

Parallel Corpora & Alignment

AARON SMITH



UPPSALA
UNIVERSITET

Machine Translation VT 2016
Uppsala, 20th April 2016

Goals for today

- What are parallel corpora and why do we need them?
- How do we create a parallel corpus?
 - Finding multilingual data
 - Sentence alignment
 - Word alignment

What is a parallel corpus?

- A (large) collection of texts in **at least two** languages
- Aligned sentence-by-sentence
- Word-alignments often also present

A three-sentence Swedish-English corpus

Är marknaden en bra, dålig eller neutral institution?

Is the market a good, bad or neutral institution?

Efter att ha genomgått kursen förväntas studenten:

It is expected that the student after taking the course will be able to:

Kursen ger också en orientering i det svenska transkriptionssystemet.

The course also provides an overview of the Swedish transcription system.

What is a parallel corpus?

- A (large) collection of texts in **at least two** languages
- Aligned sentence-by-sentence
- Word-alignments often also present

A three-sentence Swedish-English corpus

Är marknaden en bra, dålig eller neutral institution?

Is the market a good, bad or neutral institution?

Efter att ha genomgått kursen förväntas studenten:

It is expected that the student after taking the course will be able to:

Kursen ger också en orientering i det svenska transkriptionssystemet.

The course also provides an overview of the Swedish transcription system.

What is a parallel corpus?

<http://opus.lingfil.uu.se>



... the open parallel corpus

OPUS is a growing collection of translated texts from the web. In the OPUS project we try to convert and align free online data, to add linguistic annotation, and to provide the community with a publicly available parallel corpus. OPUS is based on open source products and the corpus is also delivered as an open content package. We used several tools to compile the current collection. All pre-processing is done automatically. No manual corrections have been carried out.

The OPUS collection is growing! Check this page from time to time to see new data arriving ...
Contributions are very welcome! Please contact <jorg.tiedemann@lingfil.uu.se>

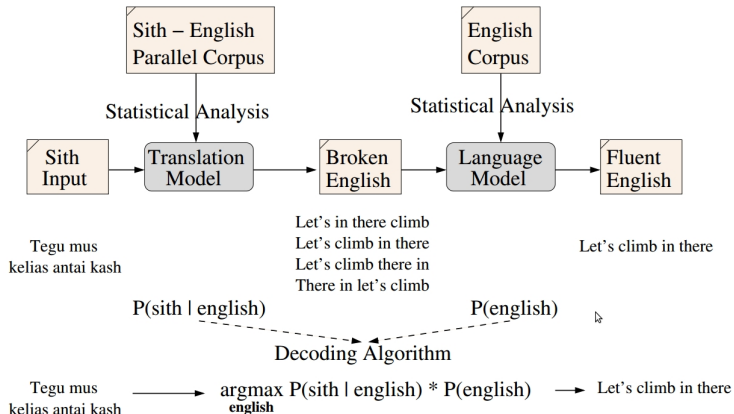
Search & download resources:

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, true = truecaser model, TM = phrase-based translation model)

corpus	doc's	sent's	src tokens	trg tokens	XCES/XML	raw	TMX	Moses	mono	raw	true	TM	dic	freq	Browser Files
OpenSubtitles2016	15036	13.3M	110.8M	88.4M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv][raw/en][raw/sv]
EBookshop	4006	2.0M	87.0M	69.3M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[sample] [xml/en][xml/sv]
OpenSubtitles2012	10851	9.9M	82.8M	66.0M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[sample]
DGT	26884	3.2M	72.7M	54.9M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	en sv		[sample] [xml/en][xml/sv]
Europarl	9385	1.9M	57.3M	51.5M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv]
OpenSubtitles2013	7346	7.2M	60.5M	47.5M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv			dic	en sv	[sample]
Europarl3	608	1.2M	36.6M	33.2M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[query] [sample] [xml/en][xml/sv]
JRC-Acquis	11745	0.8M	33.2M	26.8M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv]
EMEA	1915	1.1M	12.0M	13.7M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv]
GNOME	2152	0.7M	4.9M	4.9M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv]
Tanzil	15	0.1M	2.8M	3.1M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	en sv		[query] [sample] [xml/en][xml/sv][xml/en-sv]
OpenSubtitles	348	0.4M	3.0M	2.5M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv][xml/en-sv]
Tatoeba	1	9.8k	3.6M	0.1M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[sample] [xml/en][xml/sv]
KDE4	2161	0.2M	2.4M	0	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv]
WikiSource	66	35.0k	0.9M	0.9M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv][xml/en-sv]
Ubuntu	445	0.1M	0.8M	0.5M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv]
OpenOffice	1739	38.9k	0.4M	0.4M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv]
PHP	3283	35.7k	0.5M	75.9k	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv]
GlobalVoices	316	8.2k	0.2M	0.2M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv					[sample] [xml/en][xml/sv]
EUconst	47	10.1k	0.2M	0.1M	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv][xml/en-sv]
Books	1	3.1k	79.3k	76.6k	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	en sv		[sample] [xml/en][xml/sv][xml/en-sv]
RF	2	0.2k	4.4k	2.9k	[xc es en sv]	[en sv]	[tmx]	[moses]	en sv	en sv	en sv	en-sv	dic	en sv	[query] [sample] [xml/en][xml/sv][xml/en-sv]
total	98352	42.2M	572.8M	464.0M	42.2M		30.3M	38.2M							

What are parallel corpora used for?

From Fabienne's lecture:



What else?

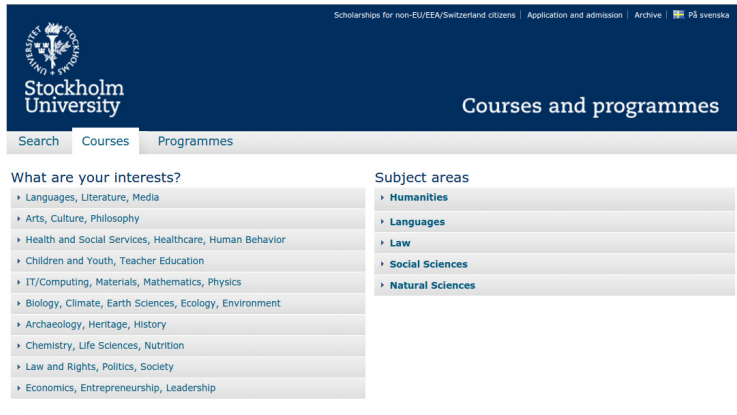
Any ideas?

How do we create a parallel corpus?


- Collect translated documents
 - Web scraping
- Pre-processing
 - Conversion to another format
 - Sentence boundary detection (segmentation)
 - Tokenization
- Alignment
 - Document alignment
 - Paragraph alignment
 - Sentence alignment
 - Word alignment

Example: Course syllabuses

<https://sisu.it.su.se/search/courses/en>



Scholarships for non-EU/EEA/Switzerland citizens | Application and admission | Archive | På svenska

 **Stockholm University**

Courses and programmes

Search | **Courses** | Programmes

What are your interests?

- › Languages, Literature, Media
- › Arts, Culture, Philosophy
- › Health and Social Services, Healthcare, Human Behavior
- › Children and Youth, Teacher Education
- › IT/Computing, Materials, Mathematics, Physics
- › Biology, Climate, Earth Sciences, Ecology, Environment
- › Archaeology, Heritage, History
- › Chemistry, Life Sciences, Nutrition
- › Law and Rights, Politics, Society
- › Economics, Entrepreneurship, Leadership

Subject areas

- › **Humanities**
- › **Languages**
- › **Law**
- › **Social Sciences**
- › **Natural Sciences**

Stockholm University

© Stockholm University, SE-106 91 Stockholm, Sweden | Phone: +46 8 16 20 00

Practical exercise

Try to align these sentences:

English	"Swedish"
Tropical Marine Biology	Tebcvfx znevaovbybtv
7.5 Higher Education	7.5 Höftxbyrcbåat
Credits	7.5 ECTS perqvgf
7.5 ECTS credits	Pebixbq
(Three credits corresponds to approximately two weeks full-time studies).	5003
Examination code	Khefra tre ra trabztåat ni qrg gebcvfxn znevan ynaqfxcnrg bpu frzfcryrg zryyna xhfgmbaraf byvxn rxblffgrz:
5003	znatebir, xbenyyeri, fwöteååfatne, nievaavatfbzeåqra bpu öccan uning.
The course covers the tropical marine landscape and the interaction between different ecosystems such as the mangroves, coral reefs, seagrass beds, run-off area and the open ocean.	Sghqrenaqr fbz haqrexåagf v beqvanevr cebi une eågg ngg trabztå zvaifg slen lggreyvtner cebi få yåatr xhefra trf.
Students who fail to achieve a pass grade in an ordinary examination have the right to take at least further four examinations, as long as the course is given.	Mrq cebi wåzfgyäyf bpxfå naqen boyvtngbevfxn xhefqrnye.
The term "examination" here is used to denote also other compulsory elements of the course.	Öiretåatforfgåzzryfre
Interim	Sghqrenaqr xna ortåen ngg rknzvangvba trabzsöef rayvtg qraan xhefcyna åira rsgre qrg ngg qra hccuöeg ngg täyyn, qbpx uöftg ger tåatre haqre ra giååefcrevbq rsgre qrg ngg haqreivafvat cå xhefra hccuöeg.
Students may request that the examination is carried out in accordance with this syllabus even after it has ceased to apply.	Fenzfgåyyna uåebz fxn töenf gvyy vatfgvghgvbaffgleryfra.
This right is limited, however, to a maximum of three occasions during a two-year-period after the end of giving the course.	Breåafavatne
A request for such examination must be sent to the departmental board.	Khefra xna rw vatå v rknzra gvyyfnzznaf zrq xhefra Tebcvfx inggraiåeg 5 c (BI3820) ryyre zbgfinenaqr.
Limitations	Öievtg
The course may not be included in a degree together with the course Management of Aquatic Resources in the Tropics 5 p (BI3820) or the equivalent.	Khefra vatåe v xnaqvngcebtentzrrg v ovbytvr zra xna bpxfå yåfnf fbz sevfgåraqr xhef.
Misc	
The course is a component of the Bachelor's Programmes in Biology and Marine Biology, and it can also be taken as an individual course.	

Practical exercise

Solution:

English	Swedish
Tropical Marine Biology	Tropisk marinbiologi
7.5 Higher Education	7.5 Högskolepoäng
Credits	7.5 ECTS credits
7.5 ECTS credits	Provkod
(Three credits corresponds to approximately two weeks full-time studies).	5003
Examination code	Kursen ger en genomgång av det tropiska marina landskapet och samspelet mellan kustzonens olika ekosystem: mangrove, korallrev, sjögräsängar, avrinningsområden och öppna havet.
5003	Studering som underkänt i ordinarie prov har rätt att genomgå minst fyra ytterligare prov så länge kursen ges.
The course covers the tropical marine landscape and the interaction between different ecosystems such as the mangroves, coral reefs, seagrass beds, run-off area and the open ocean.	Med prov jämföras också andra obligatoriska kursdelar.
Students who fail to achieve a pass grade in an ordinary examination have the right to take at least further four examinations, as long as the course is given.	Övergångsbestämmelser
The term "examination" here is used to denote also other compulsory elements of the course.	Studering kan begära att examination genomförs enligt denna kursplan även efter det att den upphört att gälla, dock högst tre gånger under en tvåårsperiod efter det att undervisning på kursen upphört.
Interim	Framställan härom ska göras till institutionsstyrelsen.
Students may request that the examination is carried out in accordance with this syllabus even after it has ceased to apply.	Begränsningar
This right is limited, however, to a maximum of three occasions during a two-year-period after the end of giving the course.	Kursen kan ej ingå i examen tillsammans med kursen Tropisk vattenvård 5 p (BI3820) eller motsvarande.
A request for such examination must be sent to the departmental board.	Övrigt
Limitations	Kursen ingår i kandidatprogrammet i biologi men kan också läsas som fristående kurs.
The course may not be included in a degree together with the course Management of Aquatic Resources in the Tropics 5 p (BI3820) or the equivalent.	
Misc	
The course is a component of the Bachelor's Programmes in Biology and Marine Biology, and it can also be taken as an individual course.	

Practical exercise

What type of alignments did we see?

- 1:1
- 2:1
- 1:0

Manual alignment

- Extremely Slow
 - We did 18 sentences in ~ 5 minutes
 - 1000 sentences in ~ 4.5 hours
 - 1,000,000 sentences in ~ 4500 hours = 188 days
- Very Accurate ($> 99\%$)

Can we do this faster without dropping accuracy significantly?

Practical exercise

What type of alignments did we see?

- 1:1
- 2:1
- 1:0

Manual alignment

- Extremely Slow
 - We did 18 sentences in ~ 5 minutes
 - 1000 sentences in ~ 4.5 hours
 - 1,000,000 sentences in ~ 4500 hours = 188 days
- Very Accurate ($> 99\%$)

Can we do this faster without dropping accuracy significantly?

Automatic sentence alignment

Gale & Church 1990: "longer sentences in one language tend to be translated into longer sentences in another language."

But how do we measure sentence length? Number of characters or number of words?

Consider the following:

English: "You know how to describe the time and space complexity of an algorithm." *13 words, 72 characters*

Finnish: "Osaat selittää, miten algoritmin aika- ja tilavaativuutta kuvataan." *8 words, 70 characters*

Automatic sentence alignment

Gale & Church 1990: "longer sentences in one language tend to be translated into longer sentences in another language."

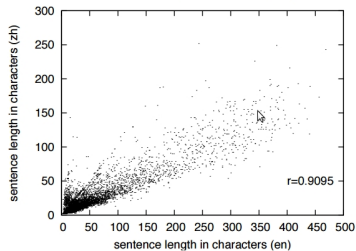
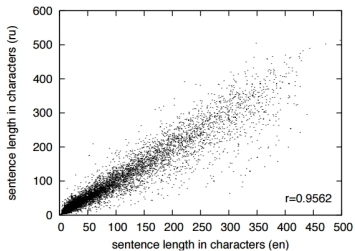
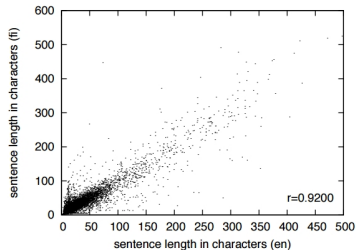
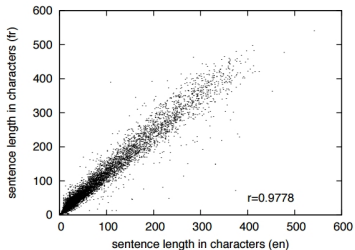
But how do we measure sentence length? Number of characters or number of words?

Consider the following:

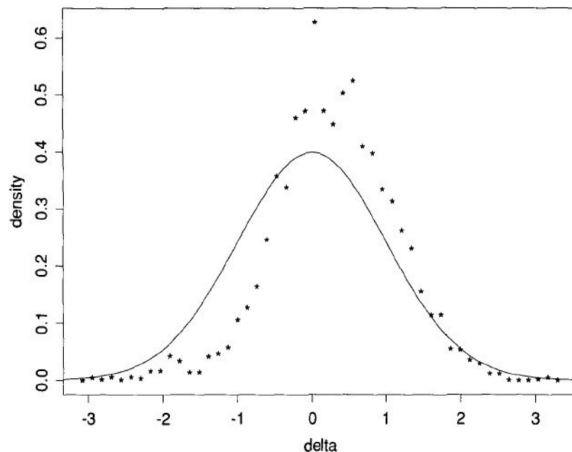
English: "You know how to describe the time and space complexity of an algorithm." *13 words, 72 characters*

Finnish: "Osaat selittää, miten algoritmin aika- ja tilavaativuutta kuvataan." *8 words, 70 characters*

Length correlation



Normal distribution



$$\delta(l_1, l_2) = (l_1 - l_2 c) / \sqrt{\frac{1}{2}(l_1 + l_2) s^2}$$

Sentence alignment model

Bayes' theorem:

$$p(\text{match}|\delta) = K \times p(\delta|\text{match}) \times p(\text{match})$$

Trick:

$$p(\delta|\text{match}) = 2(1 - p(|\delta|))$$

What about $p(\text{match})$?

Depends

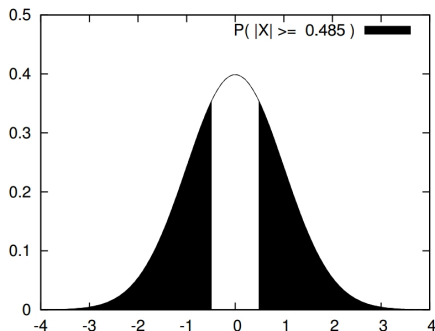
on alignment **type**:

1:1 = 0.89

1:0 or 0:1 = 0.0099

2:1 or 1:2 = 0.089

2:2 = 0.011



Sentence alignment model

- Define the cost of an alignment a_i as
$$d(a_i) = -\log p(\text{match}|\delta)$$
- Task: Find alignment $A' = (a_1, a_2, \dots)$ with **minimal total cost**
- $A' = \operatorname{argmin}_A \sum_i -\log(p(\delta|\text{match}) \times p(\text{match}))$
- We know how to calculate all these things for all possible alignments
- But there are lots of possible alignments so we need an efficient algorithm
 - Dynamic programming

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X						
1							
2							
3							
4							

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X	X					
1	X						
2							
3							
4							

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X	X					
1	X	X					
2							
3							
4							

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X	X	X				
1	X	X	X				
2	X	X	X				
3							
4							

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X	X	X	X	X	X	X
1	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X
4	X	X	X	X	X	X	

Dynamic programming

Source → Target ↓	0	1	2	3	4	5	6
0	X	X	X	X	X	X	X
1	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X
3	X	X	X	X	X	X	X
4	X	X	X	X	X	X	X

Other methods for automatic sentence alignment

- Distance-based measures work very well ($> 95\%$) for 'easy-to-align' corpora
- For more difficult corpora we need more sophisticated methods
 - Cognates
 - Dictionary look-up
 - Two-pass algorithm - align, translate, align again
- Must also consider speed vs. accuracy trade-off

How do we create a parallel corpus?

- Collect translated documents ✓
 - Web scraping
- Pre-processing ✓
 - Conversion to another format
 - Sentence boundary detection (segmentation)
 - Tokenization
- Alignment
 - Document alignment ✓
 - Paragraph alignment ✓
 - Sentence alignment ✓
 - **Word alignment**

Reminder on IBM model 1

From Fabienne's lecture:

das		Haus		ist		klein	
<i>e</i>	<i>t(e f)</i>	<i>e</i>	<i>t(e f)</i>	<i>e</i>	<i>t(e f)</i>	<i>e</i>	<i>t(e f)</i>
the	0.7	house	0.8	is	0.8	small	0.4
that	0.15	building	0.16	's	0.16	little	0.4
which	0.075	home	0.02	exists	0.02	short	0.1
who	0.05	household	0.015	has	0.015	minor	0.06
this	0.025	shell	0.005	are	0.005	petty	0.04

$$\begin{aligned} p(e, a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

Chicken and egg problem

- How do we calculate the lexical translation probabilities?
 - Maximum-likelihood estimation (i.e. counting instances from a corpus)
- But we have assumed we know the alignment
- On the other hand, we can use the translation models to figure out the most likely alignment

The problem

Given the model, we could fill the gaps in our data: given the data, we could estimate the model. To begin with, we have neither!

- Solution: Expectation Maximization (EM)

Chicken and egg problem

- How do we calculate the lexical translation probabilities?
 - Maximum-likelihood estimation (i.e. counting instances from a corpus)
- But we have assumed we know the alignment
- On the other hand, we can use the translation models to figure out the most likely alignment

The problem

Given the model, we could fill the gaps in our data: given the data, we could estimate the model. To begin with, we have neither!

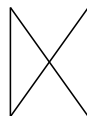
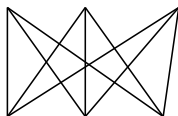
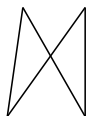
- Solution: Expectation Maximization (EM)

EM in a nutshell

- 1 Initialize the model, typically with uniform distributions
- 2 Apply the model to the data (expectation step)
- 3 Estimate the model from the data (maximization step)
- 4 Iterate steps 2-3 until convergence

EM algorithm

... la maison ... la maison blue ... la fleur ...

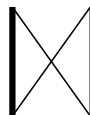
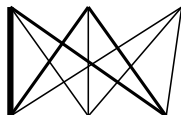


... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
- Model learns that *la*, for example, is often aligned with *the*

EM algorithm

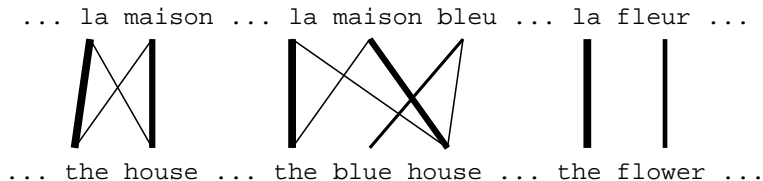
... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

- After one iteration
- Certain alignments, for example between *la* and *the*, are now more likely

EM algorithm



- After another iteration
- It becomes apparent the other alignments, such as *fleur* and *flower*, are more likely

EM algorithm

... la maison ... la maison bleu ... la fleur ...
/ | | X | |
... the house ... the blue house ... the flower ...

- Convergence
- Inherent hidden structure revealed by EM

EM algorithm

... la maison ... la maison bleu ... la fleur ...
/ | | X | |
... the house ... the blue house ... the flower ...



$$\begin{aligned}p(\text{la}|\text{the}) &= 0.453 \\p(\text{le}|\text{the}) &= 0.334 \\p(\text{maison}|\text{house}) &= 0.876 \\p(\text{bleu}|\text{blue}) &= 0.563 \\&\dots\end{aligned}$$

- Parameter estimation from aligned corpus

- Note that in the maximization step, we still don't know the correct alignment, but we have an estimate of the **probability of every possible alignment**
- To collect counts, we could just consider the alignment with the highest probability
- Even better: take a weighted average of the counts over all possible alignments

EM and the IBM models

IBM Model 1	lexical translation
IBM Model 2	adds absolute reordering model
IBM Model 3	adds fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

- EM algorithm can be applied to all IBM models
- With lower IBM models we can apply certain mathematical tricks to simplify calculations (see course textbook)
- Only with IBM Model 1 are we guaranteed to reach a global maximum

EM and the IBM models

IBM Model 1	lexical translation
IBM Model 2	adds absolute reordering model
IBM Model 3	adds fertility model
IBM Model 4	relative reordering model
IBM Model 5	fixes deficiency

- From IBM Model 3 computation becomes more expensive and sampling over high probability alignments is employed
- Typical training scheme uses all IBM models sequentially, using result from one to initialise the next
- Popular implementation: GIZA++

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT

Summary

- A parallel corpus is a collections of texts in at least two languages, with sentence and possibly word alignments
- **Step 1:** Find appropriate data
- **Step 2:** Pre-processing
- **Step 3:** Sentence alignment
 - Length-based methods such as Church and Gale perform well
 - Dynamic programming required to make search efficient
- **Step 4:** Word alignment
 - IBM models allow us to calculate the probability of possible alignments
 - Chicken and egg problem: we need alignments to calculate translation probabilities and vice-versa
 - Solution: EM algorithm
- Next up: Lab on parallel corpora and alignment, then phrase-based SMT