alignment 2	Alignment 3	Alignment 4
a b / x y	a b / x y	a b / x y	a b / x y
$p(x\|a)*p(y\|b)$	$p(x\|b)*p(y\|a)$	$p(x\|b)*p(y\|b)$	$p(x\|a)*p(y\|a)$
$p(F,a_1\|E)$	$p(F,a_2\|E)$	$p(F,a_3\|E)$	$p(F,a_4\|E)$

Normalize

$$p(a_1|E,F) = \frac{p(x|a) * p(y|b)}{\sum_{i=1}^{4} p(F, ai|E)}$$

19

Have we gotten anywhere?



20

Training without alignments

Initially assume a p(f|e) are equally probable

Repeat:

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))
- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

(Note: theoretical algorithm)

21

EM algorithm

(*something from nothing*)

General approach for calculating "hidden variables", i.e. variables without explicit labels in the data

Repeat:

E-step: Calculate the expected probabilities of the hidden variables based on the current model

M-step: Update the model based on the expected counts/probabilities

22

EM alignment

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

23

green house

casa verde

the house

la casa

What are the different p(f|e) that make up my model?

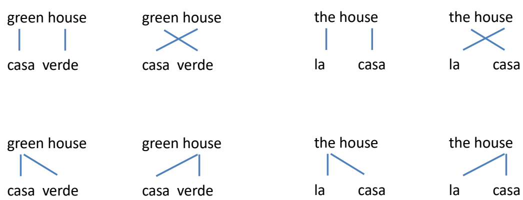
p(casa | green)	
p(verde | green)	
p(la | green)	

p(casa | house)	
p(verde | house)	
p(la | house)	

p(casa | the)	
p(verde | the)	
p(la | the)	

Technically, all combinations of foreign and English words

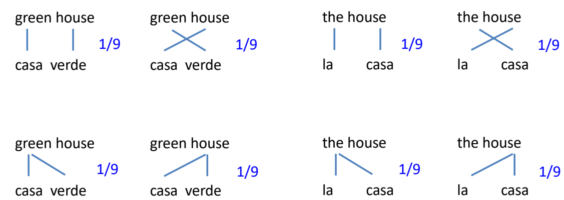
24



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

Start with all p(f|e) equally probable

25

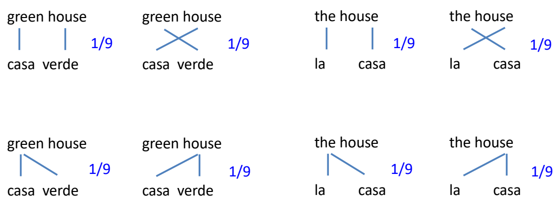


p(casa | green)	1/3	p(casa | house)	1/3	p(casa | the)	1/3
p(verde | green)	1/3	p(verde | house)	1/3	p(verde | the)	1/3
p(la | green)	1/3	p(la | house)	1/3	p(la | the)	1/3

E-step: What are the probabilities of the alignments?

1. calculate: $p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$

26



sum = 4/9

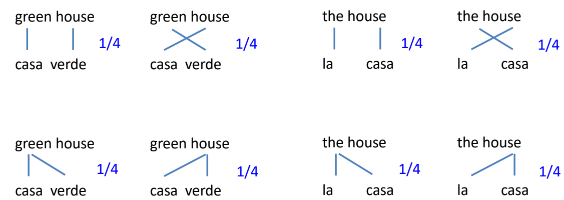
sum = 4/9

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

E-step: What are the probabilities of the alignments?

2. normalize: $p(a_i|E,F) = \frac{p(F,ai|E)}{\sum_{j=1}^{4} p(F,aj|E)}$

27

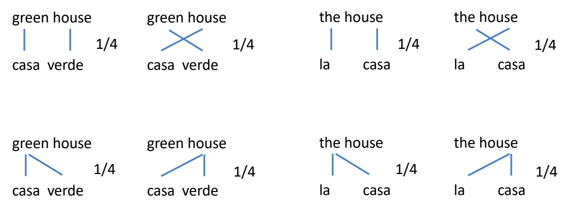


p(casa | green)	1/3	p(casa | house)	1/3	p(casa | the)	1/3
p(verde | green)	1/3	p(verde | house)	1/3	p(verde | the)	1/3
p(la | green)	1/3	p(la | house)	1/3	p(la | the)	1/3

E-step: What are the probabilities of the alignments?

2. normalize: $p(a_i|E,F) = \frac{p(F,ai|E)}{\sum_{j=1}^{4} p(F,aj|E)}$

28



M-step: What are the p(f|e) given the alignments?

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

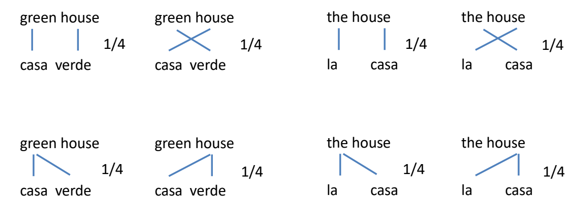
c(casa,green) = ?
c(verde,green) = ?
c(la, green) = ?

c(casa,house) ?
c(verde,house) = ?
c(la,house) = ?

c(casa,the) = ?
c(verde,the) = ?
c(la,the) = ?

First, calculate the partial counts

29



M-step: What are the p(f|e) given the alignments?

p(casa | green)	?	p(casa | house)	?	p(casa | the)	?
p(verde | green)	?	p(verde | house)	?	p(verde | the)	?
p(la | green)	?	p(la | house)	?	p(la | the)	?

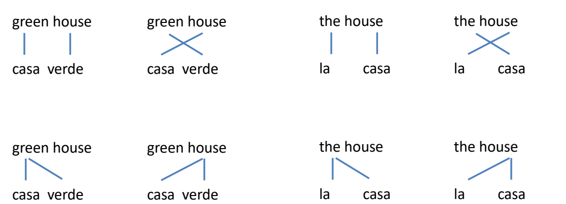
c(casa,green) = 1/4+1/4 = 1/2
c(verde,green) = 1/4+1/4 = 1/2
c(la, green) = 0

c(casa,house) = 1/4+1/4+
1/4+1/4 = 1
c(verde,house) = 1/4+1/4 = 1/2
c(la,house) = 1/4+1/4 = 1/2

c(casa,the) = 1/4+1/4 = 1/2
c(verde,the) = 0
c(la,the) = 1/4+1/4 = 1/2

Then, calculate the probabilities by normalizing the counts

30



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

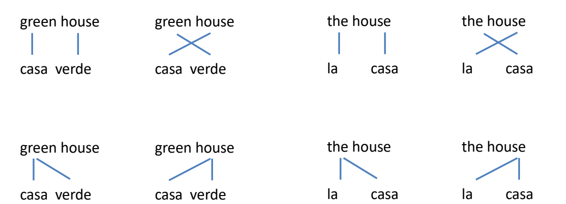
c(casa,green) = 1/4+1/4 = 1/2
c(verde,green) = 1/4+1/4 = 1/2
c(la, green) = 0

c(casa,house) = 1/4+1/4+
1/4+1/4 = 1
c(verde,house) = 1/4+1/4 = 1/2
c(la,house) = 1/4+1/4 = 1/2

c(casa,the) = 1/4+1/4 = 1/2
c(verde,the) = 0
c(la,the) = 1/4+1/4 = 1/2

Then, calculate the probabilities by normalizing the counts

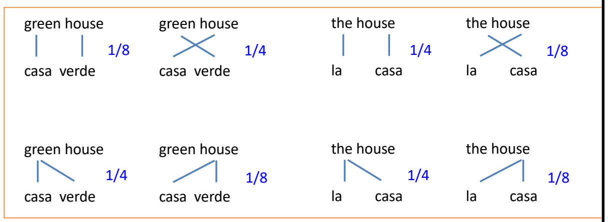
31



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

E-step: 1. what are the p(A,F|E)?

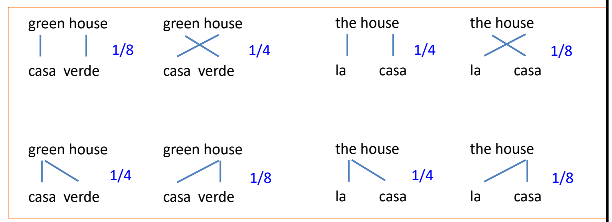
32



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

E-step: 1. what are the p(A,F|E)?

33



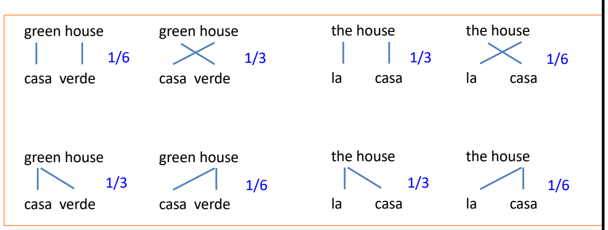
sum = (3/4) **sum = (3/4)**

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

$$p(a_i|E,F) = \frac{p(F,ai|E)}{\sum_{j=1}^{4} p(F,aj|E)}$$

E-step: 2. what are the alignments, i.e. normalize?

34



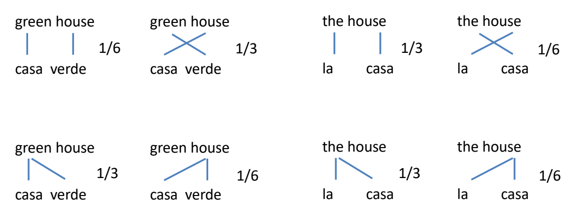
sum = (3/4) **sum = (3/4)**

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

$$p(a_i|E,F) = \frac{p(F,ai|E)}{\sum_{j=1}^{4} p(F,aj|E)}$$

E-step: 2. what are the alignments, i.e. normalize?

35



M-step: What are the p(f|e) given the alignments?

p(casa | green)		p(casa | house)		p(casa | the)	
p(verde | green)		p(verde | house)		p(verde | the)	
p(la | green)		p(la | house)		p(la | the)	

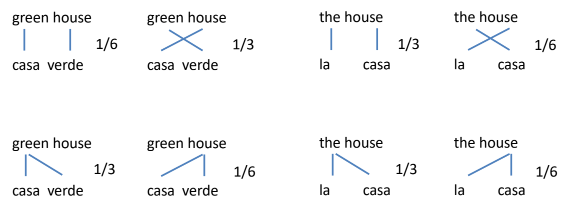
c(casa,green) = ?
c(verde,green) = ?
c(la, green) = ?

c(casa,house) = ?
c(verde,house) = ?
c(la,house) = ?

c(casa,the) = ?
c(verde,the) = ?
c(la,the) = ?

First, calculate the partial counts

36



M-step: What are the p(f|e) given the alignments?

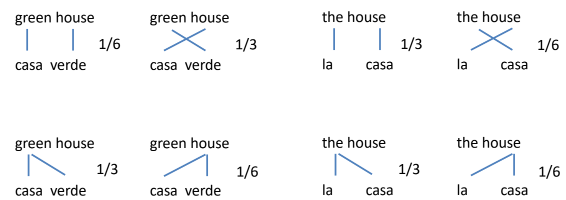
p(casa | green)		p(casa | house)		p(casa | the)	
p(verde | green)		p(verde | house)		p(verde | the)	
p(la | green)		(la | house)		p(la | the)	

c(casa,green) = 1/6+1/3 = 3/6
c(verde,green) = 1/3+1/3 = 4/6
c(la, green) = 0

c(casa,house) = 1/3+1/6+ 1/3+1/6 = 6/6
c(verde,house) = 1/6+1/6 = 2/6
c(la,house) = 1/6+1/6 = 2/6

c(casa,the) = 1/6+1/3 = 3/6
c(verde,the) = 0
c(la,the) = 1/3+1/3 = 4/6

37



M-step: What are the p(f|e) given the alignments?

p(casa | green)	?	p(casa | house)	?	p(casa | the)	?
p(verde | green)	?	p(verde | house)	?	p(verde | the)	?
p(la | green)	?	(la | house)	?	p(la | the)	?

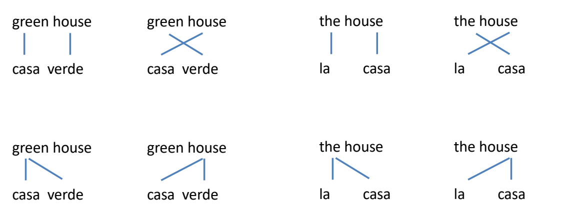
c(casa,green) = 1/6+1/3 = 3/6
c(verde,green) = 1/3+1/3 = 4/6
c(la, green) = 0

c(casa,house) = 1/3+1/6+ 1/3+1/6 = 6/6
c(verde,house) = 1/6+1/6 = 2/6
c(la,house) = 1/6+1/6 = 2/6

c(casa,the) = 1/6+1/3 = 3/6
c(verde,the) = 0
c(la,the) = 1/3+1/3 = 4/6

Then, calculate the probabilities by normalizing the counts

38



p(casa | green)	3/7	p(casa | house)	3/5	p(casa | the)	3/7
p(verde | green)	4/7	p(verde | house)	1/5	p(verde | the)	0
p(la | green)	0	p(la | house)	1/5	p(la | the)	4/7

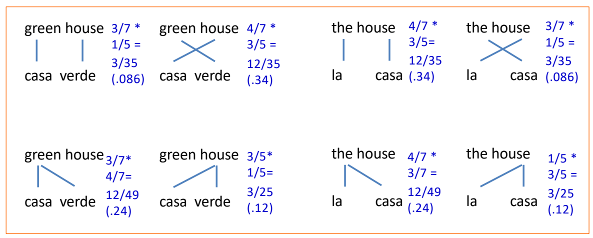
c(casa,green) = 1/6+1/3 = 3/6
c(verde,green) = 1/3+1/3 = 4/6
c(la, green) = 0

c(casa,house) = 1/3+1/6+ 1/3+1/6 = 6/6
c(verde,house) = 1/6+1/6 = 2/6
c(la,house) = 1/6+1/6 = 2/6

c(casa,the) = 1/6+1/3 = 3/6
c(verde,the) = 0
c(la,the) = 1/3+1/3 = 4/6

Then, calculate the probabilities by normalizing the counts

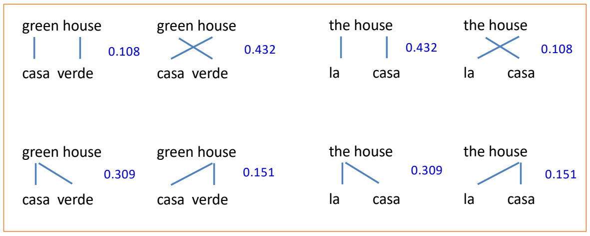
39



p(casa | green)	3/7	p(casa | house)	3/5	p(casa | the)	3/7
p(verde | green)	4/7	p(verde | house)	1/5	p(verde | the)	0
p(la | green)	0	p(la | house)	1/5	p(la | the)	4/7

E-step: 1. what are the p(A,F|E)?

40



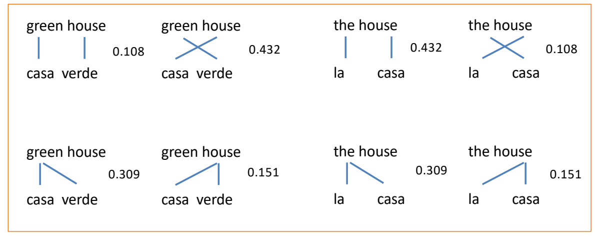
p(casa | green)	3/7
p(verde | green)	4/7
p(la | green)	0

p(casa | house)	3/5
p(verde | house)	1/5
p(la | house)	1/5

p(casa | the)	3/7
p(verde | the)	0
p(la | the)	4/7

E-step: 2. what are the alignments, i.e. normalize?

41



p(casa | green)	3/7
p(verde | green)	4/7
p(la | green)	0

p(casa | house)	3/5
p(verde | house)	1/5
p(la | house)	1/5

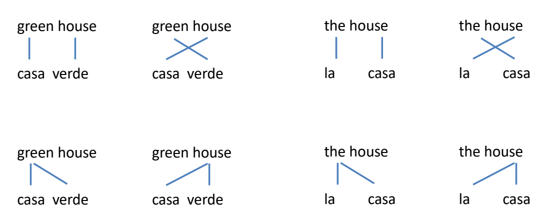
p(casa | the)	3/7
p(verde | the)	0
p(la | the)	4/7

c(casa,green) = .108+.309=0.417
c(verde,green) = .432+.309 = 0.741
c(la, green) = 0

c(casa,house) = .432+.151+ .432+.151=1.166
c(verde,house) = .108+.151=.259
c(la,house) = .108+.151=0.259

c(casa,the) = .108+.309=0.417
c(verde,the) = 0
c(la,the) = .432+.309=0.741

42



p(casa | green)	0.36
p(verde | green)	0.64
p(la | green)	0

p(casa | house)	0.69
p(verde | house)	0.15
p(la | house)	0.15

p(casa | the)	0.36
p(verde | the)	0
p(la | the)	0.64

c(casa,green) = .108+.309=0.417
c(verde,green) = .432+.309 = 0.741
c(la, green) = 0

c(casa,house) = .432+.151+ .432+.151=1.166
c(verde,house) = .108+.151=.259
c(la,house) = .108+.151=0.259

c(casa,the) = .108+.309=0.417
c(verde,the) = 0
c(la,the) = .432+.309=0.741

43

Iterate…

5 iterations

p(casa | green)	0.24
p(verde | green)	0.76
p(la | green)	0

p(casa | house)	0.84
p(verde | house)	0.08
p(la | house)	0.08

p(casa | the)	0.24
p(verde | the)	0
p(la | the)	0.76

10 iterations

p(casa | green)	0.1
p(verde | green)	0.9
p(la | green)	0

p(casa | house)	0.98
p(verde | house)	0.01
p(la | house)	0.01

p(casa | the)	0.1
p(verde | the)	0
p(la | the)	0.9

100 iterations

p(casa | green)	0.005
p(verde | green)	0.995
p(la | green)	0

p(casa | house)	~1.0
p(verde | house)	~0.0
p(la | house)	~0.0

p(casa | the)	0.005
p(verde | the)	0
p(la | the)	0.995

44

EM alignment

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

Why does it work?

45

EM alignment

E-ste

–

– der the

M-st

– ıents,

Why does it work?

46

EM alignment

Intuitively:

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

Things that co-occur will have higher probabilities

E-step

- Calculate how probable the alignments are under the current model (i.e. p(f|e))

Alignments that contain things with higher p(f|e) will be scored higher

47

An aside: estimating probabilities

What is the probability of "the" occurring in a sentence?

$$\frac{\text{number of sentences with "the"}}{\text{total number of sentences}}$$

Is this right?

48

Estimating probabilities

What is the probability of "the" occurring in a sentence?

$$\frac{\text{number of sentences with "the"}}{\text{total number of sentences}}$$

No. This is an *estimate* based on our data

This is called the maximum likelihood estimation. Why?

49

Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation picks the values for the model parameters that maximize the likelihood of the training data

You flip a coin 100 times. 60 times you get heads.

What is the MLE for heads?

p(head) = 0.60

50

Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation picks the values for the model parameters that maximize the likelihood of the training data

You flip a coin 100 times. 60 times you get heads.

What is the likelihood of the data under this model (each coin flip is a data point)?

51

MLE example

You flip a coin 100 times. 60 times you get heads.

MLE for heads: p(head) = 0.60

What is the likelihood of the data under this model (each coin flip is a data point)?

$$likelihood(data) = \prod_i p(x_i)$$

$\log(0.60^{60} * 0.40^{40}) = -67.3$

52

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

p(heads) = 0.5

log(0.50^{60} * 0.50^{40}) =-69.3

p(heads) = 0.7

– log(0.70^{60} * 0.30^{40})=-69.5

53

EM alignment: the math

The EM algorithm tries to find parameters of the model (p(f|e)) that *maximize the likelihood of the data*

In our case:

$$p(f_1 f_2 \ldots f_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \sum_{a_1} \sum_{a_2} \ldots \sum_{a_{|F|}} p(f_i \mid e_{a_i})$$

Each iteration, we increase (or keep the same) the likelihood of the data

54

Implementation details

Any concerns/issues?
Anything underspecified?

Repeat:

E-step
- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step
- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

55

Implementation details

When do we stop?

Repeat:

E-step
- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step
- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

56

Implementation details

- Repeat for a fixed number of iterations
- Repeat until parameters don't change (much)
- Repeat until likelihood of data doesn't change much

Repeat:

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

57

Implementation details

For |E| English words and |F| foreign words, how many alignments are there?

Repeat:

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

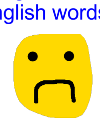
- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

58

Implementation details

Each foreign word can be aligned to any of the English words (or NULL)

$(|E|+1)^{|F|}$



Repeat:

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

59

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

$$p(f_i \mid e_{a_i}) = \frac{count(f\ aligned\text{-}to\ e)}{count(e)}$$

p(el | the) = 0.5
p(Los | the) = 0.5

60

If we had the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

```
for (E, F) in corpus:
    for aligned words (e, f) in pair (E,F):
        count(e,f) += 1
        count(e) += 1

for all (e,f) in count:
    p(f|e) = count(e,f) / count(e)
```

61

If we had the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

```
for (E, F) in corpus:
    for e in E:
        for f in F:
            if f aligned-to e:
                count(e,f) += 1
                count(e) += 1

for all (e,f) in count:
    p(f|e) = count(e,f) / count(e)
```

62

If we had the alignments

Input: corpus of English/Foreign sentence pairs along with alignment

```
for (E, F) in corpus:
    for aligned words (e, f) in pair (E,F):
        count(e,f) += 1
        count(e) += 1
```

```
for (E, F) in corpus
    for e in E:
        for f in F:
            if f aligned-to e:
                count(e,f) += 1
                count(e) += 1
```

Are these equivalent?

```
for all (e,f) in count:
    p(f|e) = count(e,f) / count(e)
```

63