CA4012

Statistical Machine Translation

Week 5: Translation Modeling

Lecturer: Dimitar Shterionov

E-mail: (dimitar.shterionov@adaptcentre.ie)

Lab tutors: Eva Vanmassenhove, Alberto Poncelas

2nd Semester, 2018-2019 Academic Year

Content

- **Introduction to Translation Models**
- Word-based Translation Model
- Word Alignment
- Estimation of Word-based Translation Model
- Exercises

Recap & Quiz

- What is the purpose of the language model?

Recap & Quiz

- What is the purpose of the language model?
- The purpose of a language model is to identify what is considered a good sentence in the target language
- That is, it measures the probability p(e) of a sentence e being a fluent sentence.

Recap & Quiz

- How do you build a language model?

Recap & Quiz

- How do you build a language model?
- Sentences in a corpus are broken down to pieces (n-grams), and then probabilities of n-grams are calculated based on their counts (MLE) – n-gram language modelling.
- This process is done based on Markov process assumptions, i.e. the probability of word w_n only depends on the previous *n-1* words.

Recap & Quiz

- What is the term given to the process of accounting for unseen n-grams in language modelling? Why is this process necessary?
- The process of accounting for unseen n-grams is known as smoothing.
- It's necessary because otherwise too many n-grams would be assigned zero probability and this will affect the calculation of the sentence probability. We cannot conclude that an n-gram has zero probability of occurring just because it hasn't been seen in the training corpus.

Recap & Quiz

- How can we say that a language model is good or bad?
- This question is about the evaluation of a language model. Perplexity (PP) is often used to evaluate a language model on a given test data. The higher the PP is, the worse the language model is.
- PP is a measure using cross-entropy of the model that can be expressed as:

$$PP = 2^{H(P_{LM})} = 2^{-\frac{1}{n}\log(p(w_1 w_2 \ldots w_n))}$$

Content

Introduction to Translation Models

Word-based Translation Model

Word Alignment

Estimation of Word-based Translation Model

Exercises

Noisy Channel Model Revisited

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f)$$

Noisy Channel Model Revisited

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f)$$

$$\hat{e} = \underset{e}{\operatorname{argmax}} \frac{p(f|e)p(e)}{p(f)}$$

Noisy Channel Model Revisited

$$\hat{e} = \operatorname*{argmax}_{e} p(e|f)$$

$$\hat{e} = \operatorname*{argmax}_{e} \frac{p(f|e)p(e)}{p(f)}$$

$$\hat{e} = \operatorname*{argmax}_{e} \frac{p(f|e)p(e)}{\cancel{p(f)}}$$

$$\hat{e} = \operatorname*{argmax}_{e} p(e)p(f|e)$$

Noisy Channel Model Revisited

$$\hat{e} = \underbrace{\operatorname*{argmax}_{e}}\, \underbrace{p(e)}\, \underbrace{p(f|e)}$$

Decoder

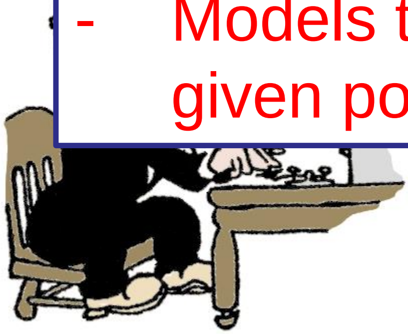
Translation Model

Language Model

Noisy Channel Model Revisited

$$\hat{e} = \underbrace{\operatorname*{argmax}_{e}}\,\underbrace{p(e)}\,\underbrace{p(f|e)}$$

The purpose of the translation model
- Models statistically the process of translation
- Encodes the faithfulness of e as a translation of f
- Models the probability of the foreign sentence given possible translations



odel

Language Model

Translation Model

- We are translating a foreign-language sentence **f** into native-language (English) sentence **e**
- Given any English sentence **e** and any foreign sentence **f**, we define the probability that **f** is a translation of **e** as: $p(\mathbf{f}|\mathbf{e})$

where the normalization condition is:

$$\sum_{\boldsymbol{f}} p(\mathbf{f}|\mathbf{e}) = 1$$

Content

- Introduction to Translation Models
- **Word-based Translation Model**
- Word Alignment
- Estimation of Word-based Translation Model
- Exercises

Estimation of Sentence Probabilities

I love the boy.
J'aime le garcon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent a la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the dog.
Je parle au garcon.

Estimation of Sentence Probabilities

I love the boy.
J'aime le garcon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent a la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the dog.
Je parle au garcon.

EXERCISE:
What is the probability of

p(J'aime le garcon|I love the boy)

p(J'aime la fille|I love the girl)

Estimation of Sentence Probabilities

I love the boy.
J'aime le garcon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent a la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the dog.
Je parle au garcon.

EXERCISE:
What is the probability of

p(J'aime le garcon|I love the boy)
= 1/6

p(J'aime la fille|I love the girl)
= 0/6

Estimation of Sentence Probabilities

Recall - Similar Problem:

How can we estimate the probability of a sentence in a specific language?

Similar Idea:

Decompose a sentence into words and estimate translation probabilities at the word level.

Lexical (Word) Translation

- How to translate a word?

Lexical (Word) Translation

- How to translate a word?
  - Dictionary look up?
    - (DE→EN) Haus: house, building, home, household, shell

Lexical (Word) Translation

- How to translate a word?
  - Dictionary look up?
    - (DE→EN) Haus: house, building, home, household, shell
  - Multiple translations: some more frequent than others
- How do we determine probabilities for possible candidate translations?

Lexical (Word) Translation

- How to translate a word?
  - Dictionary look up?
    - (DE→EN) Haus: house, building, home, household, shell
  - Multiple translations: some more frequent than others
- How do we determine probabilities for possible candidate translations?
- Collect statistics from a parallel corpus:

Translation of Haus	Count
house	8,000
building	1,600
home	200
household	150
shell	50

Estimate Translation Probabilities

Translation of Haus	Count
house	8,000
building	1,600
home	200
household	150
shell	50
Total	**10,000**

- Use relative frequencies to estimate probabilities of

$$p(house|Haus) = ?$$
$$p(building|Haus) = ?$$
$$p(home|Haus) = ?$$
$$p(household|Haus) = ?$$
$$p(shell|Haus) = ?$$

Estimate Translation Probabilities

Translation of Haus	Count
house	8,000
building	1,600
home	200
household	150
shell	50
Total	**10,000**

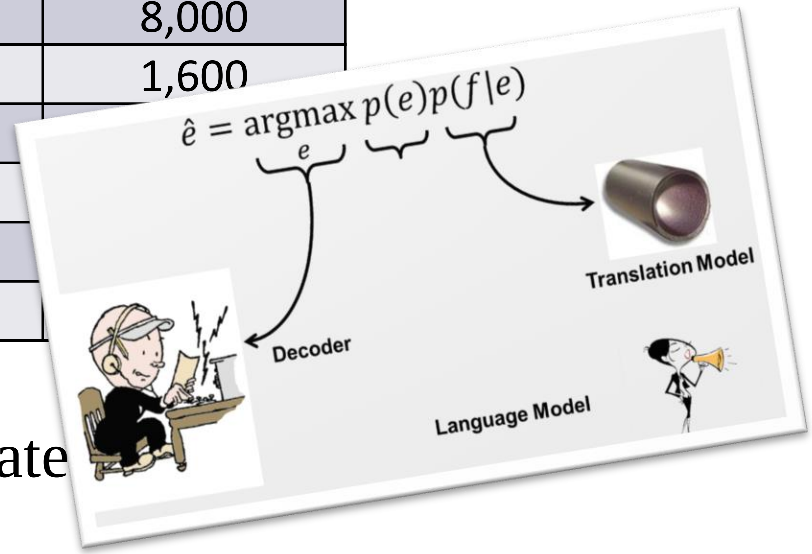
- Use relative frequencies to estimate probabilities of

$p(\text{house}|\text{Haus}) = \mathit{8000/10000}=0.8$
$p(\text{building}|\text{Haus}) = \mathit{1600/10000}=0.16$
$p(\text{home}|\text{Haus}) = \mathit{200/10000}=0.02$
$p(\text{household}|\text{Haus}) = \mathit{150/10000}=0.015$
$p(\text{shell}|\text{Haus}) = \mathit{50/10000}=0.005$

MLE: Maximum Likelihood Estimation

Estimate Translation Probabilities

Translation of Haus	Count
house	8,000
building	1,600
home	
household	
shell	
Total	



- Use relative frequencies to estimate

p(house|Haus) = *8000/10000*=0.8
p(building|Haus) = *1600/10000*=0.16
p(home|Haus) = *200/10000*=0.02
p(household|Haus) = *150/10000*=0.015
p(shell|Haus) = *50/10000*=0.005

MLE: Maximum Likelihood Estimation

Estimate Translation Probabilities

Translation of House	Count
Haus	8,000
Haushalt	2,000
Vorstellung	1000
Gebauede	800
Geschlecht	200
Total	**12,000**

- Use relative frequencies to estimate probabilities of

$$p(\text{Haus}|\text{house}) = ?$$

How would we get $p(\text{Haus}|\text{building})$?

Estimate Translation Probabilities

We estimate translation probabilities from a parallel sentence-aligned corpus.

Estimate Translation Probabilities

We estimate translation probabilities from a parallel sentence-aligned corpus.

However!

Sentences are aligned. Words are not.

We need to know which words in the source are aligned to which words in the target before we can count the co-occurrences and calculate the probabilities.

Content

Conventions

- For better understanding, notations and formulas are kept consistent with the textbook.
- By convention, given a sentence-aligned text, we refer to the input as the foreign language, and the output language as English.
- By convention, we use $p(\mathbf{e}|\mathbf{f})$ to represent the translation model although it is $p(\mathbf{f}|\mathbf{e})$ in the formula of SMT model.

Some Notations

Given a sentence-aligned text, we have the following notations:

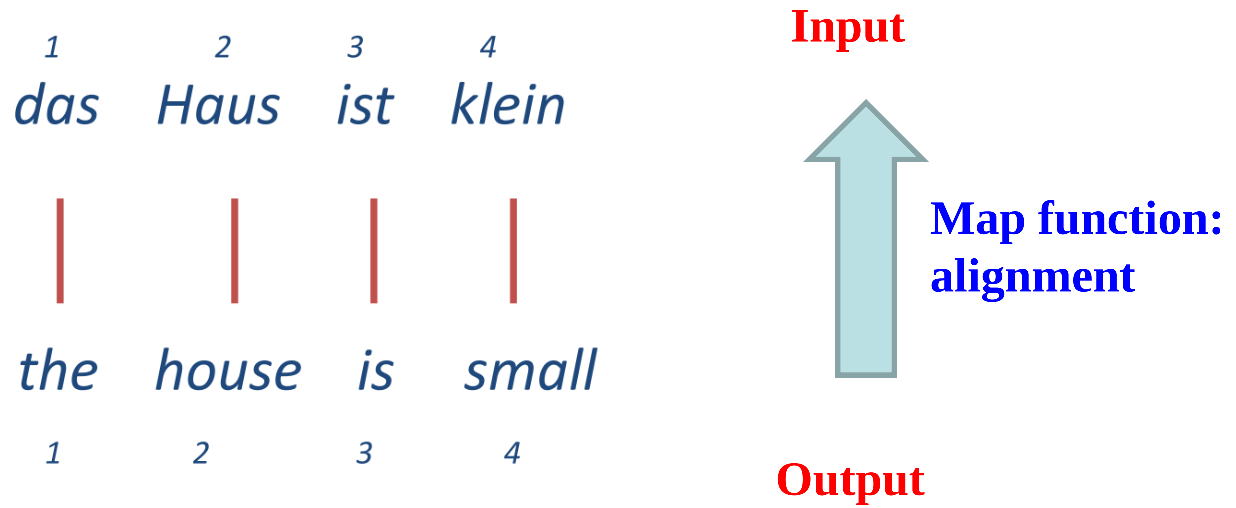
Source **f**: Foreign (e.g. German)
Target **e**: English
f: a word in **f**
e: a word in **e**
$l_{\mathbf{e}}$: length of **f** (i=1…)
$l_{\mathbf{f}}$: length of **e** (j=1…)
$t(e|f;\, \mathbf{e},\mathbf{f})$: lexical translation probability
Alignment: $a: j \rightarrow i$ or $a_j = i$

Alignment: Monotone One-to-One

Given a sentence-aligned text, we align words in one text with words in another.

Alignment: Monotone One-to-One

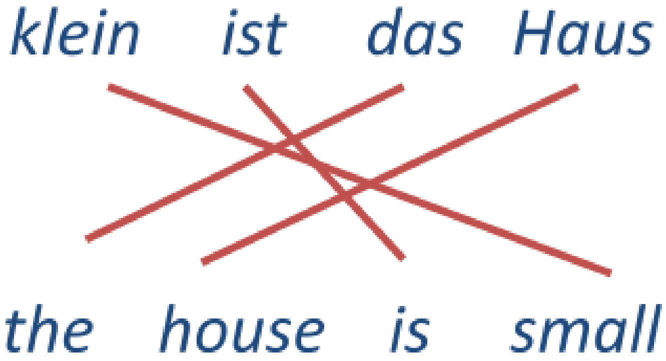
Given a sentence-aligned text, we align words in one text with words in another.



Alignment function, a $\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$

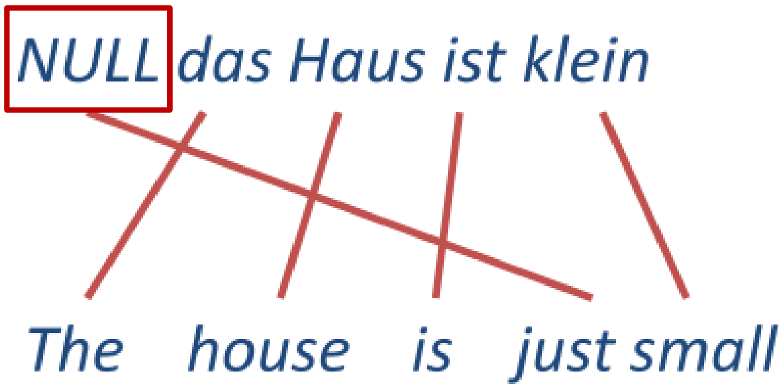
or: $\{a_j = i \mid j=1,\ldots,l_{\mathbf{e}}\,; i=1,\ldots,l_{\mathbf{f}}\}$

Alignment: Reordering



Alignment function, a $\{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$

Alignment: Spurious Words



Alignment function, a $\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$

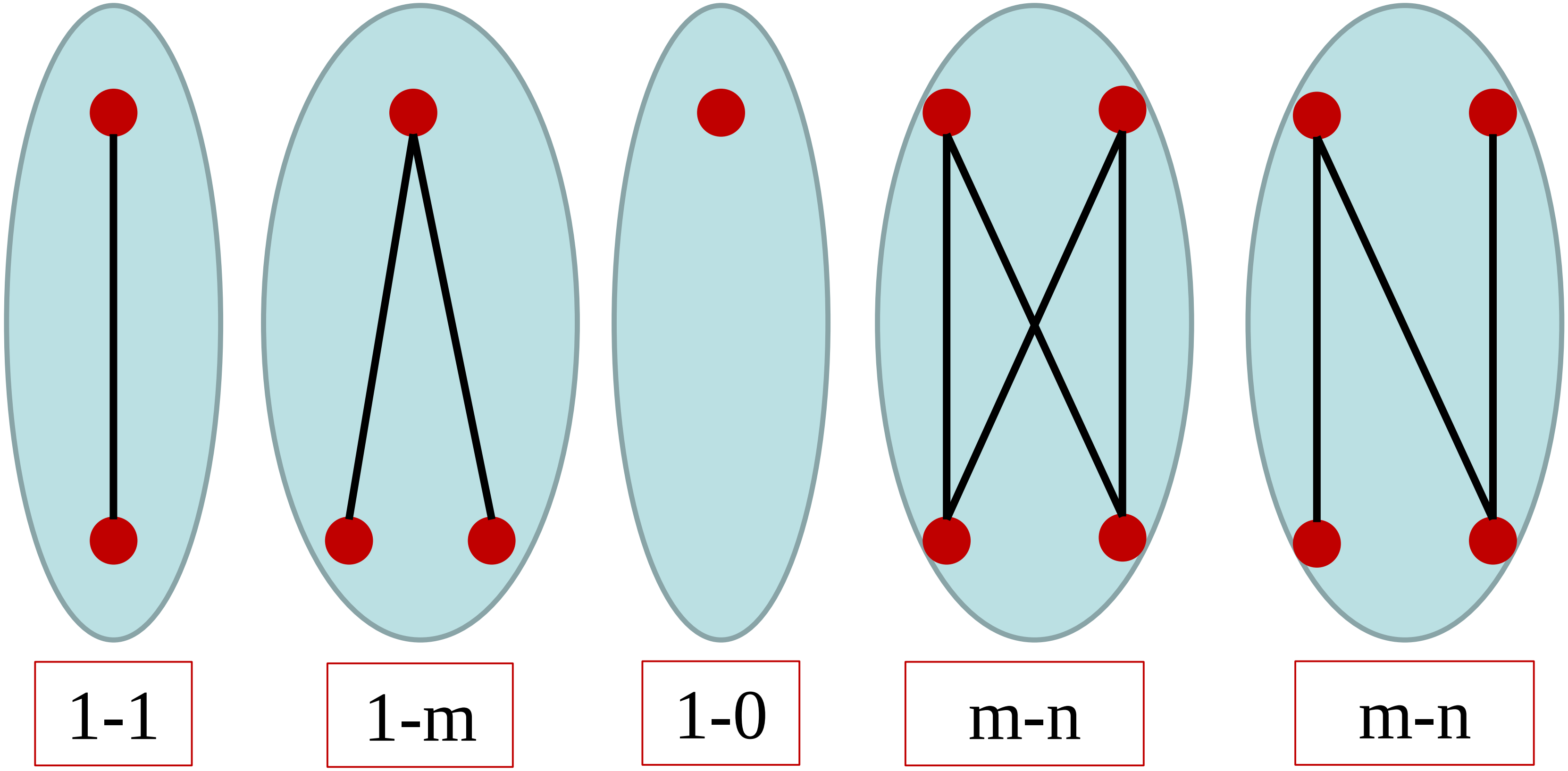
Alignment: Dropping Words

Alignment function, a $\{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$

Alignment: Many-to-One

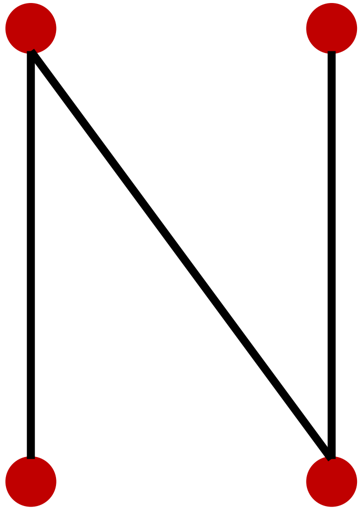
Alignment function, a $\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$

Word Alignment Patterns

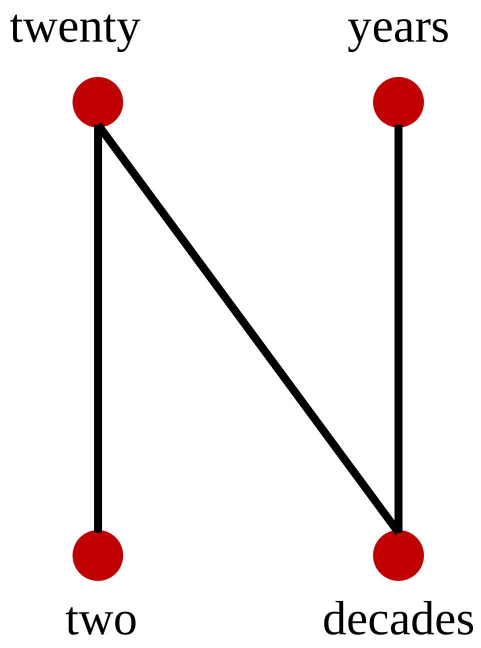


An Example of M-N

Can you image an example of this alignment pattern in English?



An Example of M-N



Content

Learning Lexical Translation Models

- We would like to estimate the lexical translation probabilities from a parallel corpus... But
- We do not have the alignments.

Learning Lexical Translation Models

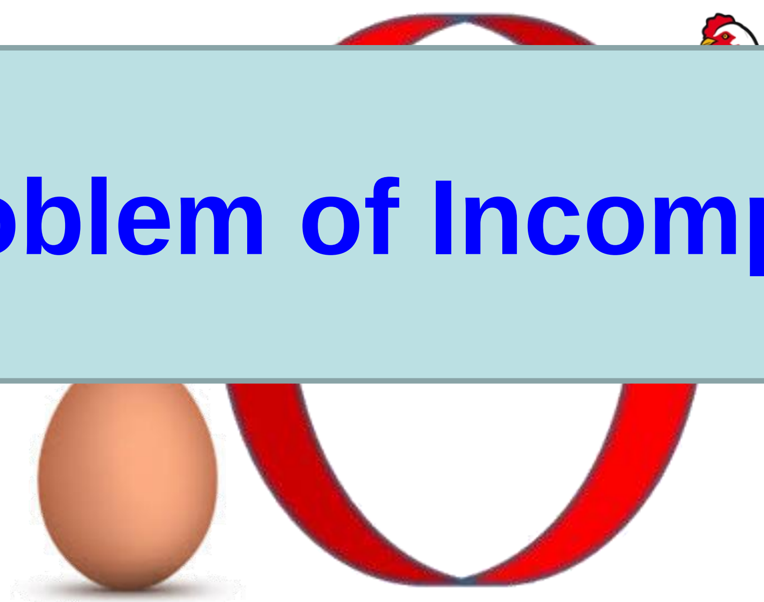
- We would like to estimate the lexical translation probabilities from a parallel corpus... But
- We do not have the alignments
  - If we had the alignments, we could estimate the lexical translation probabilities.

Learning Lexical Translation Models

- We would like to estimate the lexical translation probabilities from a parallel corpus... But
- We do not have the alignments
  - If we had the alignments, we could estimate the lexical translation probabilities.
  - If we had the probabilities, we could estimate the alignments.

Learning Lexical Translation Models

Paradox



Problem of Incomplete Data

EM Algorithm

- Incomplete data
  - if we had complete data, we could estimate the model
  - if we had the model, we could fill in the gaps in the data

EM Algorithm

- Incomplete data
  - if we had complete data, we could estimate the model
  - if we had the model, we could fill in the gaps in the data
- Expectation Maximization (EM) in a nutshell
  - initialise model parameters (e.g. uniform, random)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate

How does EM work?

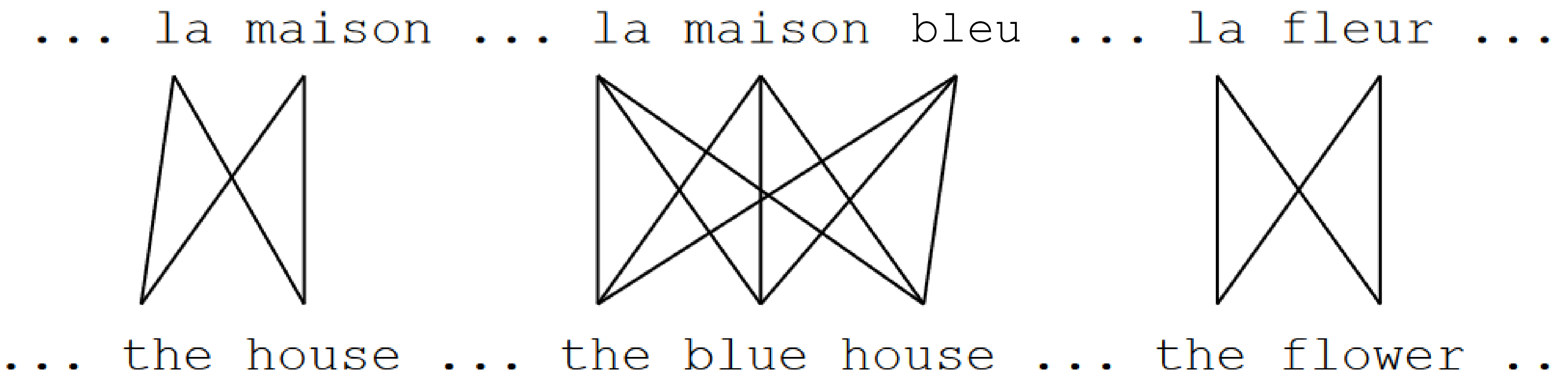
EM Algorithm consists of two steps

1. Expectation-Step: Apply model to the data
   - parts of the data are hidden (here: alignments)
   - using the model, assign probabilities of the hidden data to possible values (alignments)
2. Maximization-Step: Estimate a new model from data
   - take assigned values as <u>fact</u>
   - collect counts (weighted by probabilities)
   - estimate new model from counts

Iterate these steps until convergence

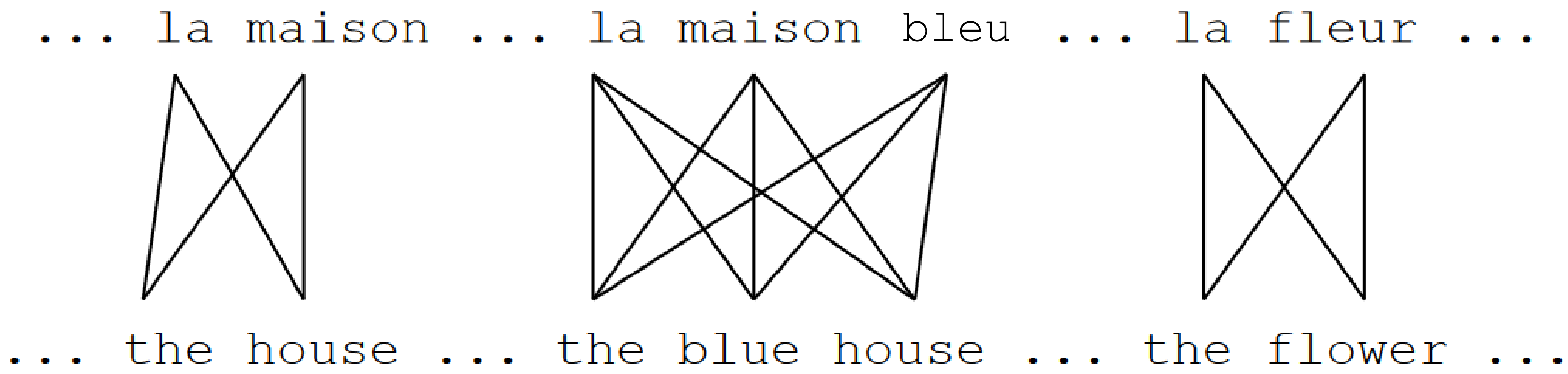
EM: Basic idea in practice

1. Initialize: all alignments are equally likely



EM: Basic idea in practice

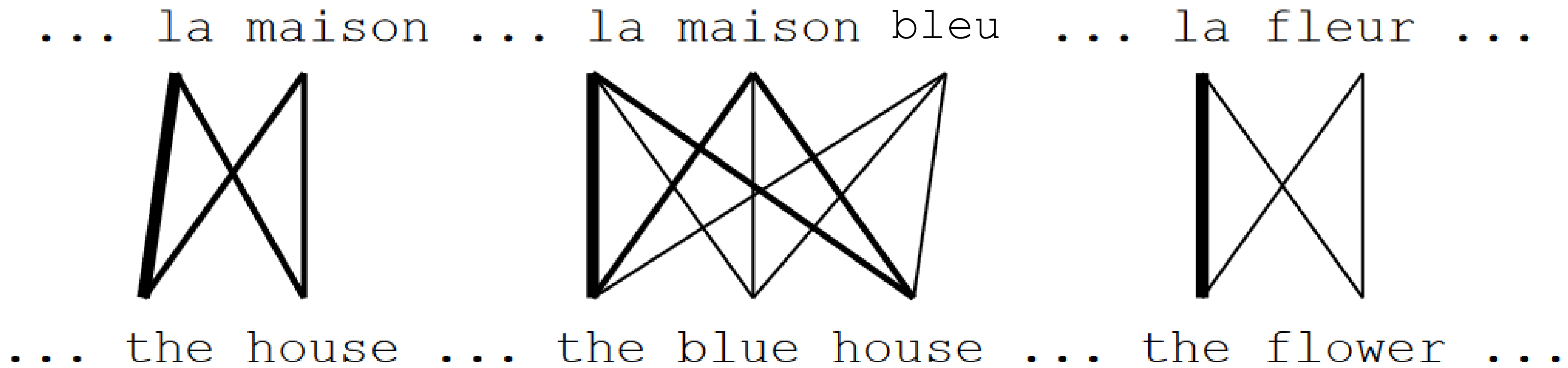
1. Initialize: all alignments are equally likely



2. Pass once and learn that la is often aligned with the

EM: Basic idea in practice

1. After one iteration alignments between la and the are more likely



2. Pass once more. It becomes apparent that alignments, e.g., between fleur and flower are more likely.

EM: Basic idea in practice

1. After n more iterations - convergence

2. Alignment and word-translation probabilities

The aim of EM

- Alignment: $p(a|\mathbf{e},\mathbf{f})$
- Translation probabilities: $t(e_j|f_{a(j)})$

EM Formularization

EM Algorithm consists of two steps

1. Expectation-Step: Apply model to the data and compute:

$$p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})} = \frac{p(\mathbf{e},a|\mathbf{f})}{\sum_a p(\mathbf{e},a|\mathbf{f})}$$

- How many alignments are there between **e** of length $l_{\mathbf{e}}$ and **f** of length $l_{\mathbf{f}}$?
- ε – a normalization constant.

$$p(\mathbf{e},a|\mathbf{f}) = \frac{\varepsilon}{(l_{\mathbf{f}}+1)^{l_{\mathbf{e}}}} \prod_{j=1}^{l_{\mathbf{e}}} t(e_j|f_{a(j)})$$

$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_{\mathbf{e}}} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_{\mathbf{f}}} t(e_j|f_i)}$$

EM Formularization

EM Algorithm consists of two steps

2. Maximization-Step: Estimate new model from data

– Compute the count function *c*:

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_{a} p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

Where: $\delta(\cdot)$ is the Kronecker function.

– Estimate the new translation probability distribution:

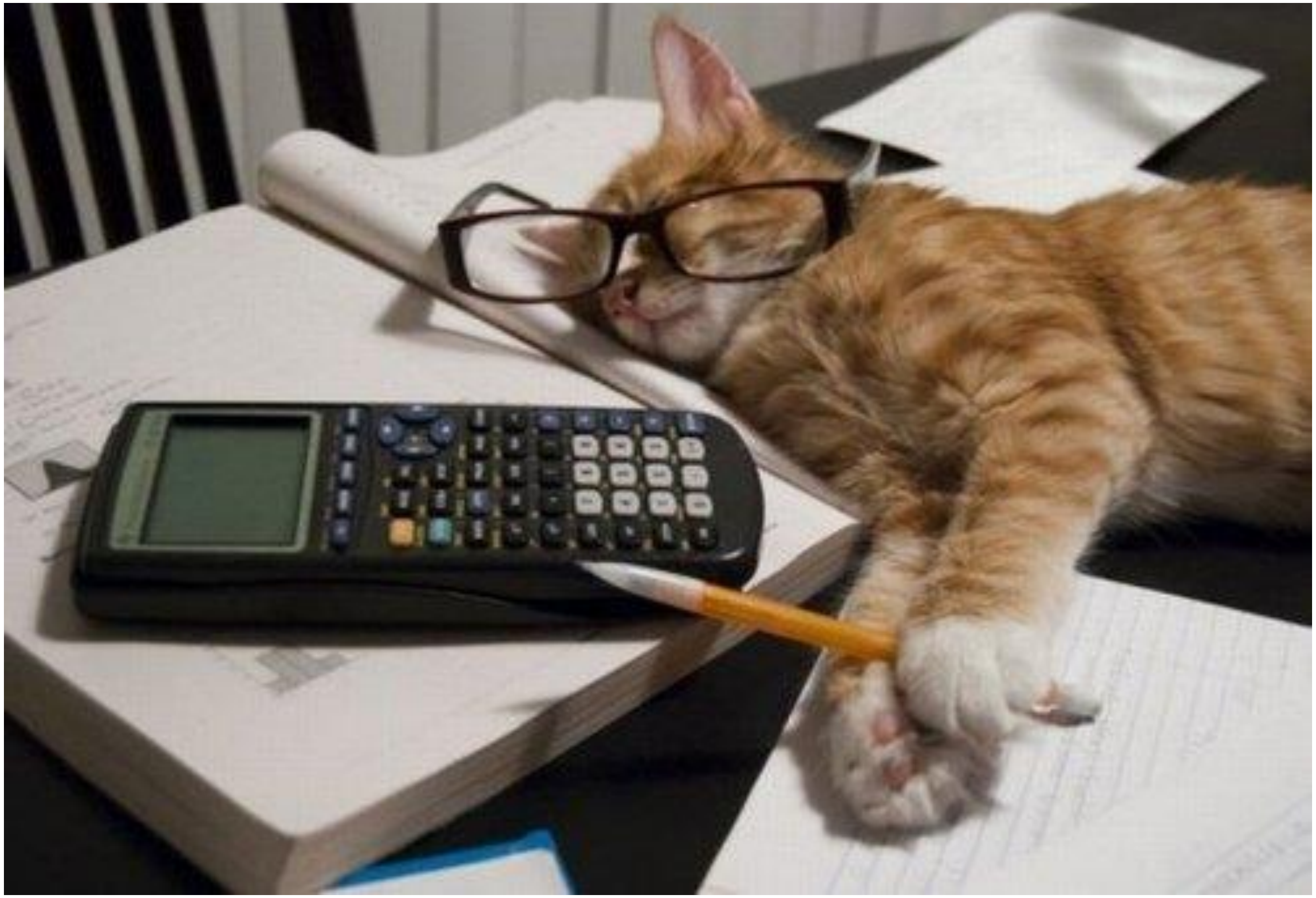
$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}{\sum_{e}\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}$$

Iterate these steps until convergence

$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_\mathbf{e}} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_\mathbf{f}} t(e_j|f_i)}$$

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}{\sum_e \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}$$



Example

Consider a parallel corpus containing just two pairs:

$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_\mathbf{e}} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_\mathbf{f}} t(e_j|f_i)}$$

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}{\sum_e \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}$$

blue house	house
maison bleu	maison

Q1: How many possible alignments in the first pair?

Q2: How many in the second pair?

Example

Consider a parallel corpus containing just two pairs:

blue house	house
maison bleu	maison

Assuming the translation direction: Fr → En

We will simplify the example by ruling out many-to-one or zero-to-one alignments.

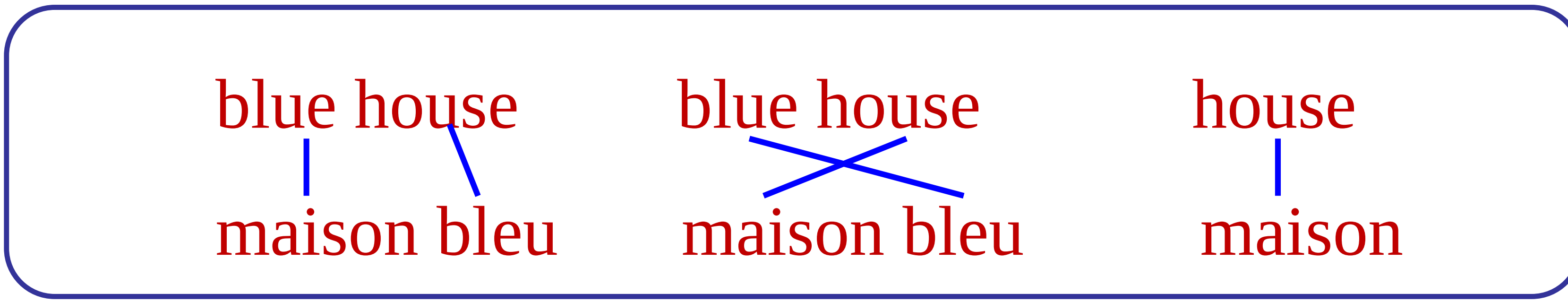
$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_\mathbf{e}} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_\mathbf{f}} t(e_j|f_i)}$$

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}{\sum_e \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}$$

Example

Consider a parallel corpus containing just two pairs:



Two possible alignments for the first pair?

One alignment for the second pair?

Step 1 (Initialisation)

Input words: {maison, bleu}

Output words: {blue, house}

Set parameter values uniformly.

- t(house|bleu) = ?
- t(house|maison) = ?
- t(blue|bleu) = ?
- t(blue|maison) = ?

Step 1 (Initialisation)

Input words: {maison, bleu}

Output words: {blue, house}

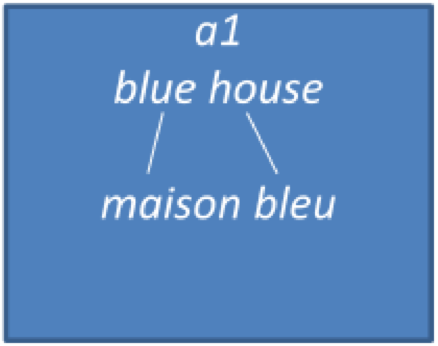
Set parameter values uniformly.

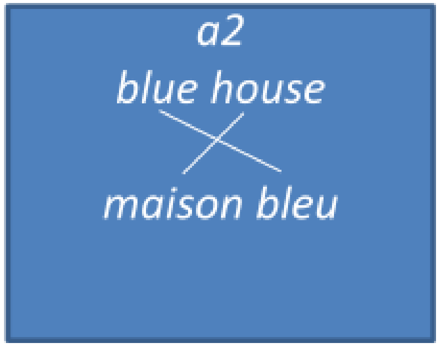
- t(house|bleu) = 1/2
- t(house|maison) = 1/2
- t(blue|bleu) = 1/2
- t(blue|maison) = 1/2

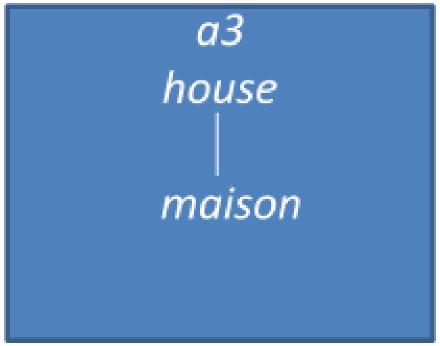
Step 2 (Expectation)

2-1: Compute the probability of **f** and **e** under alignment *a*: p(**e**, *a*|**f**)

$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_\mathbf{e}} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_\mathbf{f}} t(e_j|f_i)}$$







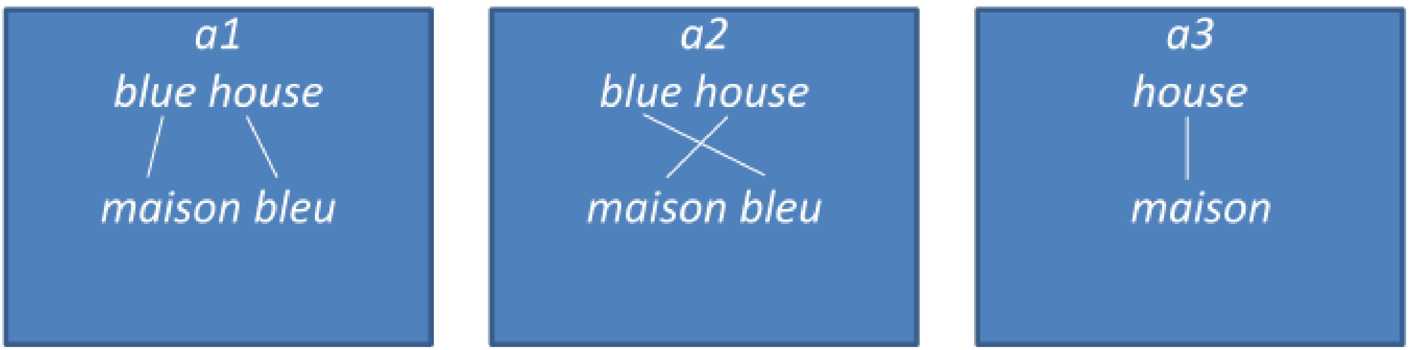
p(a1, blue house|maison bleu) = t(blue|maison) * t(house|bleu) = ½ * ½ = ¼

p(a2, blue house|maison bleu) = t(house|maison) * t(blue|bleu) = ½ * ½ = ¼

p(a3, house|maison) = t(house|maison) = ½

Step 2 (Expectation)

2-2: Normalise for all alignments - probability distribution of each of the alignment a: $p(a|\mathbf{e}, \mathbf{f})$



$$p(a|\mathbf{e},\mathbf{f}) = \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)}$$

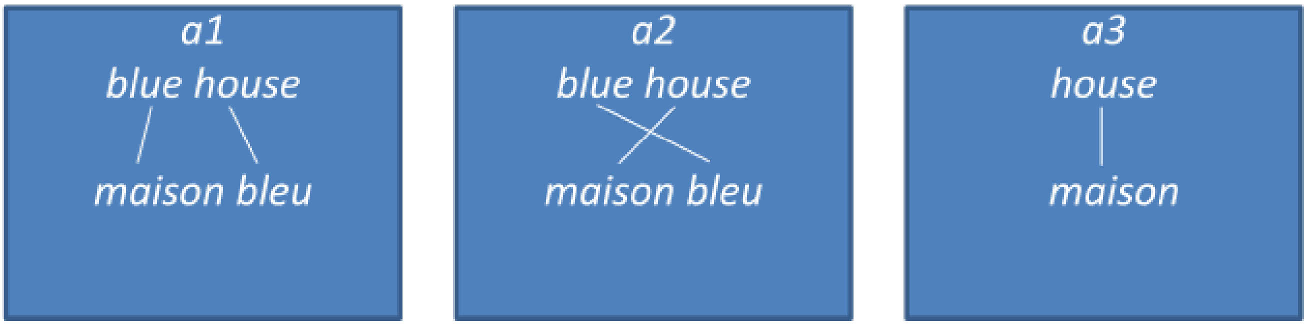
p(a1|blue house, maison bleu) = 1/4 ÷ 2/4 = 1/2

p(a2|blue house, maison bleu) = 1/4 ÷2/4 = 1/2

p(a3|house, maison) = 1/2 ÷ 1/2 = 1

Step 3 (Maximisation)

3-1: Collect fractional counts *c*:

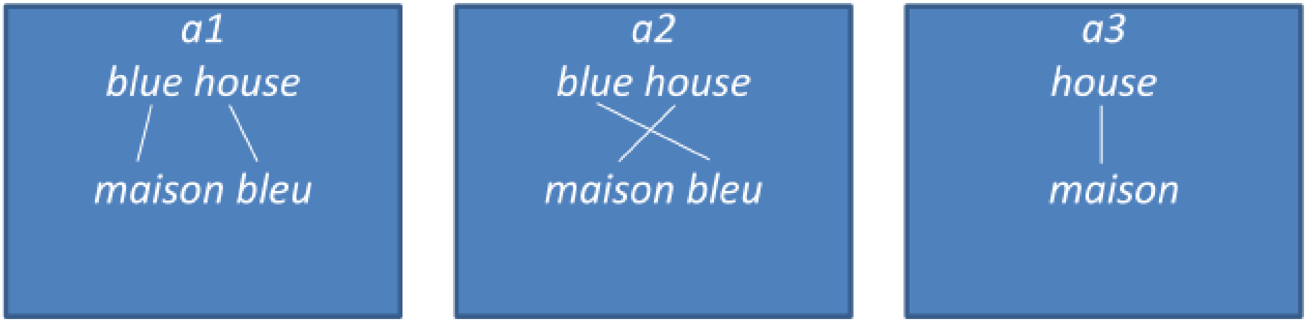


$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_{a} p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

- $c(\text{house}|\text{bleu}) = ½ * 1 = 1/2$
- $c(\text{house}|\text{maison}) = ½ *1 + 1*1 = 3/2$
- $c(\text{blue}|\text{bleu}) = ½ * 1 = 1/2$
- $c(\text{blue}|\text{maison}) = ½ *1 = 1/2$

Step 3 (Maximisation)

3-2: Normalise fractional counts to yield revised parameter values – estimate new model parameters

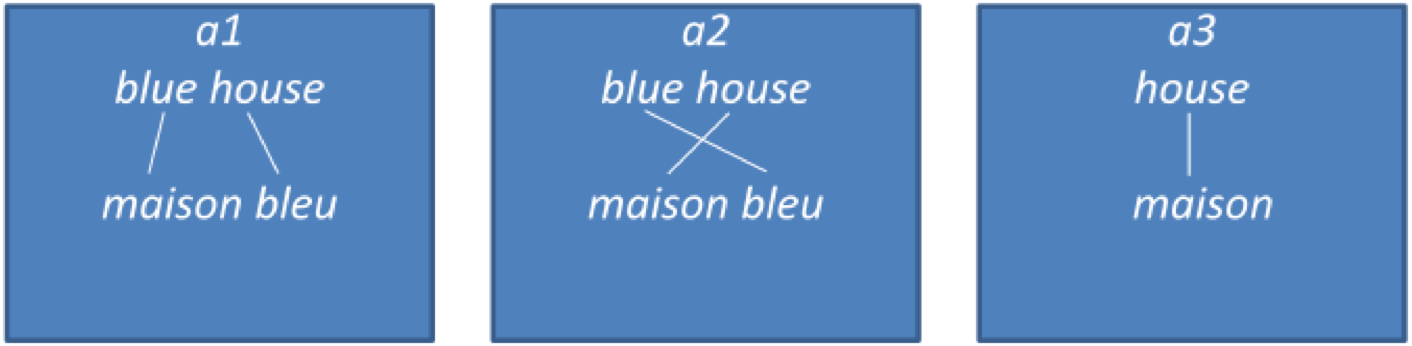


$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}{\sum_e \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f})}$$

- t(house|bleu) = 1/2 ÷ (1/2+1/2) = ½ ÷ 1 = 1/2
- t(house|maison) = 3/2 ÷ (3/2+1/2) = 3/4
- t(blue|bleu) = 1/2 ÷ (1/2+1/2)=1/2 ÷ 1 = 1/2
- t(blue|maison) = 1/2 ÷ (3/2+1/2) = 1/4

Iterate: Step 2 (Expectation)

2-1: Compute the probability of possible alignments.



$$p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})} = \frac{p(\mathbf{e},a|\mathbf{f})}{\sum_a p(\mathbf{e},a|\mathbf{f})}$$

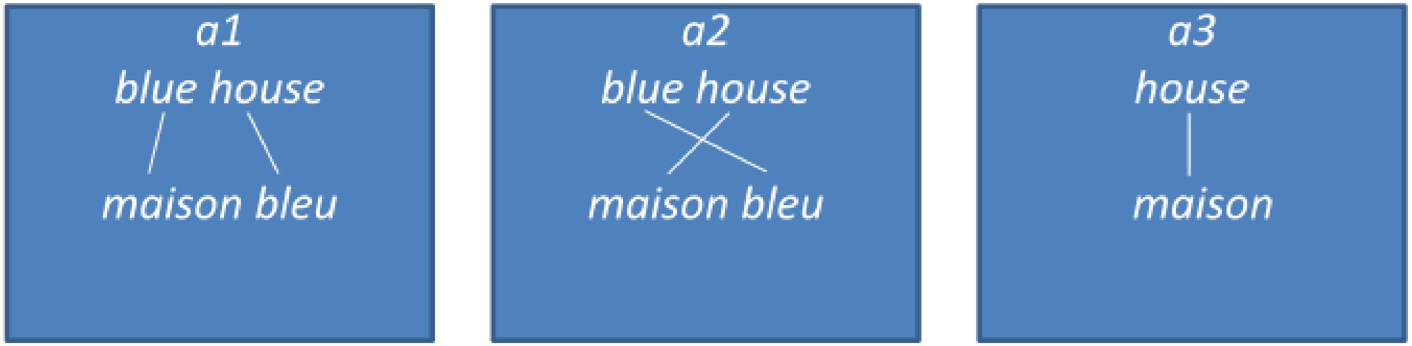
p(a1, blue house|maison bleu) = t(blue|maison) * t(house|bleu) = 1/4 * 1/2 = 1/8

p(a2, blue house|maison bleu) = t(house|maison) * t(blue|bleu) = 3/4 * 1/2 = 3/8

p(a3, house|maison) = t(house|maison) = 3/4

Iterate: Step 2 (Expectation)

2-2: Normalise for all alignments.



$$p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})} = \frac{p(\mathbf{e},a|\mathbf{f})}{\sum_a p(\mathbf{e},a|\mathbf{f})}$$

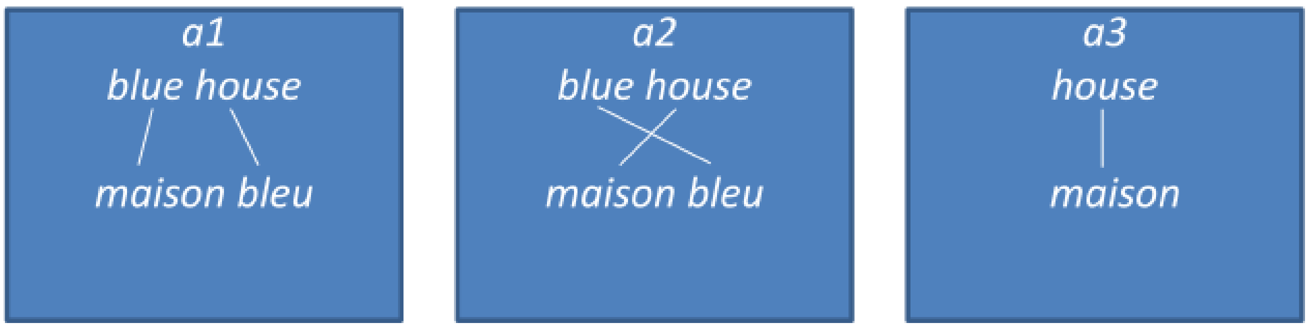
p(a1|blue house, maison bleu) = 1/8 ÷ (1/8+3/8) = 1/4

p(a2|blue house, maison bleu) = 3/8 ÷ 4/8 = 3/4

p(a3|house, maison) = 3/4 ÷ 3/4 = 1

Iterate: Step 3 (Maximisation)

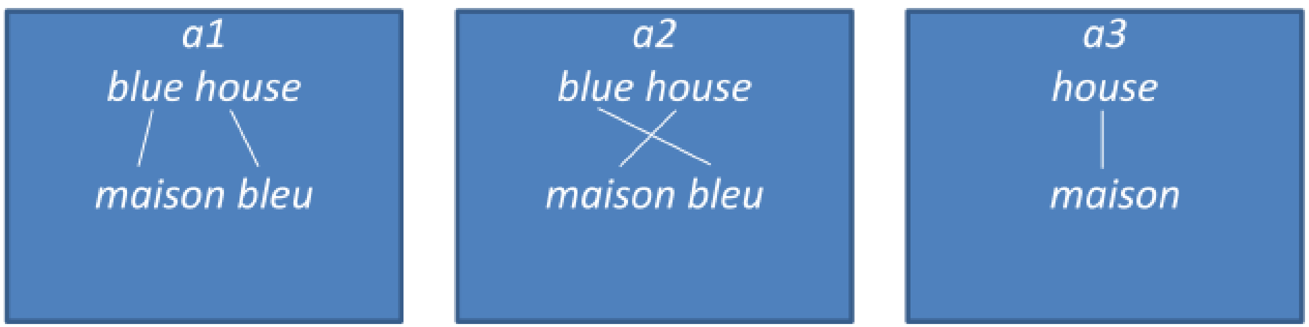
3-1: Collect fractional counts



- $c(\text{house}|\text{bleu}) = 1/4$
- $c(\text{house}|\text{maison}) = 3/4 + 1 = 7/4$
- $c(\text{blue}|\text{bleu}) = 3/4$
- $c(\text{blue}|\text{maison}) = 1/4$

Iterate: Step 3 (Maximisation)

3-2: Normalise fractional counts to yield revised parameter values



- $t(\text{house}|\text{bleu}) = 1/4 \div 1 = 1/8$
- $t(\text{house}|\text{maison}) = 7/4 \div (7/4+1/4) = 7/8$
- $t(\text{blue}|\text{bleu}) = 3/4 \div 1 = 3/4$
- $t(\text{blue}|\text{maison}) = 1/4 \div 1 = 1/4$

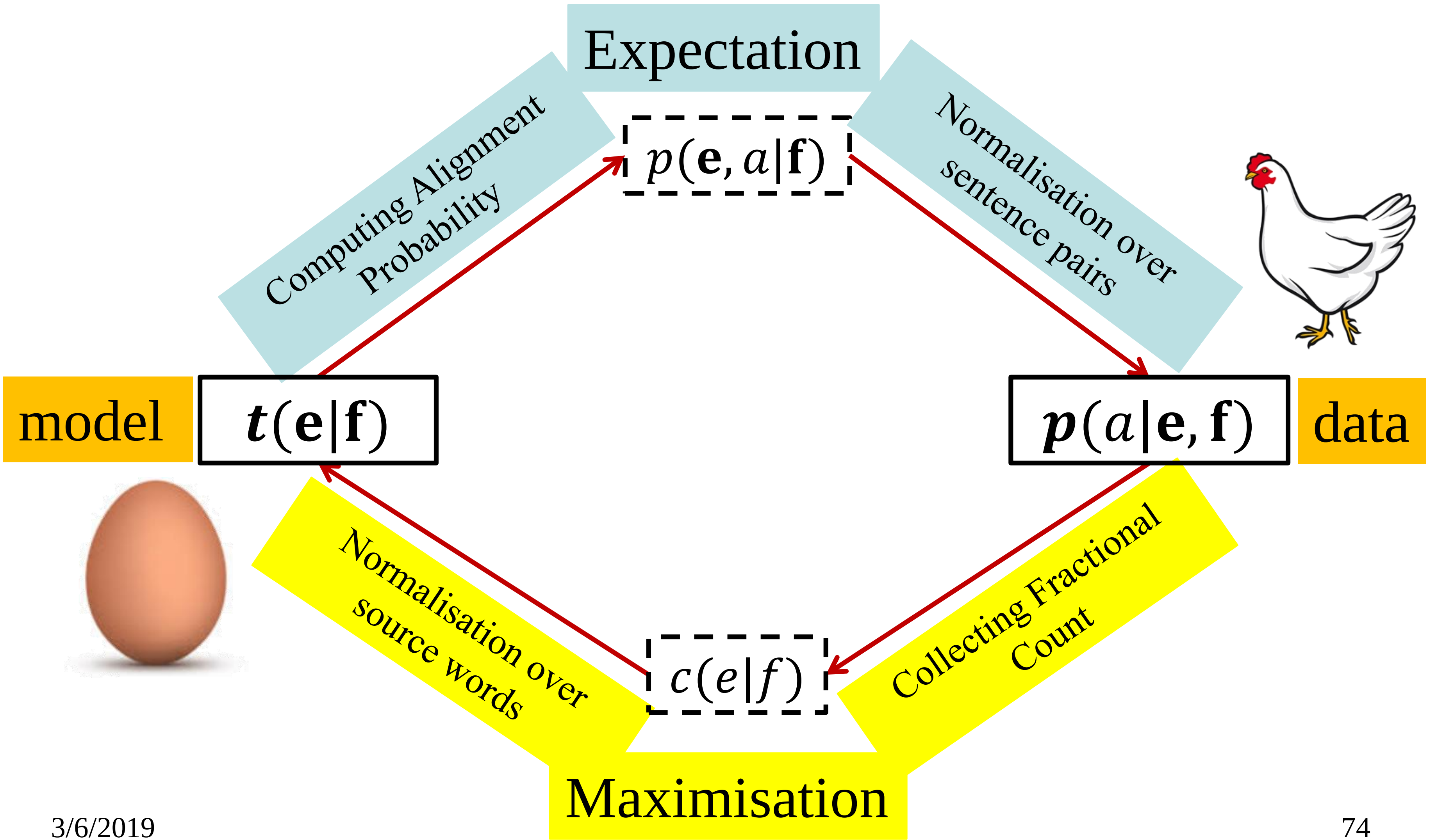
Convergence

Repeating steps 2 and 3 eventually yields:

- $t(\text{house}|\text{bleu}) = 0.0001$
- $t(\text{house}|\text{maison}) = 0.9999$
- $t(\text{blue}|\text{bleu}) = 0.9999$
- $t(\text{blue}|\text{maison}) = 0.0001$

It is proved that an EM algorithm is convergent.

EM Algorithm



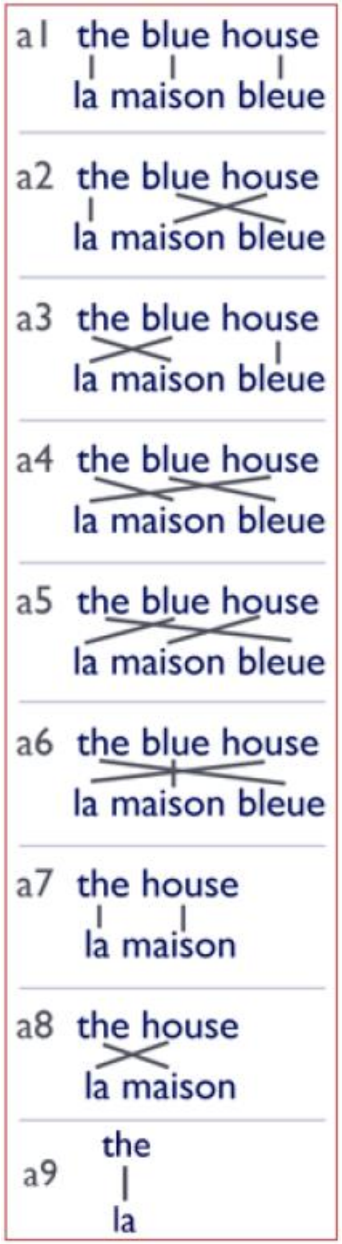
Content

Exercise

Use EM to estimate word translation probabilities (two iterations) given the following parallel corpus:

– the blue house ↔ la maison bleue
– the house ↔ la maison
– the ↔ la

Consider only the alignments on the right.
Translation direction: En→Fr



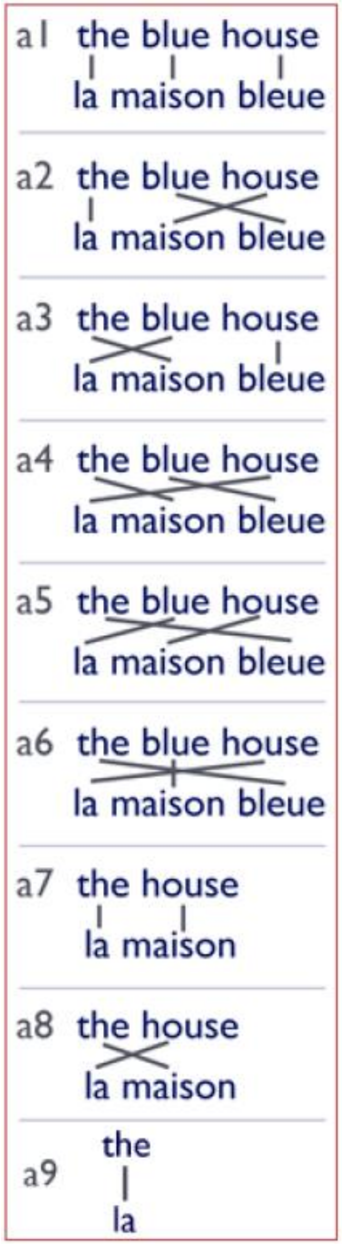
Initialisation

Input words: {the, blue, house}

Output words: {la, maison, bleue}

Set *t* parameters uniformly:

t(la|the) = 1/3	t(la|house)= 1/3
t(maison|the)= 1/3	t(maison|house)= 1/3
t(bleue|the)= 1/3	t(bleue|house) = 1/3
t(la|blue)= 1/3	
t(bleue|blue)= 1/3	
t(maison|blue)= 1/3	



Expectation (1)

t(la|the) = 1/3

t(maison|the)= 1/3

t(bleue|the)= 1/3

t(la|blue)= 1/3

t(bleue|blue)= 1/3

t(maison|blue)= 1/3

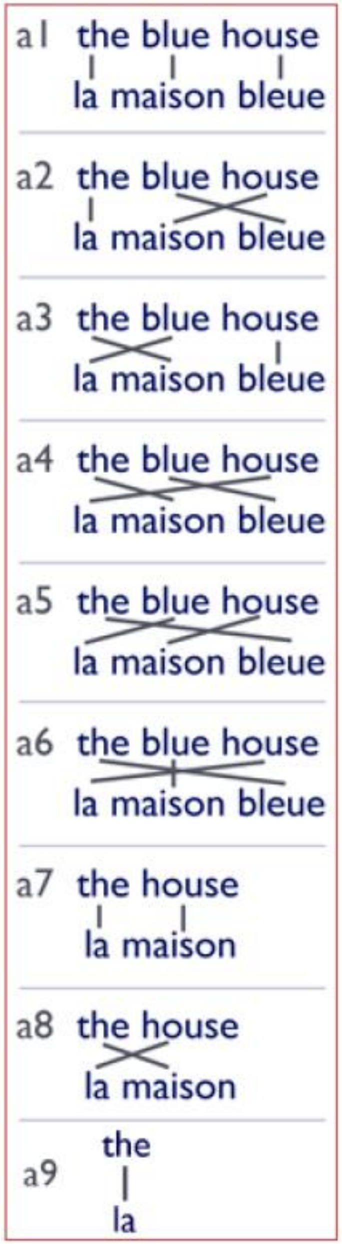
t(la|house)= 1/3

t(maison|house)= 1/3

t(bleue|house) = 1/3

Given our initial parameters, compute the probability of each of the possible alignments P(a,f|e) (illustrated in the box to the right):

- P(a1,f|e) = 1/3 (la|the) × 1/3 (maison|blue) × 1/3 (bleue|house) = 1/27
- P(a2,f|e) = 1/3 (la|the) × 1/3 (bleue|blue) × 1/3 (maison|house) = 1/27
- P(a3,f|e) = 1/3 × 1/3 × 1/3 = 1/27
- P(a4,f|e) = 1/3 × 1/3 × 1/3 = 1/27
- P(a5,f|e) = 1/3 × 1/3 × 1/3 = 1/27
- P(a6,f|e) = 1/3 × 1/3 × 1/3 = 1/27
- P(a7,f|e) = 1/3 (la|the) × 1/3 (masion|house) = 1/9
- P(a8,f|e) = 1/3 (maison|the) × 1/3 (la|house) = 1/9
- P(a9,f|e) = 1/3



Expectation (2)

- From previous step:

P(a1,f|e) = 1/27
P(a2,f|e) = 1/27
P(a3,f|e) = 1/27
P(a4,f|e) = 1/27
P(a5,f|e) = 1/27
P(a6,f|e) = 1/27

P(a7,f|e) = 1/9
P(a8,f|e) = 1/9
P(a9,f|e) = 1/3

- Normalize P(a,f|e) values to yield P(a|e,f) (normalize by sum of
- probabilities of possible alignments for the source string in question):

$P(a1|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$ ($\frac{6}{27}$ = sum over a1-a6 as they are possible alignments for the source string "the blue house")

$P(a2|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$

$P(a3|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$

$P(a4|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$

$P(a5|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$

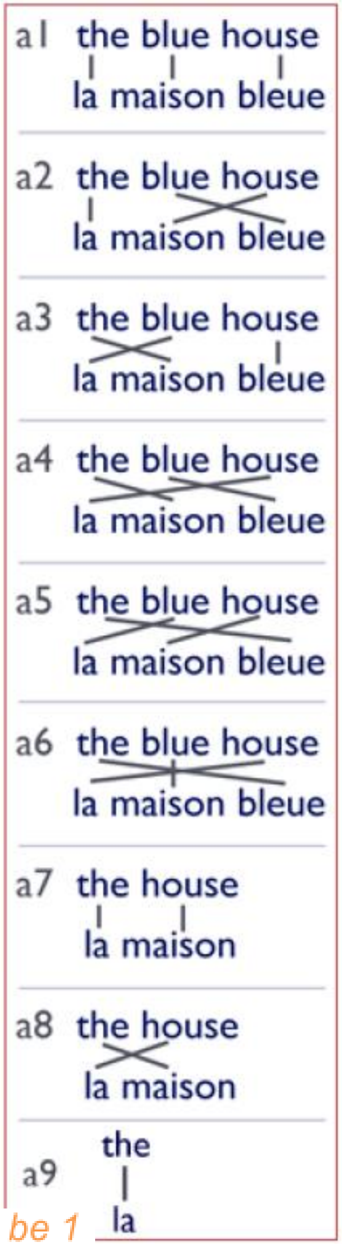
$P(a6|e,f) = \frac{1}{27} \div \frac{6}{27} = \frac{1}{6}$

$P(a7|e,f) = \frac{1}{9} \div \frac{2}{9} = \frac{1}{2}$

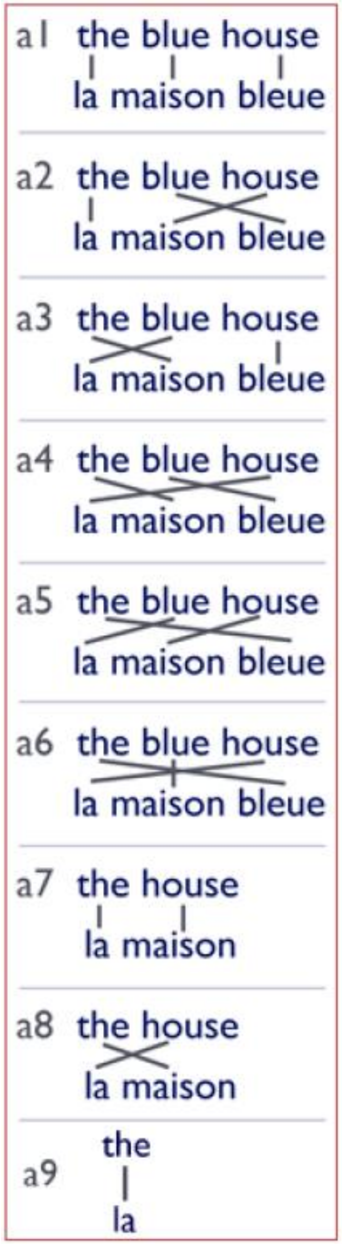
$P(a8|e,f) = \frac{1}{9} \div \frac{2}{9} = \frac{1}{2}$

$P(a9|e,f) = \frac{1}{3} \div \frac{1}{3} = 1$

only 1 alignment, therefore P(a9|e,f) will always be 1



Maximisation (1)



$P(a1|e,f) = \frac{1}{6}$

$P(a2|e,f) = \frac{1}{6}$

$P(a3|e,f) = \frac{1}{6}$

$P(a4|e,f) = \frac{1}{6}$

$P(a5|e,f) = \frac{1}{6}$

$P(a6|e,f) = \frac{1}{6}$

$P(a7|e,f) = \frac{1}{2}$

$P(a8|e,f) = \frac{1}{2}$

$P(a9|e,f) = 1$

- Collect fractional counts for each translation pair (i.e. for each translation pair, sum values of P(a|e,f) where the word pair occurs):

$$tc(la|the) = \frac{1}{6} \text{ (from a1)} + \frac{1}{6} \text{ (from a2)} + \frac{1}{2} \text{ (from a7)} + 1 \text{ (from a9)} = \frac{11}{6}$$

$$tc(maison|the) = \frac{1}{6} + \frac{1}{6} + \frac{1}{2} = \frac{5}{6}$$

$$tc(bleue|the) = \frac{1}{6} \text{ (from a5)} + \frac{1}{6} \text{ (from a6)} = \frac{2}{6}$$

$$tc(la|blue) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

$$tc(maison|blue) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

$$tc(bleue|blue) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

$$tc(la|house) = \frac{1}{6} + \frac{1}{6} + \frac{1}{2} = \frac{5}{6}$$

$$tc(maison|house) = \frac{1}{6} + \frac{1}{6} + \frac{1}{2} = \frac{5}{6}$$

$$tc(bleue|house) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$$

Maximisation (2)

$tc(la|the) = \frac{11}{6}$

$tc(maison|the) = \frac{5}{6}$

$tc(bleue|the) = \frac{2}{6}$

$tc(la|blue) = \frac{2}{6}$

$tc(maison|blue) = \frac{2}{6}$

$tc(bleue|blue) = \frac{2}{6}$

$tc(la|house) = \frac{5}{6}$

$tc(maison|house) = \frac{5}{6}$

$tc(bleue|house) = \frac{2}{6}$

- Normalize fractional counts to get revised parameters for **t**

$t(la|the) = \frac{11}{6} \div \frac{18}{6}$ (sum of counts for translation pairs where "the" occurs) $= \frac{11}{18}$

$t(maison|the) = \frac{5}{6} \div \frac{18}{6} = \frac{5}{18}$

$t(bleue|the) = \frac{2}{6} \div \frac{18}{6} = \frac{2}{18} = \frac{1}{9}$

$t(la|blue) = \frac{2}{6} \div \frac{6}{6} = \frac{2}{6} = \frac{1}{3}$

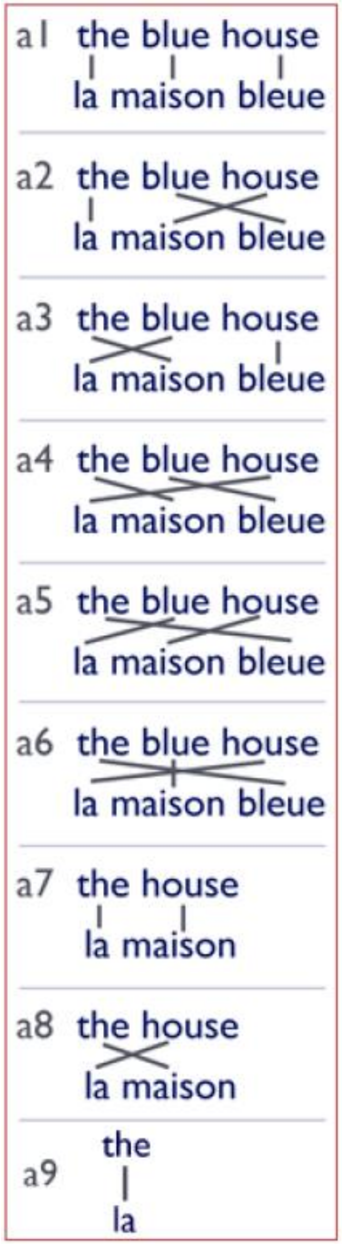
$t(maison|blue) = \frac{2}{6} \div \frac{6}{6} = \frac{1}{3}$

$t(bleue|blue) = \frac{2}{6} \div \frac{6}{6} = \frac{1}{3}$

$t(la|house) = \frac{5}{6} \div \frac{12}{6} = \frac{5}{12}$

$t(maison|house) = \frac{5}{6} \div \frac{12}{6} = \frac{5}{12}$

$t(bleue|house) = \frac{2}{6} \div \frac{12}{6} = \frac{2}{12} = \frac{1}{6}$



2nd Iteration: Expectation (1)

$t(la|the) = \frac{11}{18}$

$t(maison|the) = \frac{5}{18}$

$t(bleue|the) = \frac{1}{9}$

$t(la|blue) = \frac{1}{3}$

$t(maison|blue) = \frac{1}{3}$

$t(bleue|blue) = \frac{1}{3}$

$t(la|house) = \frac{5}{12}$

$t(maison|house) = \frac{5}{12}$

$t(bleue|house) = \frac{1}{6}$

- Given our new parameter values, re-compute the probability of each of the possible alignments P(a,f|e):

$$P(a1,f|e) = t(la|the) \times t(maison|blue) \times t(bleue|house) = \frac{11}{18} \times \frac{1}{3} \times \frac{1}{6} = \frac{11}{324}$$

$$P(a2,f|e) = \frac{11}{18} \times \frac{1}{3} \times \frac{5}{12} = \frac{55}{648}$$

$$P(a3,f|e) = \frac{5}{18} \times \frac{1}{3} \times \frac{1}{6} = \frac{5}{324}$$

$$P(a4,f|e) = \frac{5}{18} \times \frac{1}{3} \times \frac{5}{12} = \frac{25}{648}$$

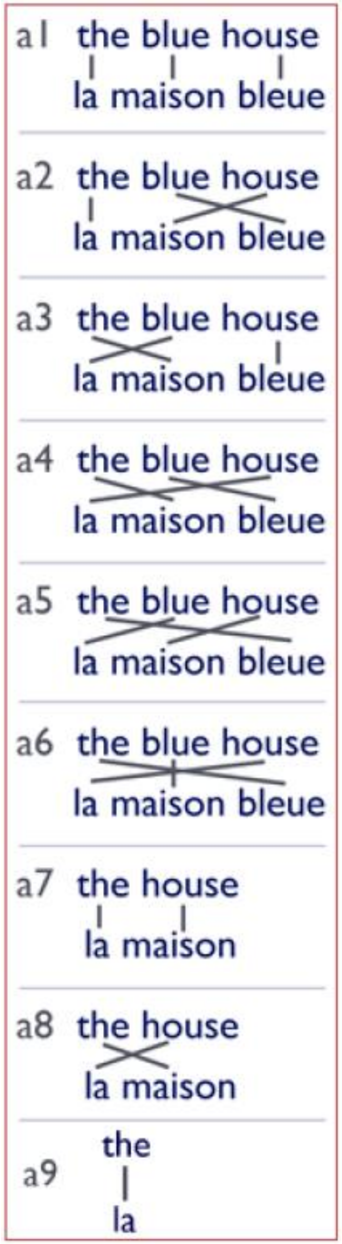
$$P(a5,f|e) = \frac{1}{9} \times \frac{1}{3} \times \frac{5}{12} = \frac{5}{324}$$

$$P(a6,f|e) = \frac{1}{9} \times \frac{1}{3} \times \frac{5}{12} = \frac{5}{324}$$

$$P(a7,f|e) = t(la|the) \times t(maison|house) = \frac{11}{18} \times \frac{5}{12} = \frac{55}{216}$$

$$P(a8,f|e) = \frac{5}{18} \times \frac{5}{12} = \frac{25}{216}$$

$$P(a9,f|e) = \frac{11}{18}$$



2^{nd} Iteration: Expectation (2)

$$P(a1,f|e) = \frac{11}{324} = \frac{22}{648} \qquad P(a6,f|e) = \frac{5}{324} = \frac{10}{648}$$

$$P(a2,f|e) = \frac{55}{648} \qquad P(a7,f|e) = \frac{55}{216}$$

$$P(a3,f|e) = \frac{5}{324} = \frac{10}{648} \qquad P(a8,f|e) = \frac{25}{216}$$

$$P(a4,f|e) = \frac{25}{648} \qquad P(a9,f|e) = \frac{11}{18}$$

$$P(a5,f|e) = \frac{5}{324} = \frac{10}{648}$$

- Normalize P(a,f|e) values to yield P(a|e,f) :

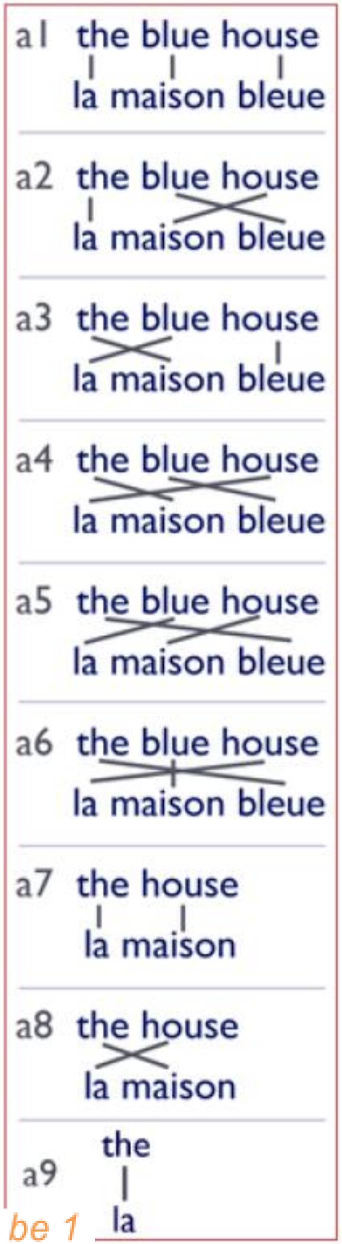
$$P(a1|e,f) = \frac{22}{648} \div \frac{132}{648} \text{ (sum a1-a6)} = \frac{22}{648} \times \frac{648}{132} = \frac{22}{132} \qquad P(a6|e,f) = \frac{10}{648} \div \frac{132}{648} = \frac{10}{132}$$

$$P(a2|e,f) = \frac{55}{648} \div \frac{132}{648} = \frac{55}{132} \qquad P(a7|e,f) = \frac{55}{216} \div \frac{80}{216} = \frac{55}{80}$$

$$P(a3|e,f) = \frac{10}{648} \div \frac{132}{648} = \frac{10}{132} \qquad P(a8|e,f) = \frac{25}{216} \div \frac{80}{216} = \frac{25}{80}$$

$$P(a4|e,f) = \frac{25}{648} \div \frac{132}{648} = \frac{25}{132} \qquad P(a9|e,f) = \frac{11}{18} \div \frac{11}{18} = 1$$

$$P(a5|e,f) = \frac{10}{648} \div \frac{132}{648} = \frac{10}{132}$$



only 1 alignment, therefore P(a9|e,f) will always be 1

2nd Iteration: Maximisation (1)

$P(a1|e,f) = \frac{22}{132} = \frac{1}{6} = \frac{8}{48} = \frac{88}{528}$

$P(a2|e,f) = \frac{55}{132} = \frac{5}{12} = \frac{20}{48} = \frac{220}{528}$

$P(a3|e,f) = \frac{10}{132} = \frac{40}{528}$

$P(a4|e,f) = \frac{25}{132} = \frac{100}{528}$

$P(a5|e,f) = \frac{10}{132} = \frac{40}{528}$

$P(a6|e,f) = \frac{10}{132}$

$P(a7|e,f) = \frac{165}{240} = \frac{11}{16} = \frac{33}{48}$

$P(a8|e,f) = \frac{75}{240} = \frac{5}{16} = \frac{165}{528}$

$P(a9|e,f) = 1 = \frac{48}{48}$

- Collect fractional counts for each translation pair:

$tc(la|the) = \frac{8}{48} + \frac{20}{48} + \frac{33}{48} + \frac{48}{48} = \frac{109}{48}$ (values from a1, a2, a7 and a9)

$tc(maison|the) = \frac{40}{528} + \frac{100}{528} + \frac{165}{528} = \frac{305}{528}$

$tc(bleue|the) = \frac{40}{528} + \frac{40}{528} = \frac{80}{528}$

$tc(la|blue) = \frac{40}{528} + \frac{40}{528} = \frac{80}{528}$

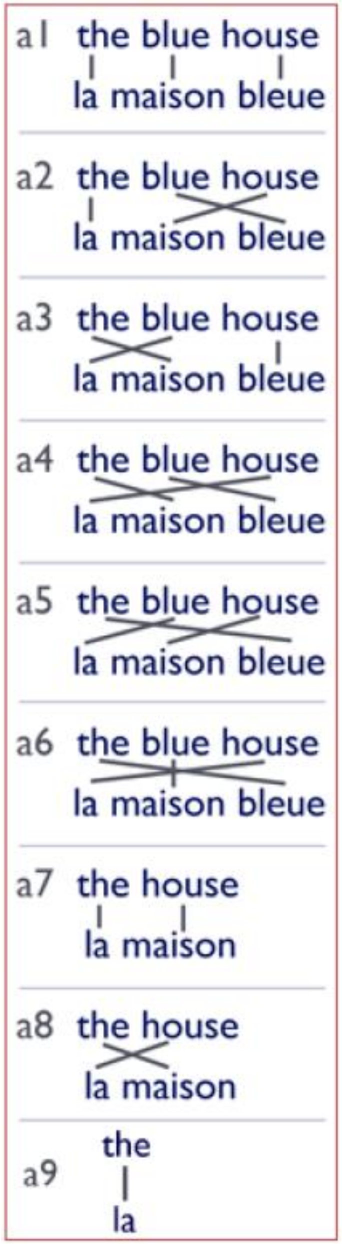
$tc(maison|blue) = \frac{88}{528} + \frac{40}{528} = \frac{128}{528}$

$tc(bleue|blue) = \frac{220}{528} + \frac{100}{528} = \frac{320}{528}$

$tc(la|house) = \frac{100}{528} + \frac{40}{528} + \frac{165}{528} = \frac{305}{528}$

$tc(maison|house) = \frac{220}{528} + \frac{40}{528} + \frac{33}{48} = \frac{623}{528}$

$tc(bleue|house) = \frac{88}{528} + \frac{40}{528} = \frac{128}{528}$



2^{nd} Iteration: Maximisation (2)

$tc(la|the) = \frac{109}{48}$

$tc(maison|the) = \frac{305}{528}$

$tc(bleue|the) = \frac{80}{528}$

$tc(la|blue) = \frac{80}{528}$

$tc(maison|blue) = \frac{128}{528}$

$tc(bleue|blue) = \frac{320}{528}$

$tc(la|house) = \frac{305}{528}$

$tc(maison|house) = \frac{623}{528}$

$tc(bleue|house) = \frac{128}{528}$

- Normalize fractional counts to get revised parameters for **t**

$$t(la|the) = \frac{109}{48} \div \left(\frac{109}{48} + \frac{305}{528} + \frac{80}{528} = \frac{1584}{528}\right) = \frac{1199}{1584} = \frac{109}{144}$$

$$t(maison|the) = \frac{305}{528} \div \frac{1584}{528} = \frac{305}{1584}$$

$$t(bleue|the) = \frac{80}{528} \div \frac{1584}{528} = \frac{80}{1584} = \frac{5}{99}$$

$$t(la|blue) = \frac{80}{528} \div \left(\frac{80}{528} + \frac{128}{528} + \frac{320}{528} = \frac{1}{1}\right) = \frac{80}{528} = \frac{5}{33}$$

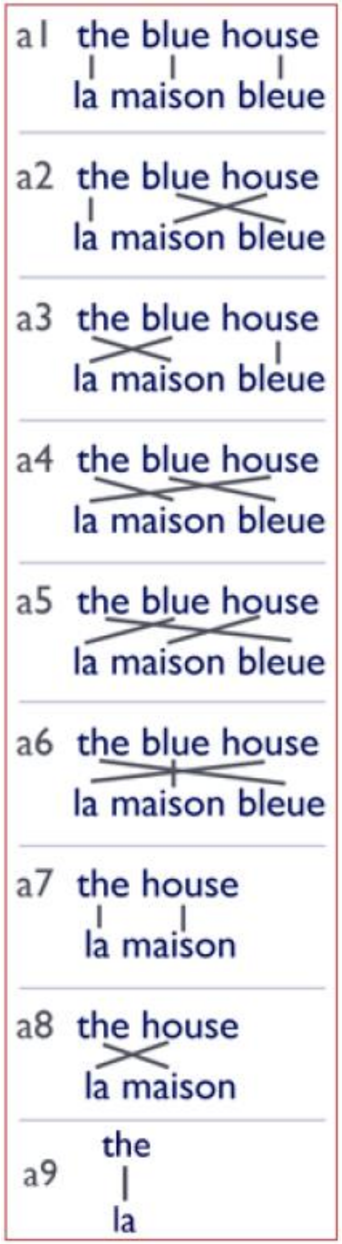
$$t(maison|blue) = \frac{128}{528} \div 1 = \frac{8}{33}$$

$$t(bleue|blue) = \frac{320}{528} \div 1 = \frac{20}{33}$$

$$t(la|house) = \frac{305}{528} \div \left(\frac{305}{528} + \frac{623}{528} + \frac{128}{528} = \frac{1056}{528}\right) = \frac{305}{1056}$$

$$t(maison|house) = \frac{623}{528} \div \frac{1056}{528} = \frac{623}{1056}$$

$$t(bleue|house) = \frac{128}{528} \div \frac{1056}{528} = \frac{128}{1056}$$



EM: Convergence

- After the second iteration, our **t** values are:

$t(la|the) = \frac{109}{144} = \mathbf{0.7569}$

$t(maison|the) = \frac{305}{1584} = 0.1926$

$t(bleue|the) = \frac{5}{99} = 0.0505$

$t(la|blue) = \frac{5}{33} = 0.1515$

$t(maison|blue) = \frac{8}{33} = 0.2424$

$t(bleue|blue) = \frac{20}{33} = \mathbf{0.6061}$

$t(la|house) = \frac{305}{1056} = 0.2888$

$t(maison|house) = \frac{623}{1056} = \mathbf{0.5810}$

$t(bleue|house) = \frac{128}{1056} = 0.1212$

- We continue EM until our **t** values converge
- It is clear to see already, after 2 iterations, how some translation candidates are (correctly) becoming more likely then others

Discussion

Statistical Machine Translation, Philipp Koehn (2010), CUP, Cambridge, UK.
http://www.statmt.org/book/errata.html