

CS447: Natural Language Processing

<http://courses.engr.illinois.edu/cs447>

# Lecture 22:

# Statistical

# Machine Translation

Julia Hockenmaier

*juliahmr@illinois.edu*

3324 Siebel Center

# Projects and Literature Reviews

*First* report due Nov 15

(PDF written in LaTeX; no length restrictions;  
submission through Compass)

Purpose of this first report:

Check-in to make sure that you're on track  
(or, if not, that we can spot problems)

Get feedback from your peers

Rubrics for the *final* reports (due on Reading Day):

<https://courses.engr.illinois.edu/CS447/LiteratureReviewRubric.pdf>

<https://courses.engr.illinois.edu/CS447/FinalProjectRubric.pdf>

# Projects and Literature Reviews

## Guidelines for first **Project Report**:

What is your project about?

What are the relevant papers you are building on?

What data are you using?

What evaluation metric will you be using?

What models will you implement/evaluate?

What is your to-do list?

## Guidelines for first **Literature Review Report**:

What is your literature review about?

(What task or what kind of models?

Do you have any specific questions or focus?)

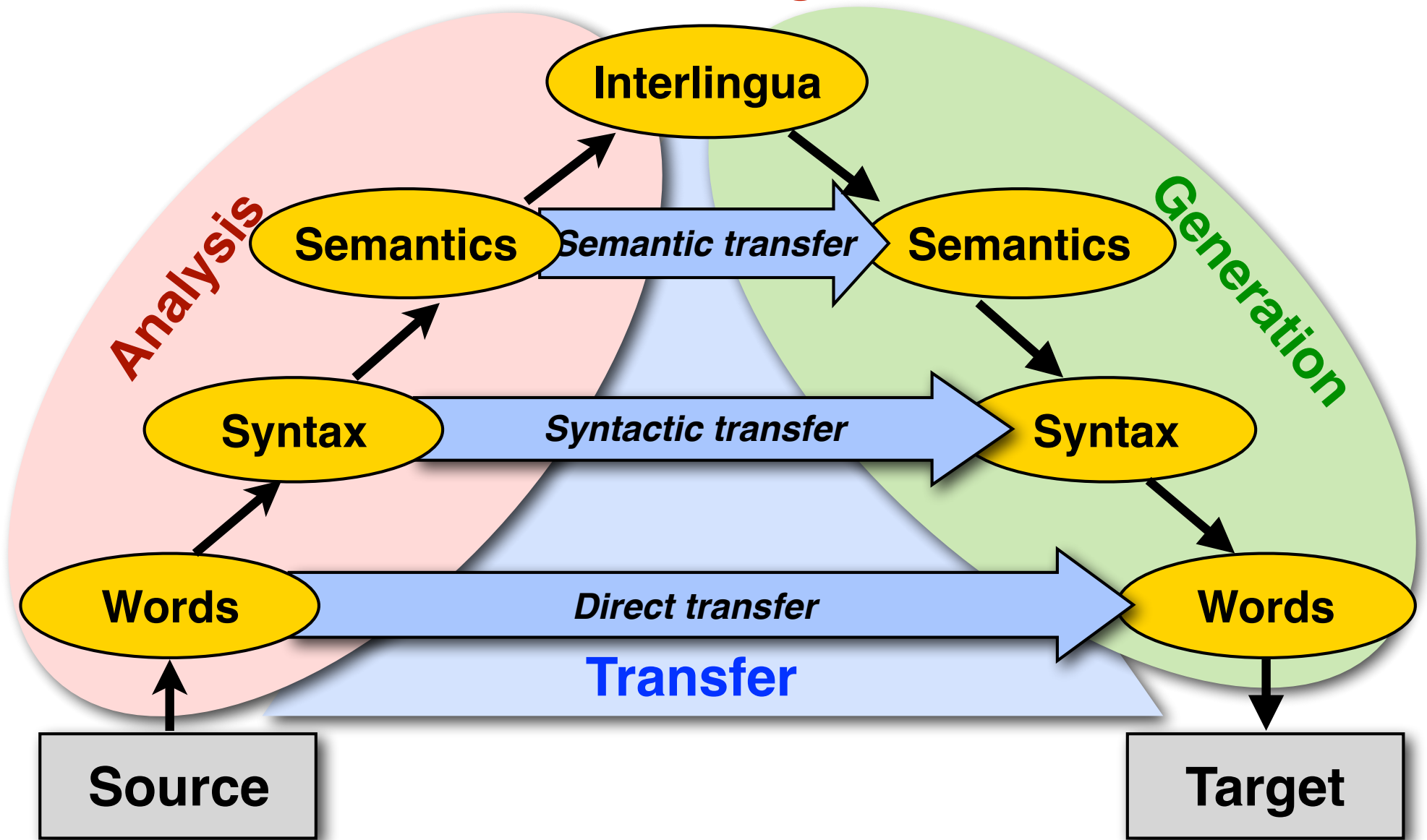
What are the papers you will review?

(If you already have it, give a brief summary of each of them)

What's your to-do list?

# Machine translation approaches

# The Vauquois triangle



# Direct translation

Maria non dió una bofetada a la bruja verde.

## 1. Morphological analysis of source string

Maria non<sub>Neg</sub> dar<sub>3sgF-Past</sub> una bofetada a la bruja verde  
(usually, a complete morphological analysis)

## 2. Lexical transfer (using a translation dictionary):

Mary not slap<sub>3sgF-Past</sub> to the witch green.

## 3. Local reordering:

Mary not slap<sub>3sgF-Past</sub> the green witch.

## 4. Morphology:

Mary did not slap the green witch.

# Limits of direct translation: Phrasal reordering

## Adverb placement in German:

The green witch is at home this week.

Diese Woche ist die grüne Hexe zuhause.



## Japanese SOV order:

He adores listening to music

Kare ha ongaku wo kiku no ga daisuki desu



## PPs in Chinese:

Jackie Cheng went to Hong Kong

Cheng Long dao Xianggang qu



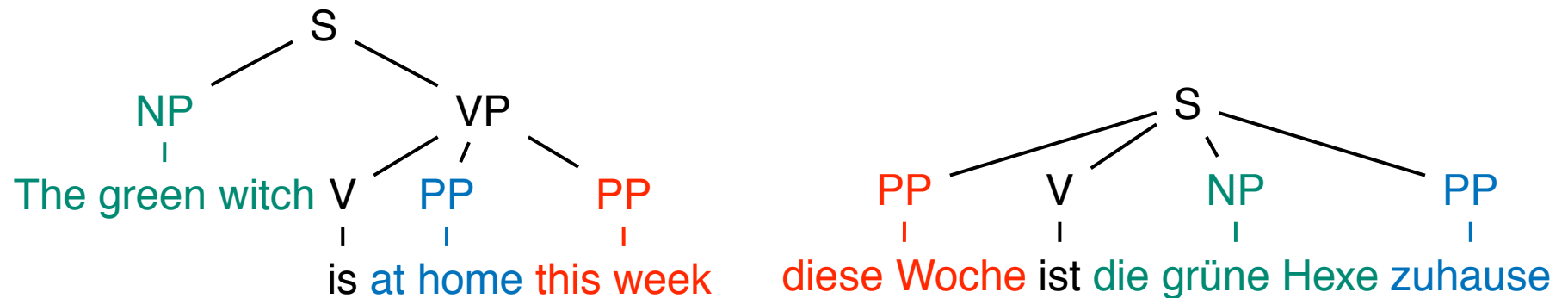
# Syntactic transfer

Requires a syntactic parse of the source language, followed by reordering of the tree

## Local reordering:



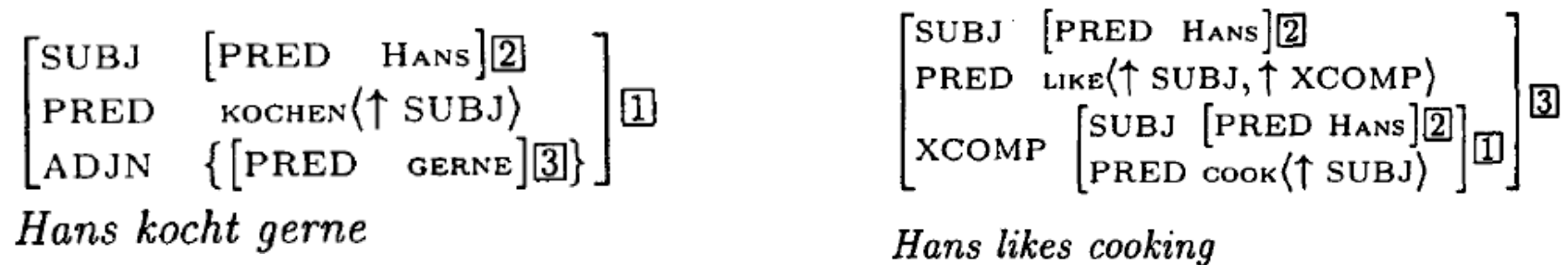
## Nonlocal reordering:



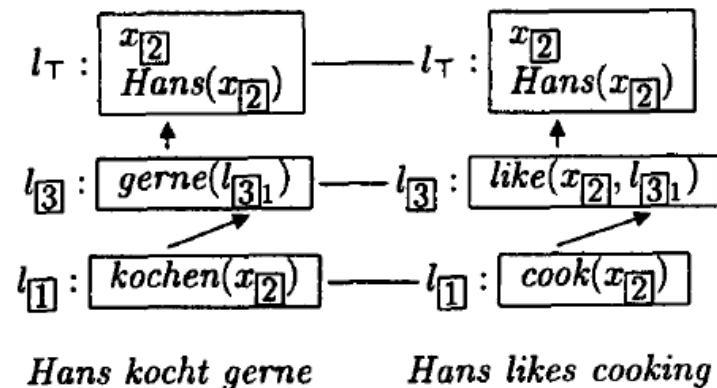


# Semantic transfer

Done at the level of **predicate-argument structure**  
(some people call this syntactic transfer too...):



or at the level of **semantic representations** (e.g. DRSs):



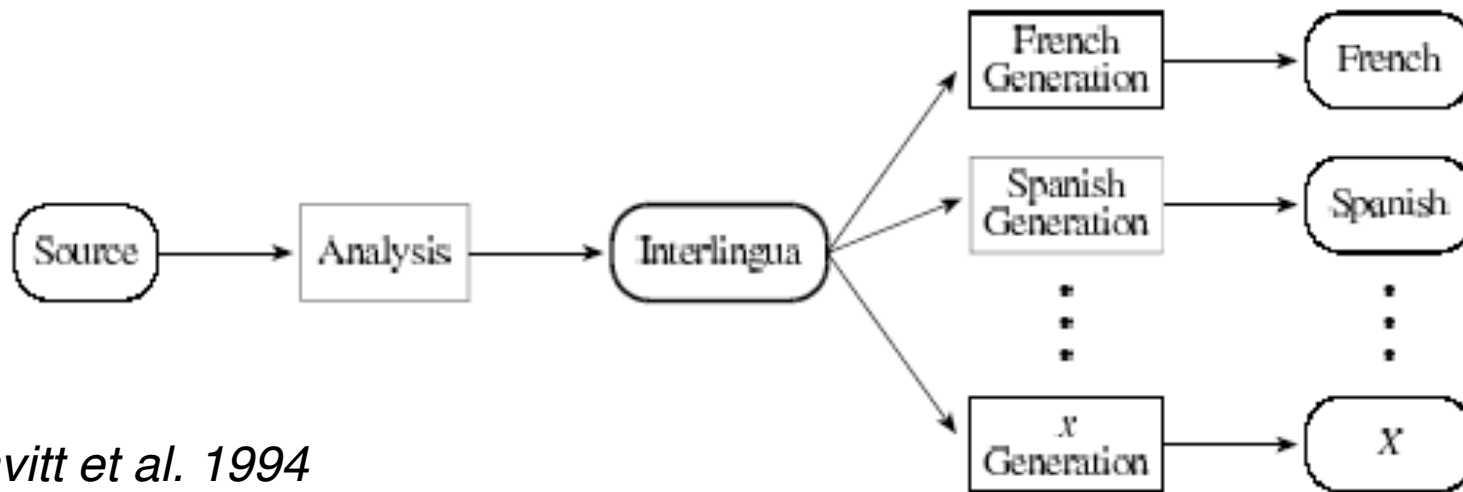
*Dorna et al. 1998*

# Interlingua approaches

Based on the assumption that there is one **common meaning representation** (e.g. predicate logic) that abstracts away from *any* difference in surface realization.

Semantic transfer: each language produces its own meaning representation

Was thought useful for multilingual translation



*Leavitt et al. 1994*

# Statistical Machine Translation with the noisy channel model

# The noisy channel model

Assume we want to **translate from Chinese to English**.

Chinese is the **source**, English is the **target**

The “**noisy channel**” **metaphor** assumes that the Chinese source text is English text that passed through a “noisy channel” which scrambled the English text such that it came out in Chinese.

... the **noisy channel** is a probability model  $P(\textit{Chinese} \mid \textit{English})$

... the **target** text is the **input to the noisy channel**

... the **source** text is the **output of the noisy channel**

... translation is the task of recovering the most likely original input (i.e. the target text) to the noisy channel, given the output (source text)

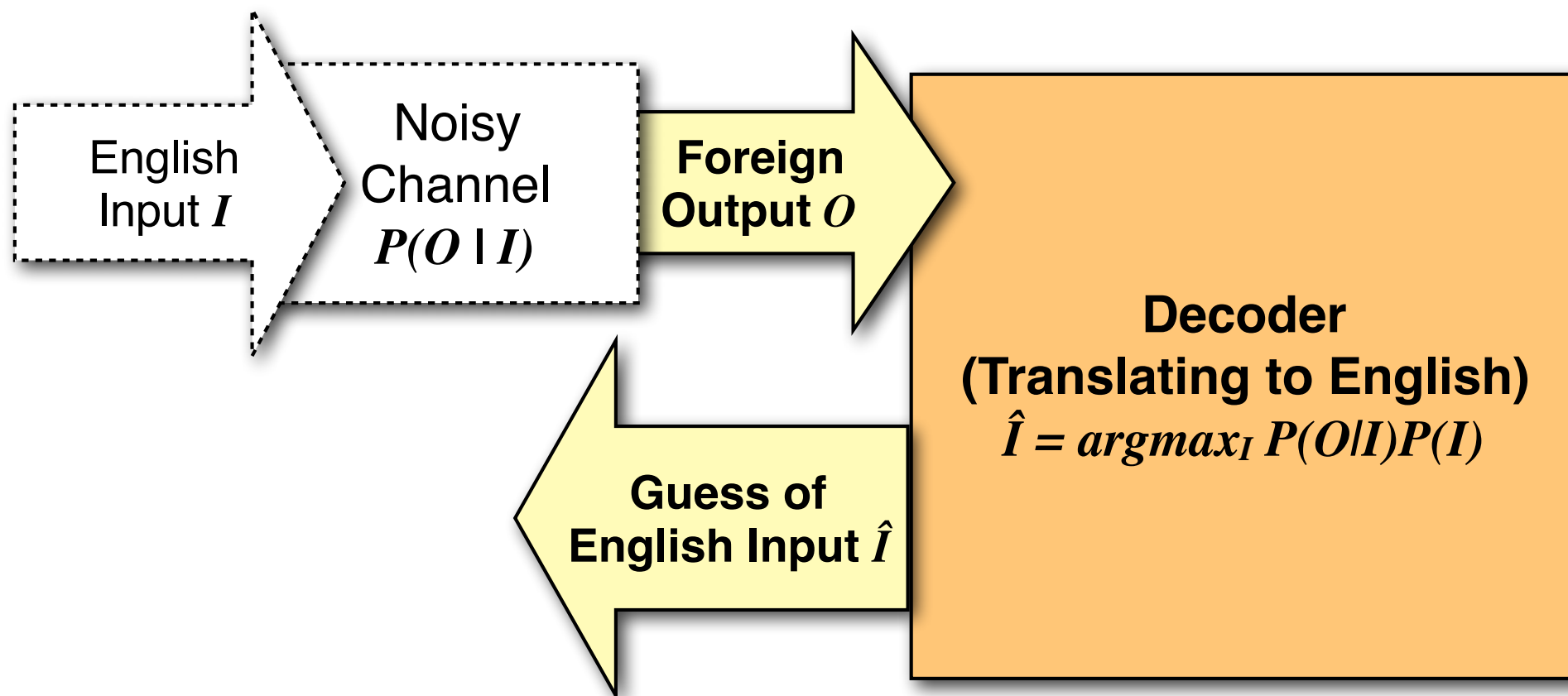
... this also requires a model  $P(\textit{English})$  over the possible input messages

$$\operatorname{argmax}_{Eng} P(Eng \mid Chin) = \operatorname{argmax}_{Eng} \underbrace{P(Chin \mid Eng)}_{\text{Translation Model}} \times \underbrace{P(Eng)}_{\text{LanguageModel}}$$

# The noisy channel model

## Translating from Chinese to English:

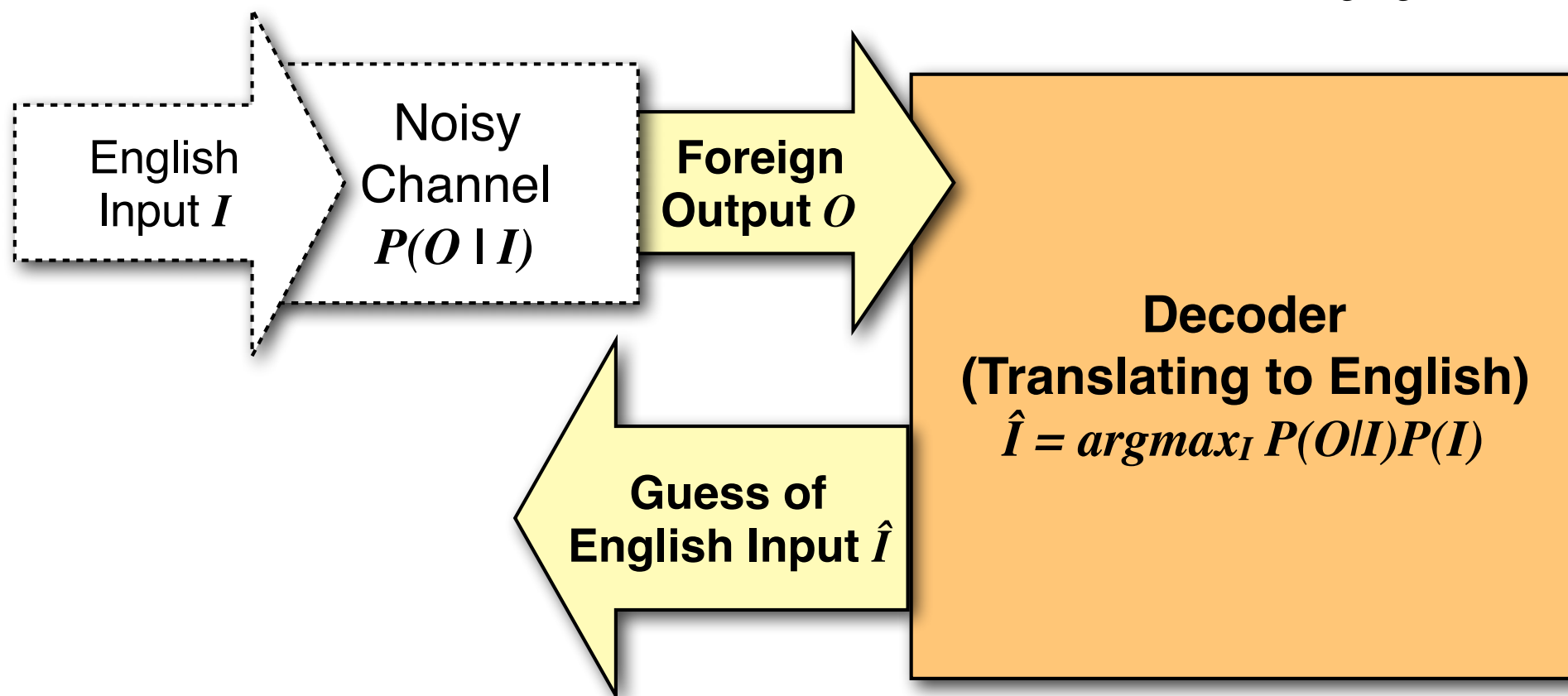
$$\operatorname{argmax}_{Eng} P(Eng|Chin) = \operatorname{argmax}_{Eng} \underbrace{P(Chin|Eng)}_{\text{Translation Model}} \times \underbrace{P(Eng)}_{\text{LanguageModel}}$$



# The noisy channel model

Translating from Chinese (Source) to English (Target):

$$\operatorname{argmax}_{Eng} P(Eng|Chin) = \operatorname{argmax}_{Eng} \underbrace{P(Chin|Eng)}_{\text{Translation Model}} \times \underbrace{P(Eng)}_{\text{LanguageModel}}$$



# The noisy channel model

This is really just an application of **Bayes' rule**:

$$\begin{aligned}\hat{E} &= \arg \max_E P(E|F) \\ &= \arg \max_E \frac{P(F|E) \times P(E)}{P(F)} \\ &= \arg \max_E \underbrace{P(F|E)}_{\text{Translation Model}} \times \underbrace{P(E)}_{\text{Language Model}}\end{aligned}$$

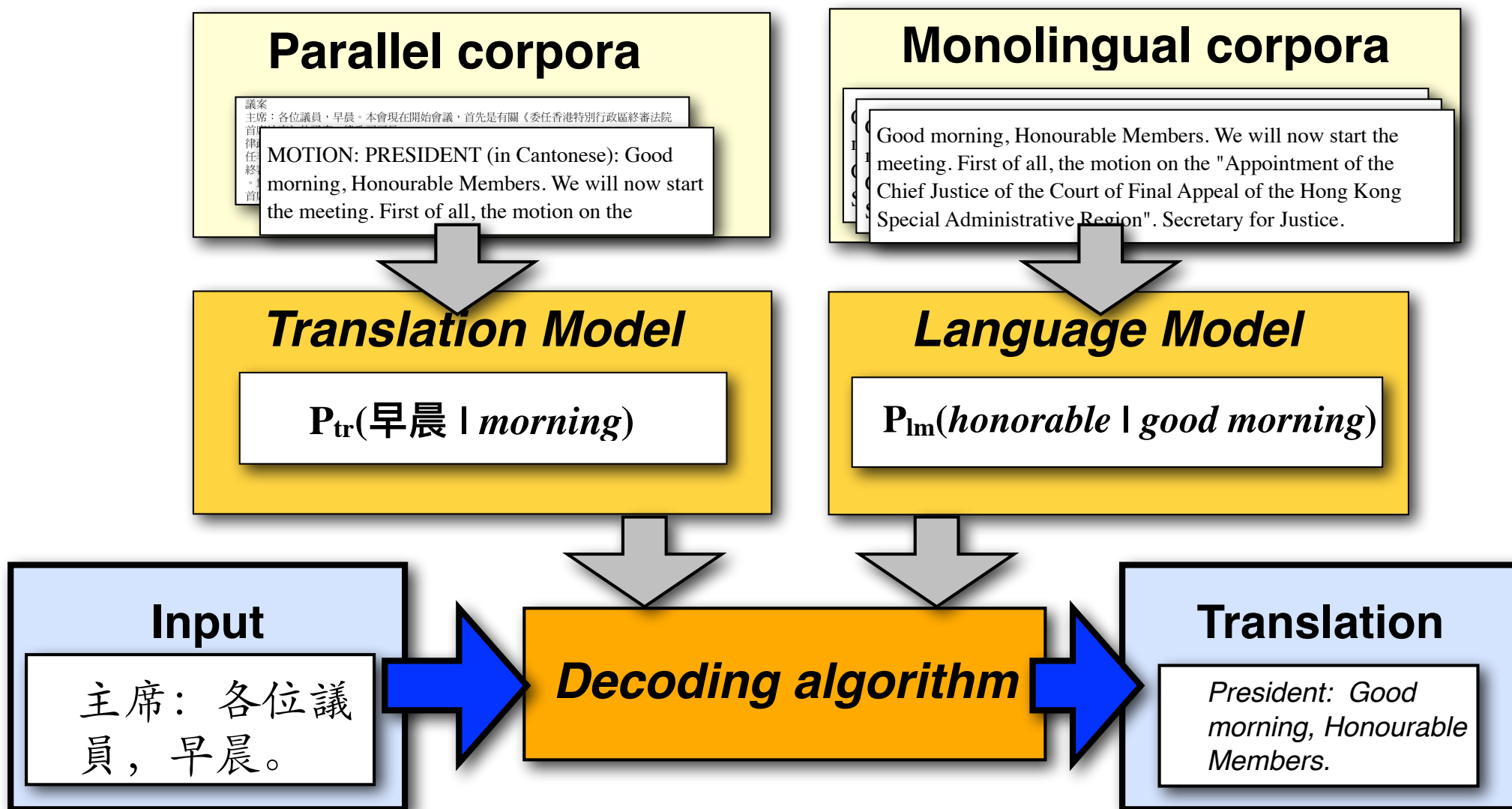
The **translation model**  $P(F|E)$  is intended to capture the **faithfulness of the translation**.

It needs to be trained on a **parallel corpus**

The **language model**  $P(E)$  is intended to capture the **fluency of the translation**.

It can be trained on a (very large) **monolingual corpus**

# Statistical MT with the noisy channel model





# *n*-gram language models for MT

With training on data from the web and clever parallel processing (MapReduce/Bloom filters),  $n$  can be quite large

- Google (2007) uses 5-grams to 7-grams,
- This results in huge models, but the effect on translation quality levels off quickly:

## Size of models

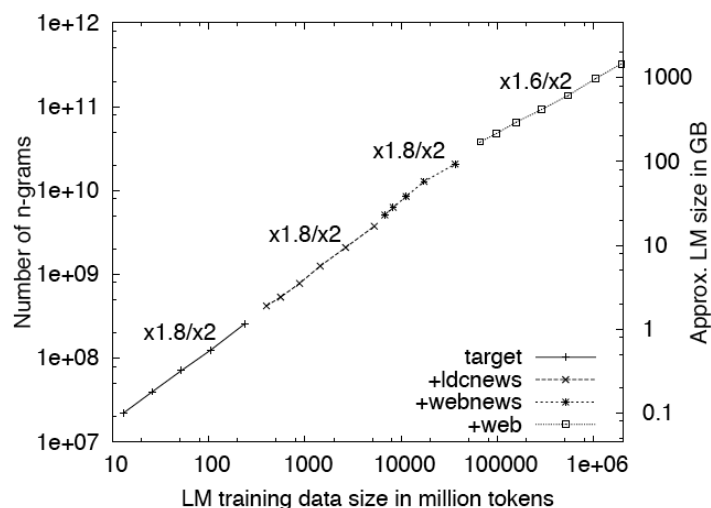


Figure 3: Number of  $n$ -grams (sum of unigrams to 5-grams) for varying amounts of training data.

## Effect on translation quality

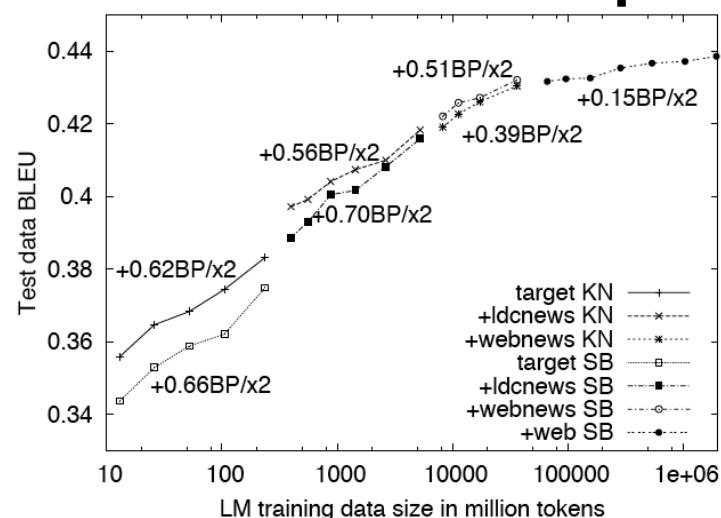


Figure 5: BLEU scores for varying amounts of data using Kneser-Ney (KN) and Stupid Backoff (SB).

# Translation probability $P(fp_i | ep_i)$

Phrase translation probabilities can be obtained from a **phrase table**:

EP	FP	count
green witch	grüne Hexe	...
at home	zu Hause	10534
at home	daheim	9890
is	ist	598012
this week	diese Woche	....

This requires **phrase alignment** on a parallel corpus.

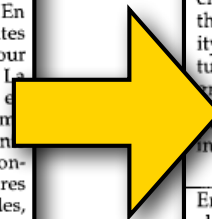
# Creating parallel corpora

A **parallel corpus** consists of the same text in two (or more) languages.

Examples: Parliamentary debates: Canadian Hansards; Hong Kong Hansards, Europarl; Movie subtitles (OpenSubtitles)

In order to train translation models, we need to **align the sentences** (Church & Gale '93)

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.



English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

# IBM models

First statistical MT models, based on noisy channel:

Translate from source  $f$  to target  $e$

via a **translation model**  $P(f | e)$  and a **language model**  $P(e)$

The translation model goes **from target  $e$  to source  $f$**

via **word alignments**  $a$ :  $P(f | e) = \sum_a P(f, a | e)$

Original purpose: Word-based translation models

Today: Can be used to obtain word alignments,  
which are then used to obtain phrase alignments  
for phrase-based translation models

## Sequence of 5 translation models

Model 1 is too simple to be used by itself,

but can be trained very easily on parallel data.

# IBM translation models: assumptions

The model “generates” the ‘foreign’ source sentence  $f$  conditioned on the ‘English’ target sentence  $e$  by the following stochastic process:

1. Generate the **length** of the source  $f$  with probability  $p = \dots$
2. Generate the **alignment** of the source  $f$  to the target  $e$  with probability  $p = \dots$
3. Generate the **words** of the source  $f$  with probability  $p = \dots$

# Word alignment

# Word alignment

John loves Mary.

↓ ↓ ↓  
*Jean aime Marie.*

... that John loves Mary.

↓ ↓ ↘ ↙  
*... dass John Maria liebt.*

	Jean	aime	Marie
John			
loves			
Mary			

	dass	John	Maria	liebt
that				
John				
loves				
Mary				

# Word alignment

	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary									
did									
not									
slap									
the									
green									
witch									



# Word alignment

	Marie	a	traversé	le	lac	à	la	nage
Mary								
swam								
across								
the								
lake								

# Word alignment

		Source							
Target		Marie	a	traversé	le	lac	à	la	nage
	Mary								
	swam								
	across								
	the								
	lake								

**One target word** can be aligned to **many source words**.

# Word alignment

		Source							
Target		Marie	a	traversé	le	lac	à	la	nage
	Mary								
	swam								
	across								
	the								
	lake								

**One target word** can be aligned to **many source words**.  
But **each source word** can only be aligned to **one target word**.

# Word alignment

		Source							
Target		Marie	a	traversé	le	lac	à	la	nage
	Mary								
	swam								
	across								
	the								
	lake								

Alignments are **not symmetric!**

**Some source words** may **not align** to **any target words**.

# Word alignment

Source

Target

	Marie	a	traversé	le	lac	à	la	nage
NULL								
Mary								
swam								
across								
the								
lake								

Variant: assume a NULL word in the target language.

# Representing word alignments

		1	2	3	4	5	6	7	8
		Marie	a	traversé	le	lac	à	la	nage
0	NULL								
1	Mary								
2	swam								
3	across								
4	the								
5	lake								



Position	1	2	3	4	5	6	7	8
Foreign	Marie	a	traversé	le	lac	à	la	nage
Alignment	1	3	3	4	5	0	0	2

Every source word  $f[i]$  is aligned to one target word  $e[j]$  (incl. NULL). We represent alignments as a vector  $\mathbf{a}$  (of the same length as the source) with  $\mathbf{a}[i] = j$

# The IBM alignment models

# The IBM models

Use the noisy channel (Bayes rule) to get the best (most likely) target translation  $e$  for source sentence  $f$ :

$$\arg \max_e P(e|f) = \arg \max_e P(f|e)P(e)$$

noisy channel

The translation model  $P(f|e)$  requires alignments  $a$

$$P(f|e) = \sum_{a \in \mathcal{A}(e,f)} P(f, a|e)$$

marginalize (=sum)  
over all alignments  $a$

Generate  $f$  and the alignment  $a$  with  $P(f, a|e)$ :

$$P(f, a|e) = \underbrace{P(m|e)}_{\text{Length: } |f|=m} \prod_{j=1}^m \underbrace{P(a_j | a_{1..j-1}, f_{1..j-1}, m, e)}_{\text{Word alignment } a_j} \underbrace{P(f_j | a_{1..j} f_{1..j-1}, e, m)}_{\text{Translation } f_j}$$

$m = \# \text{ words}$   
in  $f_j$

probability of  
alignment  $a_j$

probability  
of word  $f_j$



# Model parameters

## Length probability $P(m \mid n)$ :

What's the probability of generating a source sentence of length  $m$  given a target sentence of length  $n$ ?

Count in training data

## Alignment probability: $P(\mathbf{a} \mid m, n)$ :

Model 1 assumes all alignments have the same probability:

For each position  $a_1 \dots a_m$ , pick one of the  $n+1$  target positions uniformly at random

## Translation probability: $P(f_j = lac \mid a_j = i, e_i = lake)$ :

In Model 1, these are the only parameters we have to learn.

# IBM model 1: Generative process

For each target sentence  $e = e_1..e_n$  of length  $n$ :

0	1	2	3	4	5
NULL	Mary	swam	across	the	lake

1. **Choose a length  $m$**  for the source sentence (e.g  $m = 8$ )

Position	1	2	3	4	5	6	7	8
----------	---	---	---	---	---	---	---	---

2. **Choose an alignment  $a = a_1...a_m$**  for the source sentence

Each  $a_j$  corresponds to a word  $e_i$  in  $e$ :  $0 \leq a_j \leq n$

Position	1	2	3	4	5	6	7	8
Alignment	1	3	3	4	5	0	0	2

3. **Translate each target word  $e_{a_j}$**  into the source language

Position	1	2	3	4	5	6	7	8
Alignment	1	3	3	4	5	0	0	2
Translation	Marie	a	traversé	le	lac	à	la	nage

# IBM model 1: details

The **length probability** is constant:  $P(m | e) = \epsilon$

The **alignment probability** is uniform

( $n$  = length of target string):  $P(a_i | e) = 1/(n+1)$

The **translation probability** depends only on  $e_{a_i}$

(the corresponding target word):  $P(f_i | e_{a_i})$

$$\begin{aligned} P(\mathbf{f}, \mathbf{a} | \mathbf{e}) &= \underbrace{P(m | \mathbf{e})}_{\text{Length: } |\mathbf{f}|=m} \prod_{j=1}^m \underbrace{P(a_j | a_{1..j-1}, f_{1..j-1}, m, \mathbf{e})}_{\text{Word alignment } a_j} \underbrace{P(f_j | a_{1..j} f_{1..j-1}, \mathbf{e}, m)}_{\text{Translation } f_j} \\ &= \epsilon \prod_{j=1}^m \frac{1}{n+1} P(f_j | e_{a_j}) \\ &= \frac{\epsilon}{(n+1)^m} \prod_{j=1}^m P(f_j | e_{a_j}) \end{aligned}$$

# Finding the best alignment

How do we find the **best alignment** between **e** and **f**?

$$\begin{aligned}\hat{\mathbf{a}} &= \arg \max_{\mathbf{a}} P(\mathbf{f}, \mathbf{a} | \mathbf{e}) \\ &= \arg \max_{\mathbf{a}} \frac{\epsilon}{(n+1)^m} \prod_{j=1}^m P(f_j | e_{a_j}) \\ &= \arg \max_{\mathbf{a}} \prod_{j=1}^m P(f_j | e_{a_j})\end{aligned}$$

$$\hat{a}_j = \arg \max_{a_j} P(f_j | e_{a_j})$$

# Learning translation probabilities

The only parameters that need to be learned are the **translation probabilities**  $P(f | e)$

$$P(f_j = lac \mid e_i = lake)$$

If the training corpus had word alignments, we could simply count how often ‘lake’ is aligned to ‘lac’:

$$P(lac \mid lake) = \text{count}(lac, lake) / \sum_w \text{count}(w, lake)$$

But we don’t have word alignments.

So, instead of relative frequencies, we have to use *expected* relative frequencies:

$$P(lac \mid lake) = \langle \text{count}(lac, lake) \rangle / \langle \sum_w \text{count}(w, lake) \rangle$$

# Training Model 1 with EM

The only parameters that need to be learned are the **translation probabilities**  $P(f | e)$

We use the **EM algorithm** to estimate these parameters from a corpus with  $S$  sentence pairs  $s = \langle f^{(s)}, e^{(s)} \rangle$  with alignments  $\mathcal{A}(f^{(s)}, e^{(s)})$

- **Initialization:** guess  $P(f | e)$
- **Expectation step:** compute expected counts

$$\langle c(f, e) \rangle = \sum_{s \in S} \langle c(f, e | e^{(s)}, f^{(s)}) \rangle$$

- **Maximization step:** recompute probabilities  $P(f | e)$

$$\hat{P}(f | e) = \frac{\langle c(f, e) \rangle}{\sum_{f'} \langle c(f', e) \rangle}$$

# Expectation-Maximization (EM)

1. Initialize a first model,  $M_0$

2. **Expectation (E) step:**

Go through training data to gather expected counts  
 $\langle \text{count}(lac, lake) \rangle$

3. **Maximization (M) step:**

Use expected counts to compute a new model  $M_{i+1}$   
 $P_{i+1}(lac | lake) = \langle \text{count}(lac, lake) \rangle / \langle \sum w \text{count}(w, lake) \rangle$

4. **Check for convergence:**

Compute log-likelihood of training data with  $M_{i+1}$   
If the difference between new and old log-likelihood  
smaller than a threshold, stop. Else go to 2.

# The E-step

Compute the expected count  $\langle c(f, e | \mathbf{f}, \mathbf{e}) \rangle$ :

$$\langle c(f, e | \mathbf{f}, \mathbf{e}) \rangle = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{f}, \mathbf{e})} P(\mathbf{a} | \mathbf{f}, \mathbf{e}) \cdot \underbrace{c(f, e | \mathbf{a}, \mathbf{e}, \mathbf{f})}_{\text{How often are } f, e \text{ aligned in } \mathbf{a} ?}$$

$$P(\mathbf{a} | \mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{a}, \mathbf{f} | \mathbf{e})}{P(\mathbf{f} | \mathbf{e})} = \frac{P(\mathbf{a}, \mathbf{f} | \mathbf{e})}{\sum_{\mathbf{a}'} P(\mathbf{a}', \mathbf{f} | \mathbf{e})}$$

$$P(\mathbf{a}, \mathbf{f} | \mathbf{e}) = \prod_j P(f_j | e_{a_j})$$

$$\langle c(f, e | \mathbf{f}, \mathbf{e}) \rangle = \sum_{\mathbf{a} \in \mathcal{A}(\mathbf{f}, \mathbf{e})} \frac{\prod_j P(f_j | e_{a_j})}{\sum_{\mathbf{a}'} \prod_j P(f_j | e_{a'_j})} \cdot c(f, e | \mathbf{a}, \mathbf{e}, \mathbf{f})$$



# Other translation models

Model 1 is a very simple (and not very good) translation model.

IBM models 2-5 are more complex. They take into account:

- “**fertility**”: the number of foreign words generated by each target word
- the **word order** and **string position** of the aligned words

# Today's key concepts

Why is machine translation hard?

Linguistic divergences: morphology, syntax, semantics

Different approaches to machine translation:

Vauquois triangle

Statistical MT: Noisy Channel, IBM Model 1 (more on this next time)