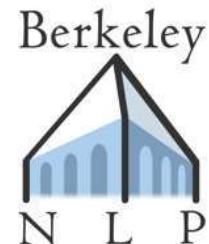


Machine Translation



Dan Klein
UC Berkeley

Many slides from John DeNero and Philip Koehn

Translation Task

- Text is both the input and the output.
- Input and output have roughly the same information content.
- Output is more predictable than a language modeling task.
- Lots of naturally occurring examples (but not much metadata).

Translation Examples

English-German News Test 2013 (a standard dev set)

Republican leaders justified their policy by the need to combat electoral fraud.

Die Führungskräfte der Republikaner
| | | |
The Executives of the republican

rechtfertigen ihre Politik mit der
| | | |
justify your politics With of the

Notwendigkeit , den Wahlbetrug zu
| | | |
need , the election fraud to

bekämpfen .
| |
fight .

Variety in Translations?

Human-generated reference translation

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomists got to know this incident 4 days later. This small planet is 50m in diameter. The astonomists are hard to find it for it comes from the direction of sun.

A commercial system from 2002

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

Google Translate, 2020

An asteroid that was large enough to destroy a medium-sized city, swept across the earth at a short distance of 463,000 kilometers, but was not detected early. Astronomers learned about it four days later. The asteroid is about 50 meters in diameter and comes from the direction of the sun, making it difficult for astronomers to spot it.

From <https://catalog.ldc.upenn.edu/LDC2003T17>

Evaluation

BLEU Score

BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty (harshly penalizes translations shorter than the reference).

$$\text{Matched}_i = \sum_{t_i} \min \left\{ C_h(t_i), \max_j C_j(t_i) \right\}$$

If "of the" appears twice in hypothesis h but only at most once in a reference, then only the first is "correct"

$$P_i = \frac{\text{Matched}_i}{H_i}$$

"Clipped" precision of n-gram tokens

$$B = \exp \left\{ \min \left(0, \frac{n - L}{n} \right) \right\}$$

Brevity penalty only matters if the hypothesis **corpus** is shorter than the sum of (shortest) references.

$$\text{BLEU} = B \left(\prod_{i=1}^4 P_i \right)^{\frac{1}{4}}$$

BLEU is a mean of clipped precisions, scaled down by the brevity penalty.

Evaluation with BLEU

In this sense, the measures will partially undermine the American democratic system.

In this sense, these measures partially undermine the democratic system of the United States.



BLEU = 26.52, 75.0/40.0/21.4/7.7 (BP=1.000, ratio=1.143, hyp_len=16, ref_len=14)

(Papineni et al., 2002) BLEU: a method for automatic evaluation of machine translation.

Corpus BLEU Correlations with Average Human Judgments

These are ecological correlations over multiple segments; segment-level BLEU scores are noisy.

Commercial machine translation providers seem to all perform human evaluations of some sort.

(Ma et al., 2019) Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

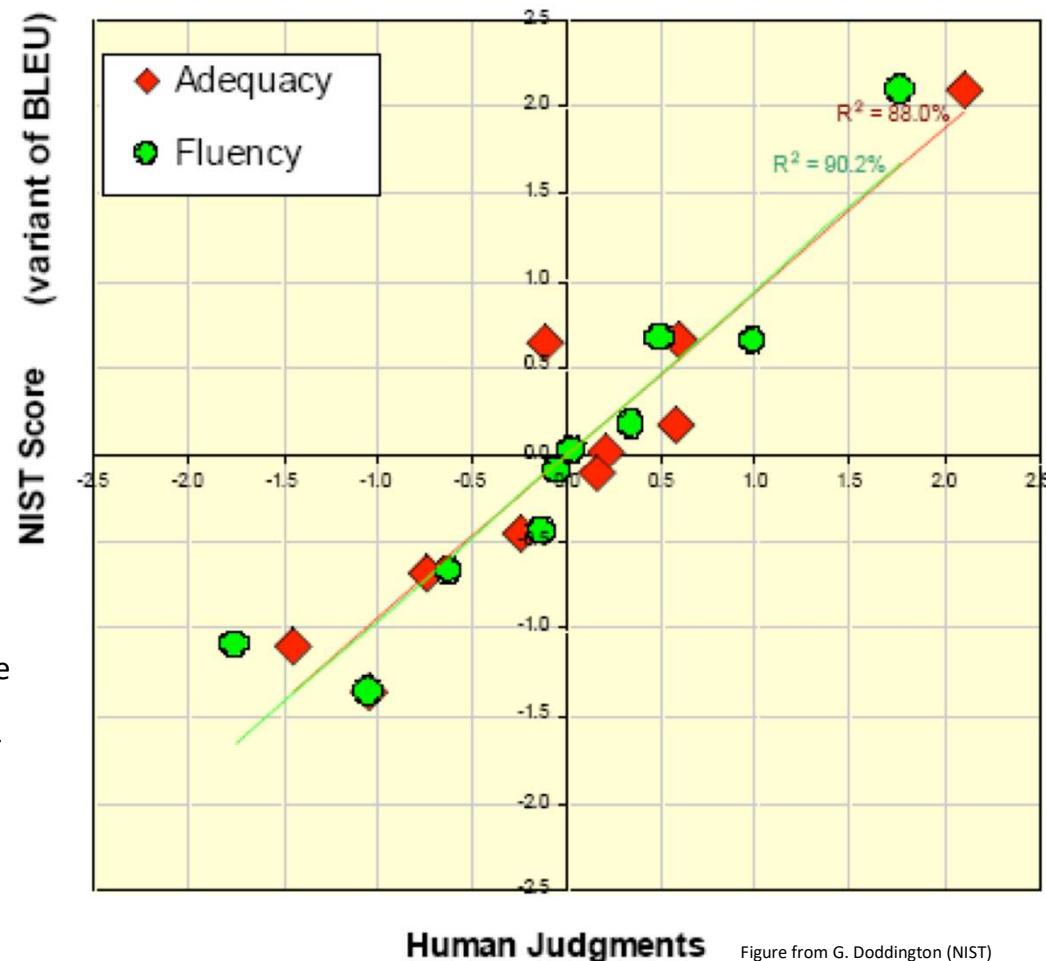


Figure from G. Doddington (NIST)

Human Evaluations

Direct assessment: adequacy & fluency

- Monolingual: Ask humans to compare machine translation to a human-generated reference. (Easier to source annotators)
- Bilingual: Ask humans to compare machine translation to the source sentence that was translated. (Compares to human quality)
- Annotators can assess segments (sentences) or whole documents.
- Segments can be assessed with or without document context.

Ranking assessment:

- Raters are presented with 2 or more translations.
- A human-generated reference may be provided, along with the source.
- "In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences." (Laubli et al., 2018)

Editing assessment: How many edits required to reach human quality

(Laubli et al., 2018) Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

1/12 documents, 4 items left in document WMT20DocSrcDA #214: Doc. #seattle_times.7674-2 English → German (deutsch)

Below you see a document with 6 sentences in English and their corresponding candidate translations in German (deutsch). Score each candidate translation in the document context, answering the question:
How accurately does the candidate text (right column, in bold) convey the original semantics of the source text (left column) in the document context?
You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Expand all items Expand unannotated Collapse all items

▼ Man gets prison after woman finds bullet in her skull	Der Mann wird gefangen, nachdem die Frau in ihrem Schädel geschossen ist	✓
▼ A Georgia man has been sentenced to 25 years in prison for shooting his girlfriend, who didn't realize she survived a bullet to the brain until she went to the hospital for treatment of headaches.	Ein georgischer Mann wurde zu 25 Jahren Gefängnis verurteilt, weil er seinen Freund geschossen hat, der nicht gewusst hatte, dass er eine Kugel ins Gehirn überlebte, bis er in das Krankenhaus zur Behandlung	✓
^ News outlets report 39-year-old Jerrontae Cain was sentenced Thursday on charges including being a felon in possession of a gun in the 2017 attack on 42-year-old Nicole Gordon.	Nachrichtenagenturen-Bericht 39-jährige Jerrontae Cain wurde am Donnerstag wegen Anklage verurteilt, darunter ein Felon im Besitz einer Waffe beim Angriff auf 42-jährige Nicole Gordon im Jahr 2017.	●

← Not at all | | | Perfectly →

Reset Submit

0/10 blocks, 10 items left in block WMT21CTRA #285: Segment #341 English → German (deutsch)

Fakhfakh stepped down the same day the party filed a no-confidence motion against him.
— Source text

How accurately does each of the candidate text(s) below convey the original semantics of the source text above?

Fakhfakh trat am selben Tag zurück, **an dem** die Partei **einen Misstrauensantrag** gegen ihn einreichte.

← Not at all Perfectly →
--

Fachfakh trat am selben Tag zurück, **als** die Partei **ein Misstrauensvotum** gegen ihn einreichte.

← Not at all Perfectly →
--

Reset Show/Hide diff. Match sliders Submit

(Akhbardeh et al., 2021) Findings of the 2021 Conference on Machine Translation

Translationese and Evaluation

Translated text can: (Baker et al., 1993; Graham et al., 2019)

- be more explicit than the original source
- be less ambiguous
- be simplified (lexical, syntactically and stylistically)
- display a preference for conventional grammaticality
- avoid repetition
- exaggerate target language features
- display features of the source language

"If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved."
(Toral et al., 2018)

(Baker et al., 1993) Corpus linguistics and translation studies: Implications and applications.

(Graham et al., 2019) Translationese in Machine Translation Evaluation.

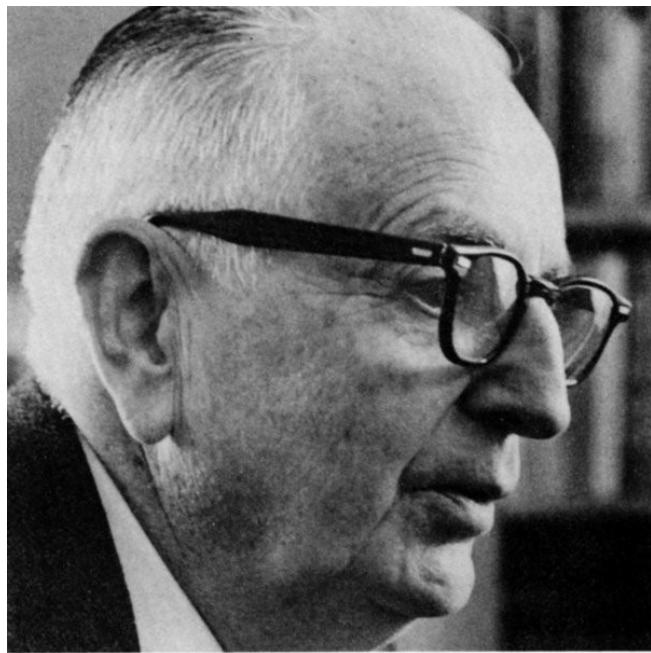
(Toral et al., 2018) Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation

WMT 2019 Evaluation

2019 segment-in-context direct assessment (Barrault et al, 2019):

- ✓ German to English: many systems are tied with human performance;
- ✗ English to Chinese: all systems are outperformed by the human translator;
- ✗ English to Czech: all systems are outperformed by the human translator;
- ✗ English to Finnish: all systems are outperformed by the human translator;
- ✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance;
- ✗ English to Gujarati: all systems are outperformed by the human translator;
- ✗ English to Kazakh: all systems are outperformed by the human translator;
- ✗ English to Lithuanian: all systems are outperformed by the human translator;
- ✓ English to Russian: Facebook-FAIR is tied with human performance.

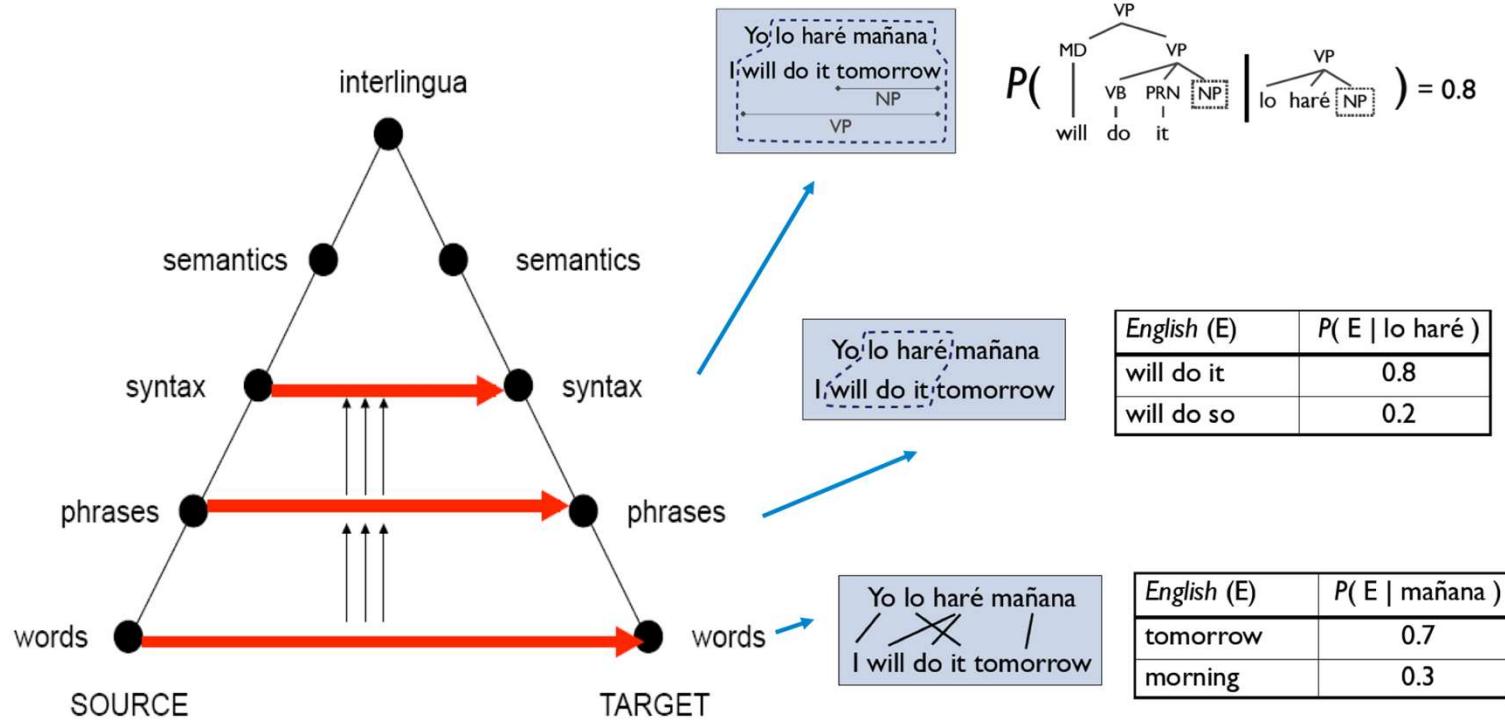
Statistical Machine Translation (1990 - 2015)



When I look at an article in Russian, I say:
“This is really written in English, but it has
been coded in some strange symbols. I
will now proceed to decode.”

Warren Weaver (1949)

Levels of Transfer: Vauquois Triangle (1968)



Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will get to it later

See you later

He will do it

Parallel corpus gives translation examples

I will do it gladly

Yo lo haré de muy buen grado

You will see later

Después lo verás

Machine translation system:

Source language

Yo lo haré después

NOVEL SENTENCE

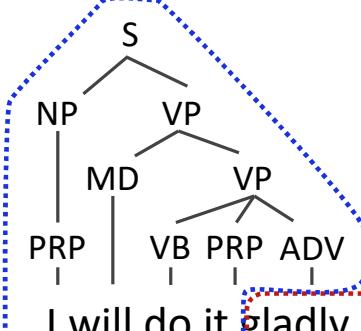
Model of translation

Target language

I will do it later

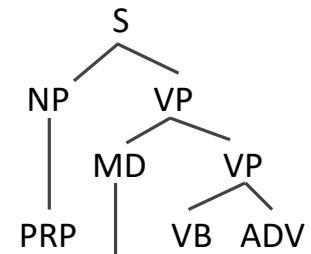
Stitching Together Fragments

Parallel corpus gives translation examples



I will do it gladly

Yo lo haré de muy buen grado



You will see later

Después lo verás

Machine translation system:

S
ADV
Yo lo haré **después**

Model of
translation

S
ADV
I will do it **later**

Evolution of the Noisy Channel Model

$$P(e|f) \propto P(f|e) \cdot P(e)$$

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

$$P(e|f) \propto P(f|e)^{\phi_{\text{tm}}} \cdot P(e)^{\phi_{\text{lm}}}$$

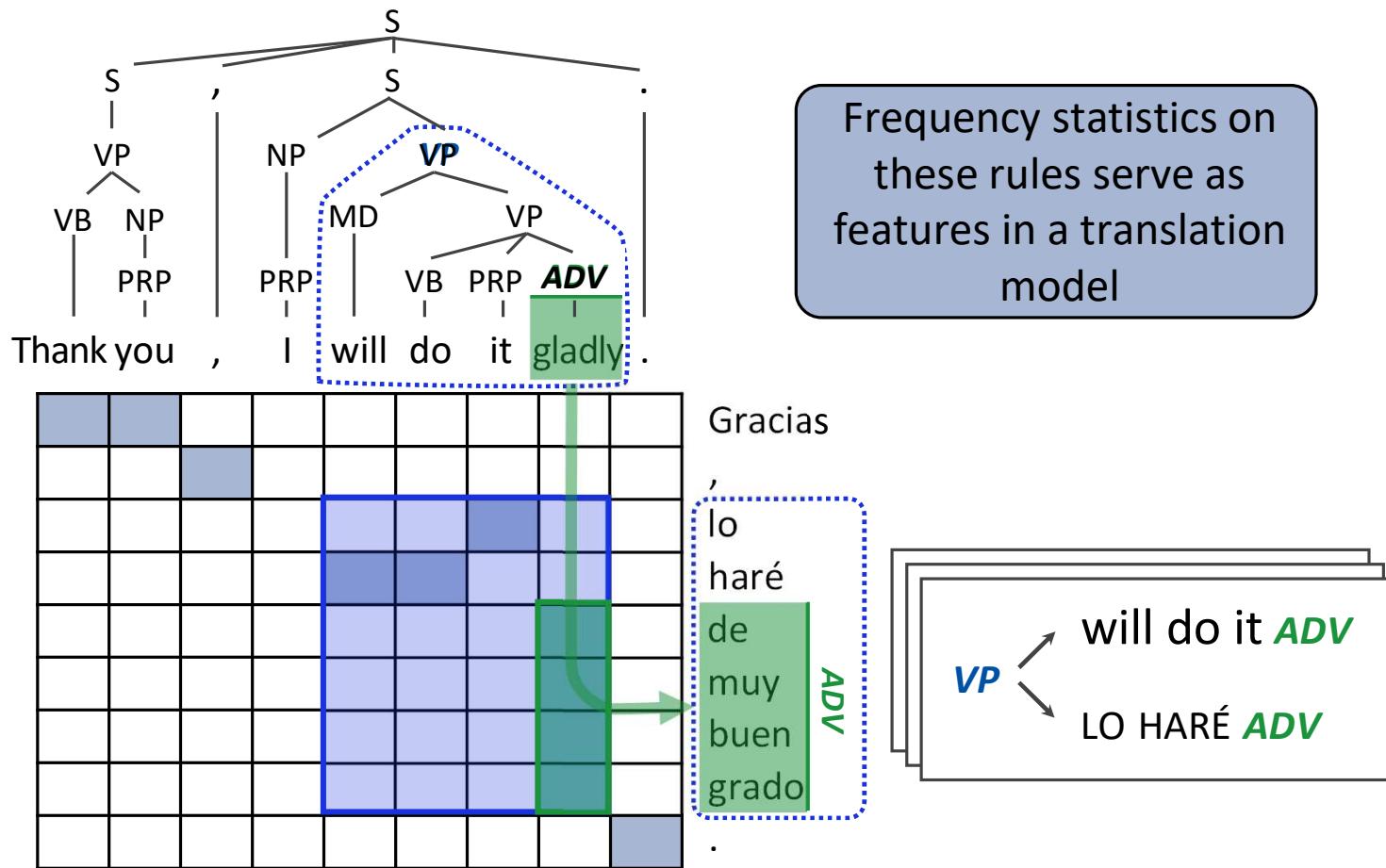
$$P(e|f) \propto \exp \left\{ \sum_i w_i \cdot f_i(e, f) \right\}$$

Chosen to minimize loss

E.g., $\log P(e)$

Word Alignment and Phrase Extraction

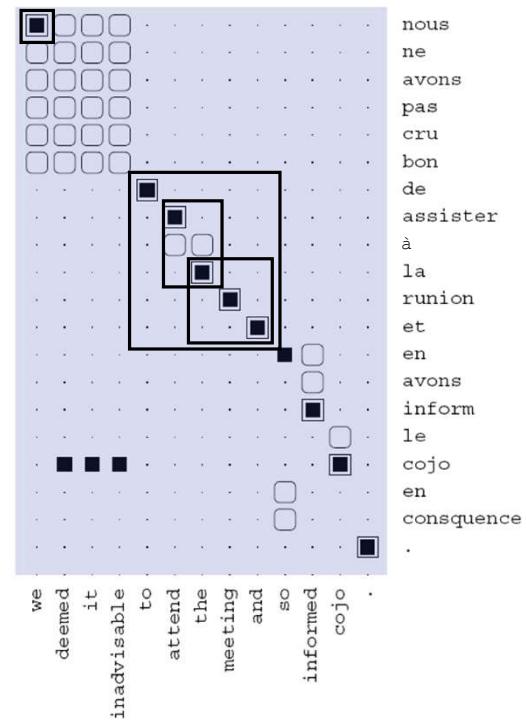
Extracting Translation Rules



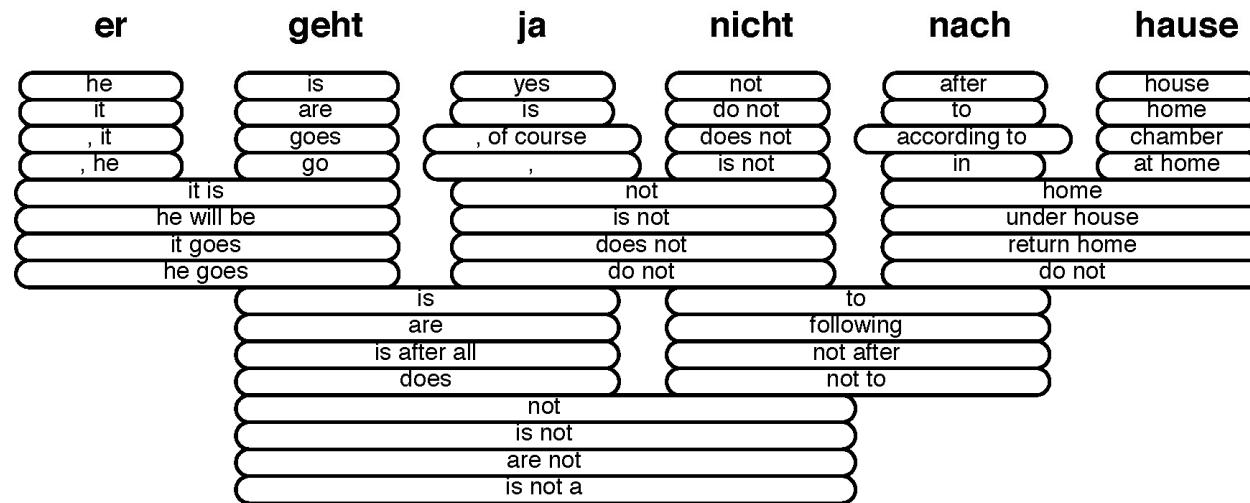
Counting Aligned Phrases

d'assister à la reunion et ||| to attend the meeting and
assister à la reunion ||| attend the meeting
la reunion et ||| the meeting and
nous ||| we
...

- Relative frequencies are the most important features in a phrase-based or syntax-based model.
- Scoring a phrase under a lexical model is the second most important feature.
- Estimation does not involve choosing among segmentations of a sentence into phrases.

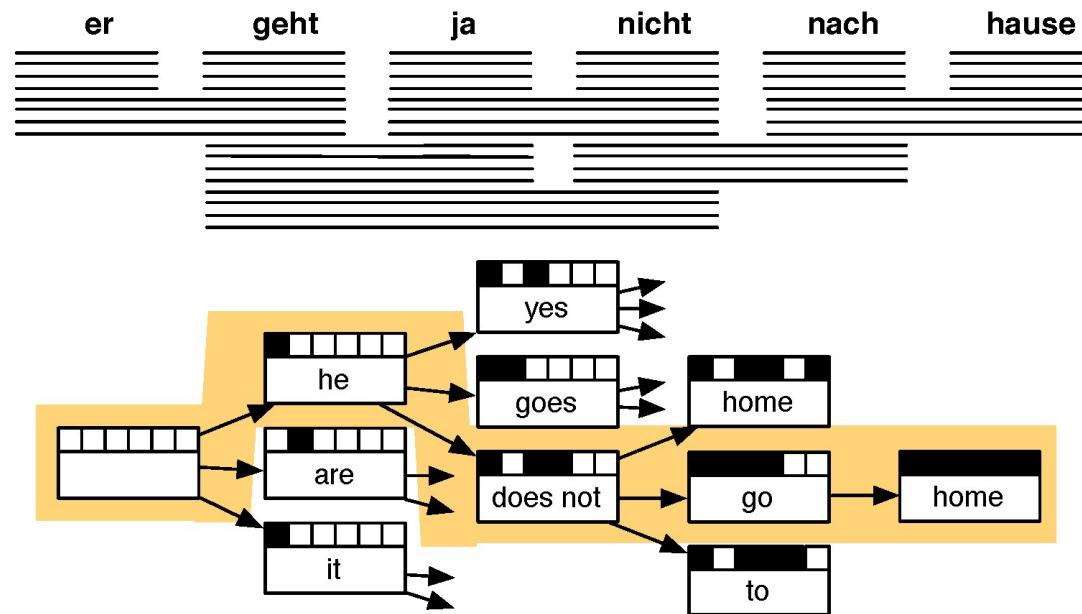


Translation Options



- Many translation options to choose from
 - in Europarl phrase table: 2727 matching phrase pairs for this sentence
 - by pruning to the top 20 per phrase, 202 translation options remain

Decoding: Find Best Path



Phrase-Based Decoding

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员

the	7 people	including	by some	and	the russian	the	the astronauts	,
it	7 people included		by france	and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french	and the russian	the fifth		.
these	7 among	including from		the french and	of the russian	of	space	members .
that	7 persons	including from the		of france and to	russian	of the	aerospace	members .
	7 include		from the	of france and	russian	astronauts		. the
	7 numbers include		from france	and russian	of astronauts who			."
	7 populations include		those from france	and russian		astronauts .		
	7 deportees included		come from	france	and russia	in	astronautical	personnel ;
	7 philtrum	including those from		france and	russia	a space		member
		including representatives from		france and the	russia	astronaut		
		include	came from	france and russia		by cosmonauts		
		include representatives from		french	and russia		cosmonauts	
		include	came from france		and russia 's		cosmonauts .	
		includes	coming from	french and	russia 's		cosmonaut	
				french and russian	's	astronavigation	member .	
				french	and russia	astronauts		
					and russia 's		special rapporteur	
				, and	russia		rapporteur	
					, and russia		rapporteur .	
					, and russia			
				or	russia 's			

Word Alignments

Word Alignment

Given a sentence pair, which words correspond to each other?

michael
geht
davon
aus
,

dass

er

im

haus

bleibt

michael

assumes

that

he

will

stay

in

the

house

Word Alignment?

	john	wohnt	hier	nicht
john	██████	██	██	██
does	██	██ ?	██	██ ?
not	██	██	██	██████
live	██	██████	██	██
here	██	██	██████	██

Is the English word **does** aligned to
the German **wohnt** (verb) or **nicht** (negation) or neither?

Word Alignment?

	john	biss	ins	grass
john	█			
kicked		█		
the			█	
bucket				█

How do the idioms **kicked the bucket** and **biss ins grass** match up?
Outside this exceptional context, **bucket** is never a good translation for **grass**

Lexical Translation / Word Alignment Models

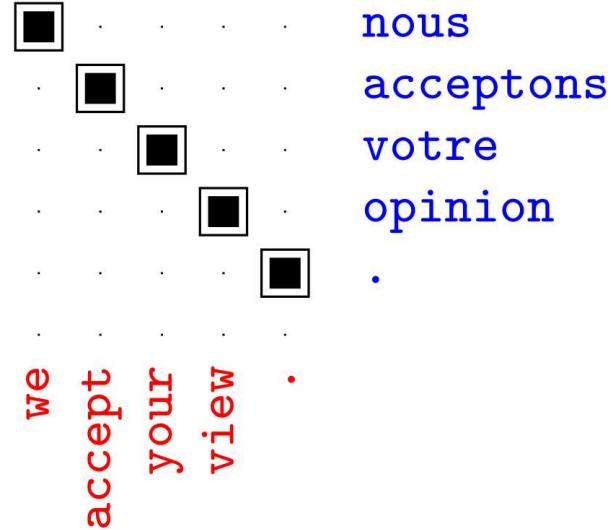


Unsupervised Word Alignment

- Input: a *bitext*: pairs of translated sentences

nous acceptons votre opinion .
we accept your view .

- Output: *alignments*: pairs of translated words
 - When words have unique sources, can represent as a (forward) alignment function a from French to English positions

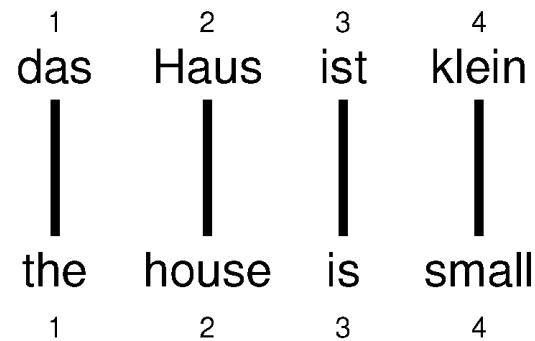


Word Alignment

- Even today models are often built on the IBM alignment models
- Create probabilistic word-level translation models
- The models incorporate latent (unobserved) word alignments
- Optimize the probability of the observed words
- Use the imputed alignments to reveal word-level correspondence
- Throw out the translation models themselves

Alignment

- In a parallel text (or when we translate), we align words in one language with the words in the other



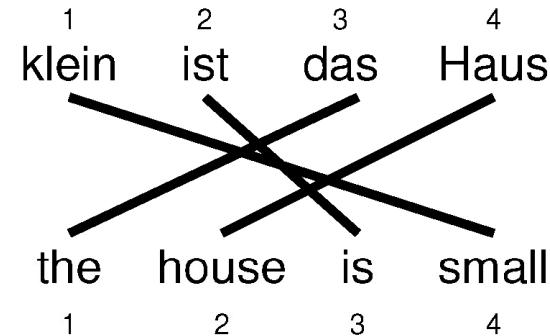
- Word positions are numbered 1–4

Alignment Function

- Formalizing alignment with an alignment function
- Mapping an English target word at position i to a German source word at position j with a function $a : i \rightarrow j$
- Example
$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Reordering

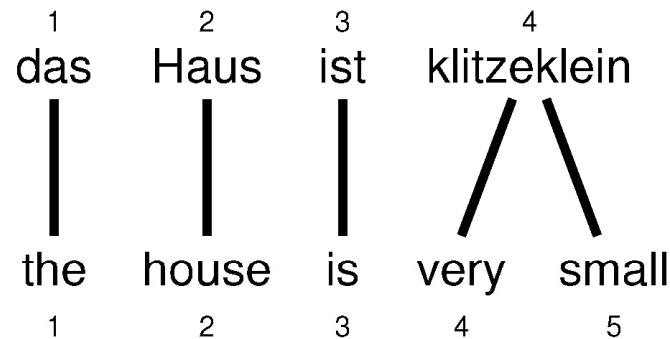
Words may be reordered during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-Many Translation

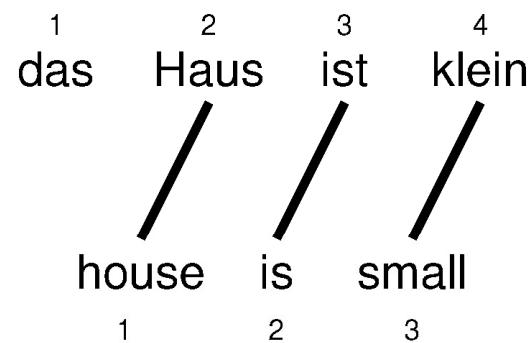
A source word may translate into multiple target words



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

Dropping Words

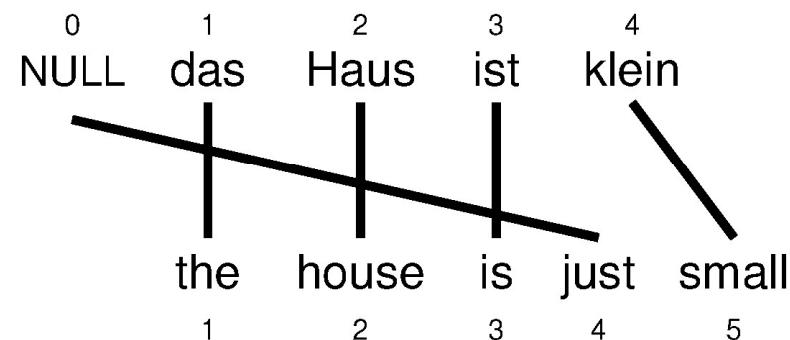
Words may be dropped when translated
(German article **das** is dropped)



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting Words

- Words may be added during translation
 - The English **just** does not have an equivalent in German
 - We still need to map it to something: special NULL token



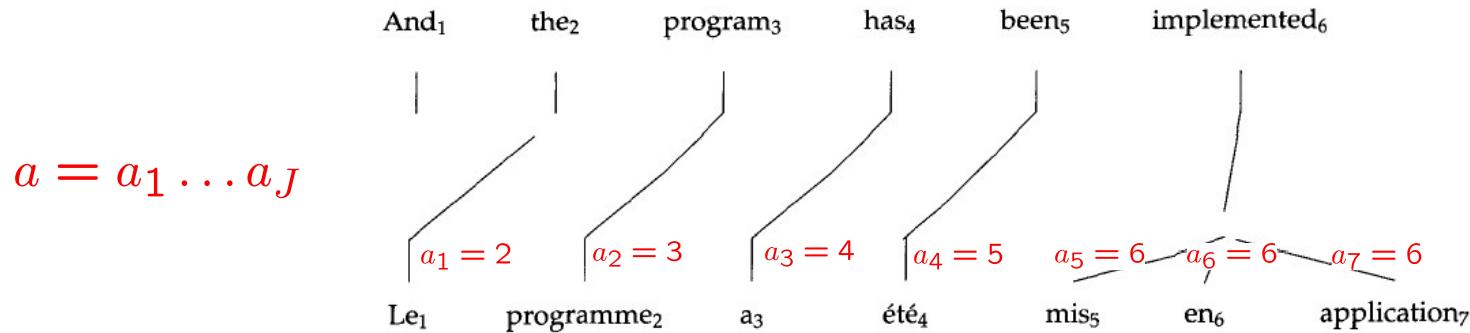
$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

IBM Model 1: Allocation



IBM Model 1 (Brown 93)

- Alignments: a hidden vector called an *alignment* specifies which English source is responsible for each French target word.



$$\begin{aligned} P(f, a | e) &= \prod_j P(a_j = i) P(f_j | e_i) \\ &= \prod_j \frac{1}{I+1} P(f_j | e_i) \end{aligned}$$

$$P(f | e) = \sum_a P(f, a | e)$$

Example

das	
<i>e</i>	<i>t(e f)</i>
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus	
<i>e</i>	<i>t(e f)</i>
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

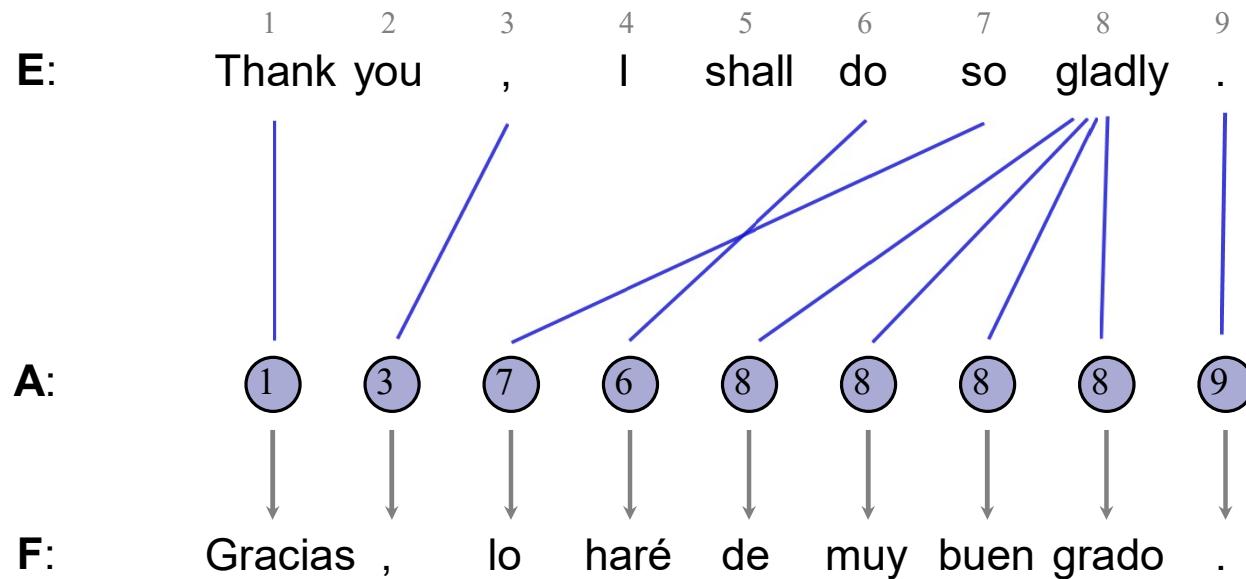
ist	
<i>e</i>	<i>t(e f)</i>
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein	
<i>e</i>	<i>t(e f)</i>
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

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



IBM Models 1/2



Model Parameters

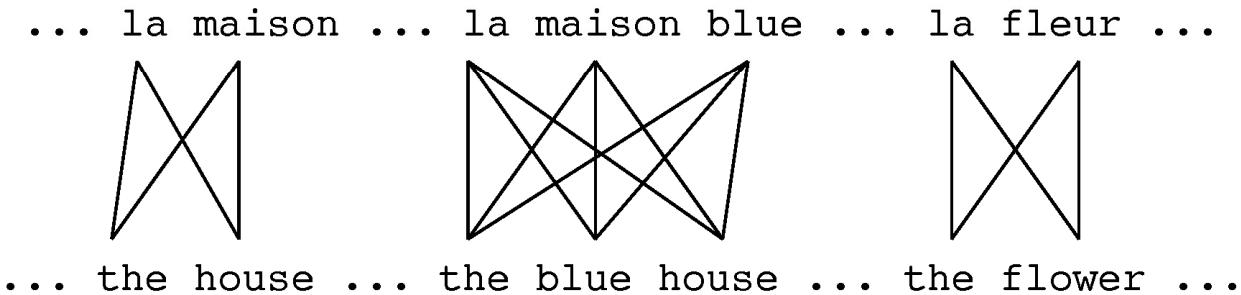
Translation: $P(F_1 = \text{Gracias} | E_{A1} = \text{Thank})$ *Alignment:* $P(A_2 = 3)$

Expectation Maximization

EM Algorithm

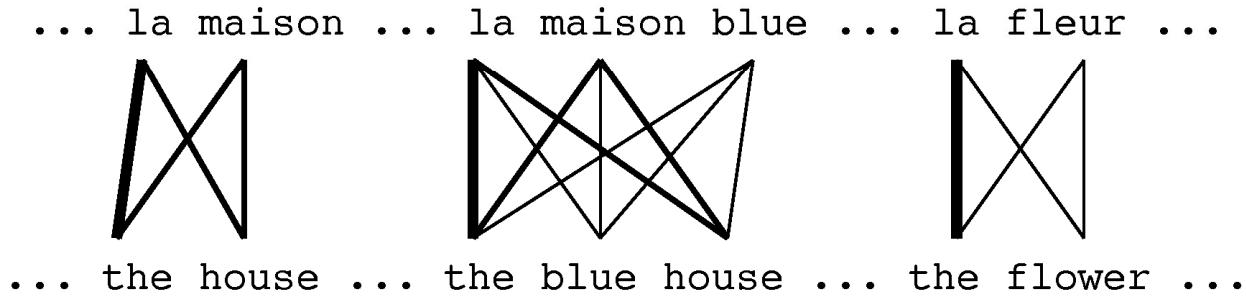
- Incomplete data
 - if we had *complete data*, would could estimate *model*
 - if we had *model*, we could fill in the *gaps in the data*
- Expectation Maximization (EM) in a nutshell
 1. initialize model parameters (e.g. uniform)
 2. assign probabilities to the missing data
 3. estimate model parameters from completed data
 4. iterate steps 2–3 until convergence

EM Algorithm



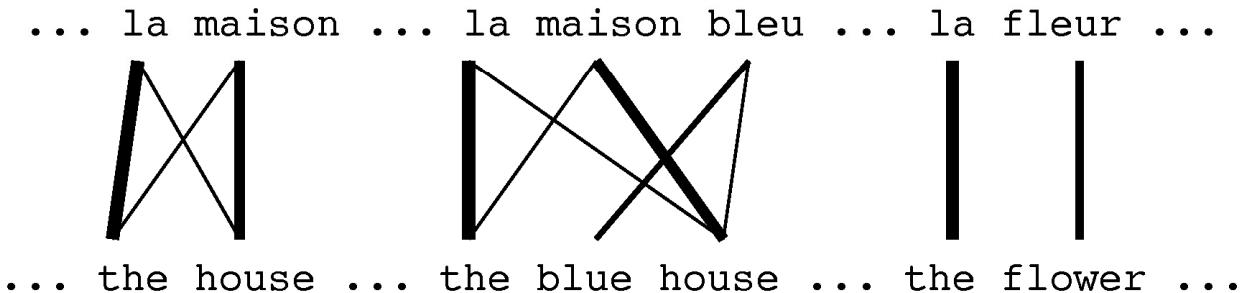
- Initial step: all alignments equally likely
- Model learns that, e.g., **la** is often aligned with **the**

EM Algorithm



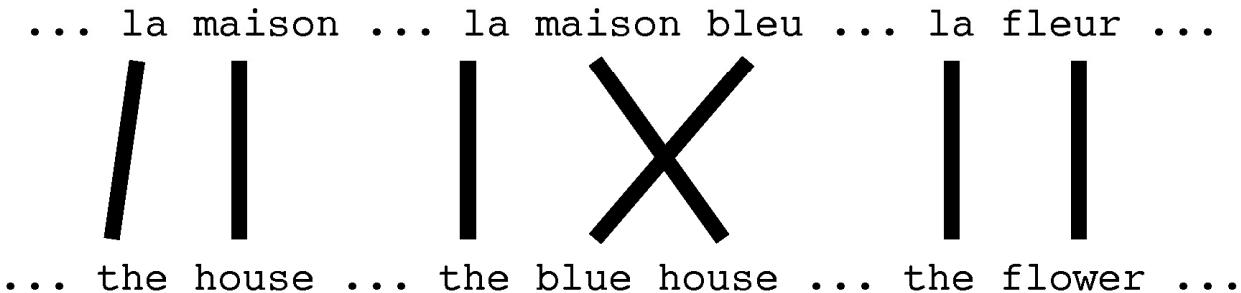
- After one iteration
- Alignments, e.g., between **la** and **the** are more likely

EM Algorithm



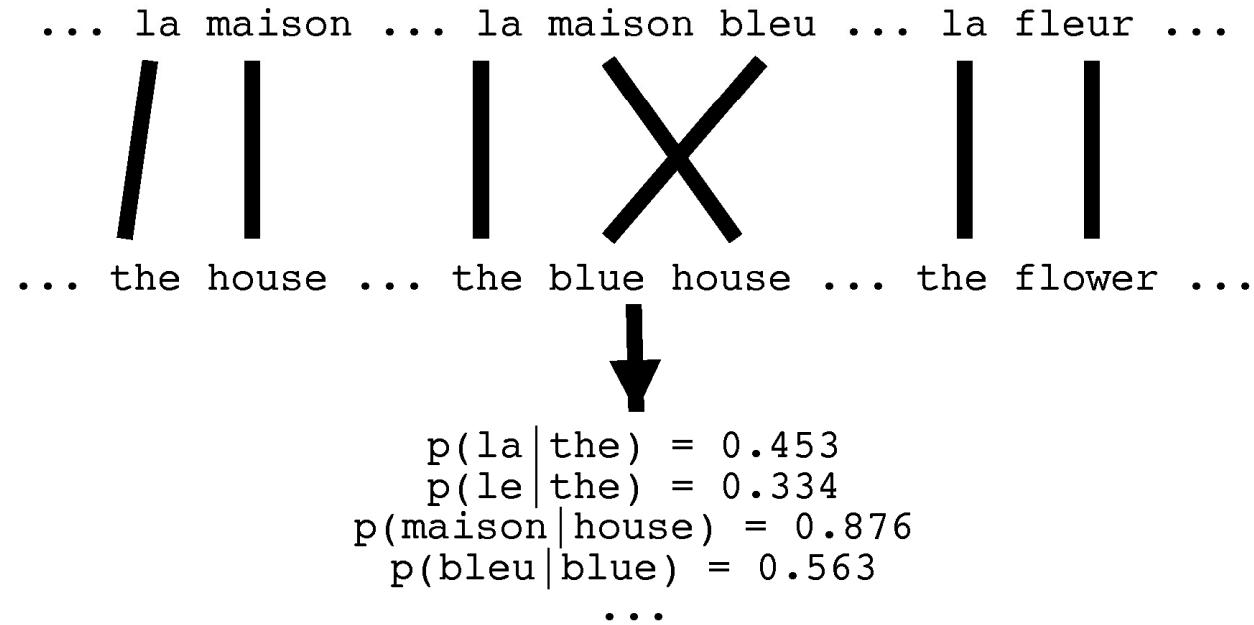
- After another iteration
- It becomes apparent that alignments, e.g., between **fleur** and **flower** are more likely (pigeon hole principle)

EM Algorithm



- Convergence
- Inherent hidden structure revealed by EM

EM Algorithm



- Parameter estimation from the aligned corpus

IBM Model 1 and EM

- EM Algorithm consists of two steps
- Expectation-Step: Apply model to the data
 - parts of the model are hidden (here: alignments)
 - using the model, assign probabilities to possible values
- Maximization-Step: Estimate model from data
 - take assigned values as fact
 - collect counts (weighted by probabilities)
 - estimate model from counts
- Iterate these steps until convergence

IBM Model 1 and EM

- We need to be able to compute:
 - Expectation-Step: probability of alignments
 - Maximization-Step: count collection

IBM Model 1 and EM

- **Probabilities**

$$\begin{array}{ll} p(\text{the}|\text{la}) = 0.7 & p(\text{house}|\text{la}) = 0.05 \\ p(\text{the}|\text{maison}) = 0.1 & p(\text{house}|\text{maison}) = 0.8 \end{array}$$

- **Alignments**



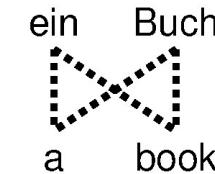
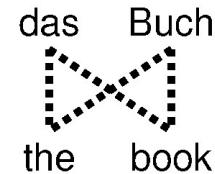
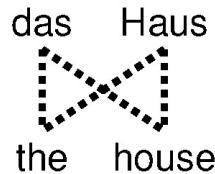
$$p(\mathbf{e}, a | \mathbf{f}) = 0.56 \quad p(\mathbf{e}, a | \mathbf{f}) = 0.035 \quad p(\mathbf{e}, a | \mathbf{f}) = 0.08 \quad p(\mathbf{e}, a | \mathbf{f}) = 0.005$$

$$p(a | \mathbf{e}, \mathbf{f}) = 0.824 \quad p(a | \mathbf{e}, \mathbf{f}) = 0.052 \quad p(a | \mathbf{e}, \mathbf{f}) = 0.118 \quad p(a | \mathbf{e}, \mathbf{f}) = 0.007$$

- **Counts**

$$\begin{array}{ll} c(\text{the}|\text{la}) = 0.824 + 0.052 & c(\text{house}|\text{la}) = 0.052 + 0.007 \\ c(\text{the}|\text{maison}) = 0.118 + 0.007 & c(\text{house}|\text{maison}) = 0.824 + 0.118 \end{array}$$

Convergence



<i>e</i>	<i>f</i>	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

Perplexity

- How well does the model fit the data?
- Perplexity: derived from probability of the training data according to the model

$$\log_2 PP = - \sum_s \log_2 p(\mathbf{e}_s | \mathbf{f}_s)$$

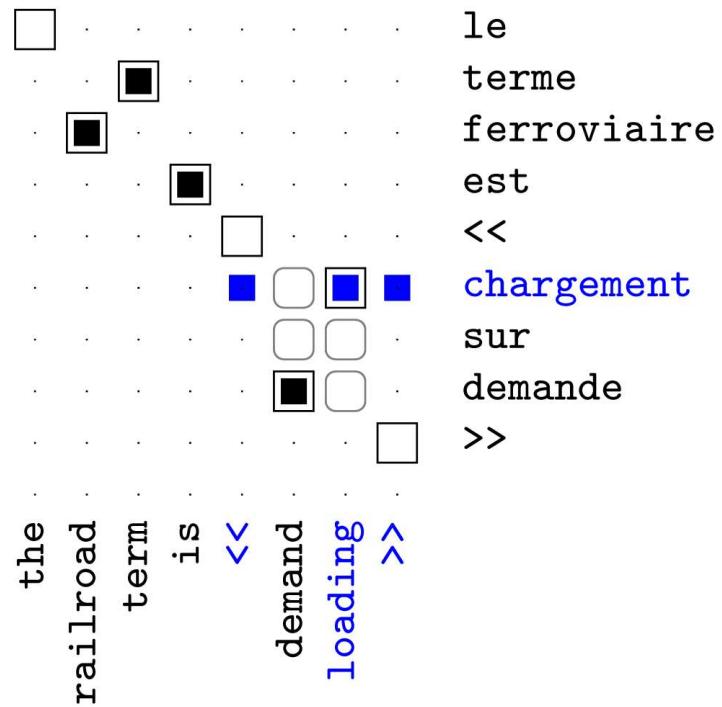
- Example ($\epsilon=1$)

	initial	1st it.	2nd it.	3rd it.	...	final
$p(\text{the haus} \text{das haus})$	0.0625	0.1875	0.1905	0.1913	...	0.1875
$p(\text{the book} \text{das buch})$	0.0625	0.1406	0.1790	0.2075	...	0.25
$p(\text{a book} \text{ein buch})$	0.0625	0.1875	0.1907	0.1913	...	0.1875
perplexity	4095	202.3	153.6	131.6	...	113.8



Problems with Model 1

- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
 - Training data: 1.1M sentences of French-English text, Canadian Hansards
 - Evaluation metric: alignment error Rate (AER)
 - Evaluation data: 447 hand-aligned sentences

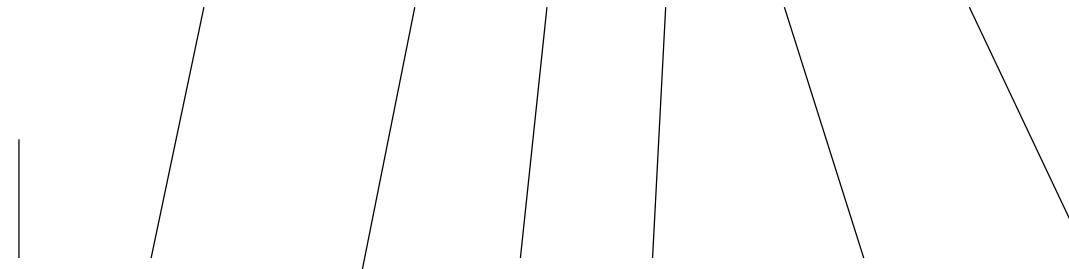


IBM Model 2: Global Monotonicity



Monotonic Translation

Japan shaken by two new quakes

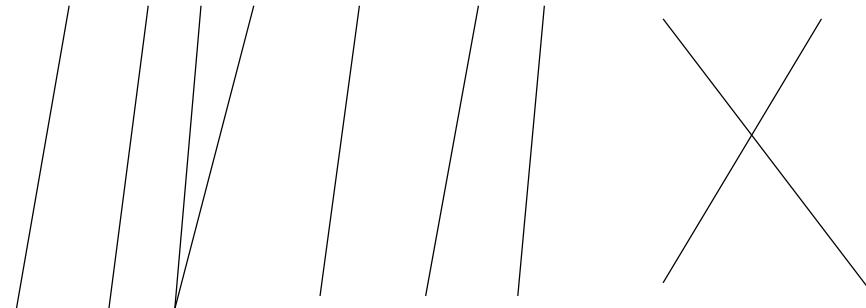


Le Japon secoué par deux nouveaux séismes



Local Order Change

Japan is at the junction of four tectonic plates



Le Japon est au confluent de quatre plaques tectoniques



IBM Model 2

- Alignments tend to the diagonal (broadly at least)

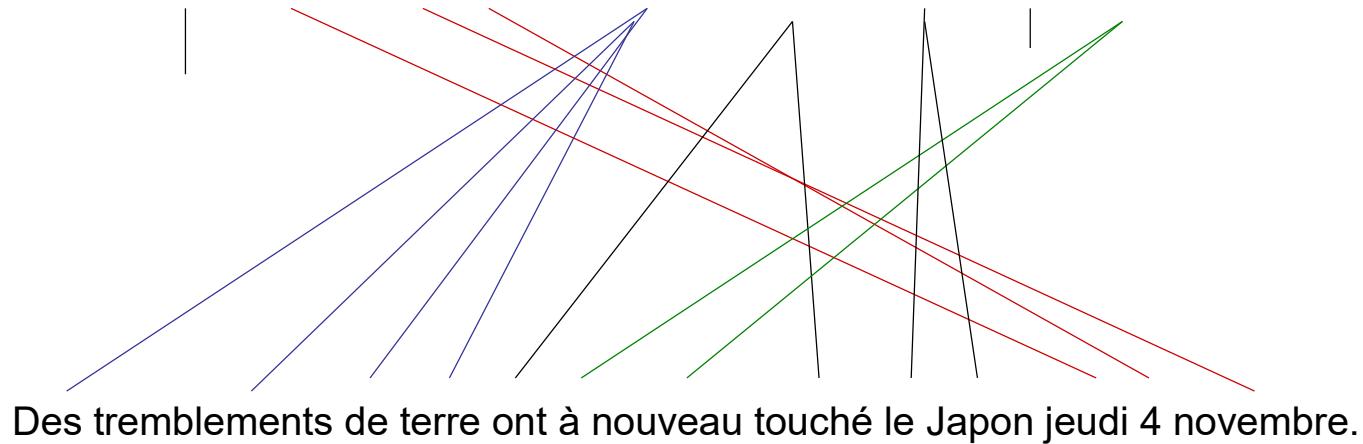
$$\begin{aligned} P(f, a|e) = \prod_j & P(a_j = i | j, I, J) P(f_j | e_i) \\ & P(\text{dist} = i - j \frac{I}{J}) \\ & \frac{1}{Z} e^{-\alpha(i - j \frac{I}{J})} \end{aligned}$$

HMM Model: Local Monotonicity



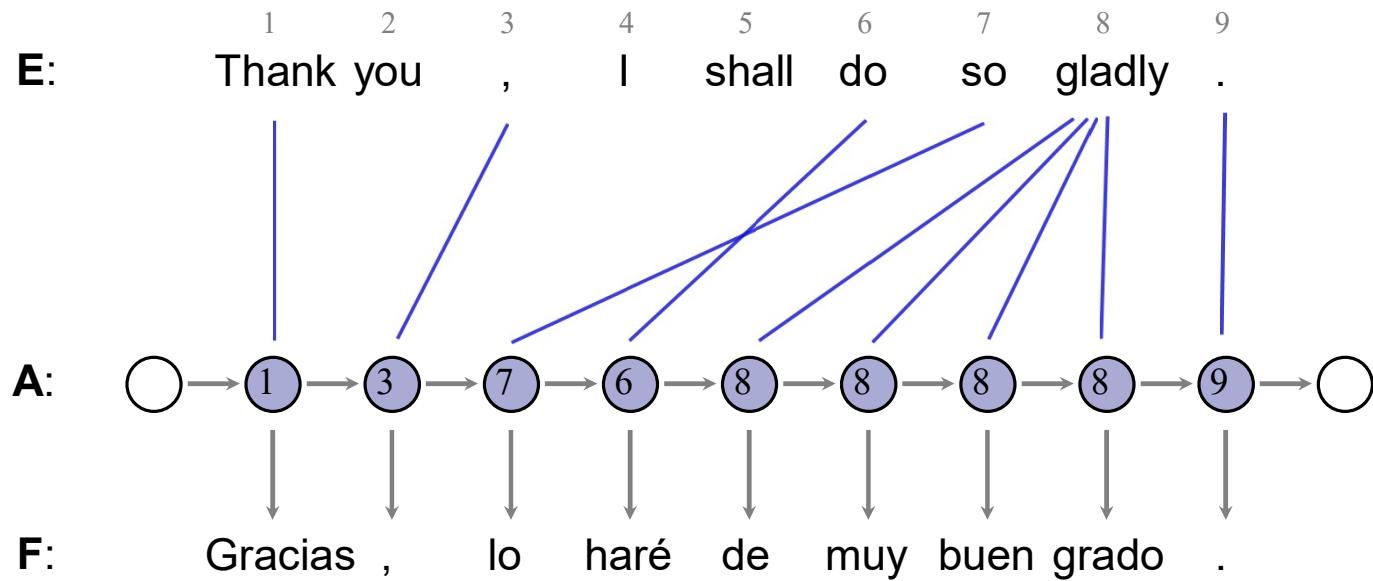
Phrase Movement

On Tuesday Nov. 4, earthquakes rocked Japan once again





The HMM Model



Model Parameters

Emissions: $P(F_1 = \text{Gracias} | E_{A1} = \text{Thank})$ *Transitions:* $P(A_2 = 3 | A_1 = 1)$



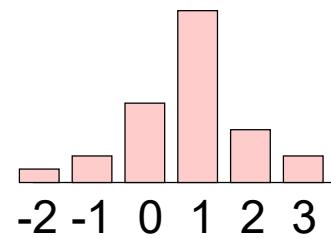
The HMM Model

- Model 2 preferred global monotonicity
- We want local monotonicity:
 - Most jumps are small
- HMM model (Vogel 96)

f	$t(f e)$
nationale	0.469
national	0.418
nationaux	0.054
nationales	0.029

$$P(f, a|e) = \prod_j P(a_j|a_{j-1}) P(f_j|e_i)$$

$P(a_j - a_{j-1})$ 

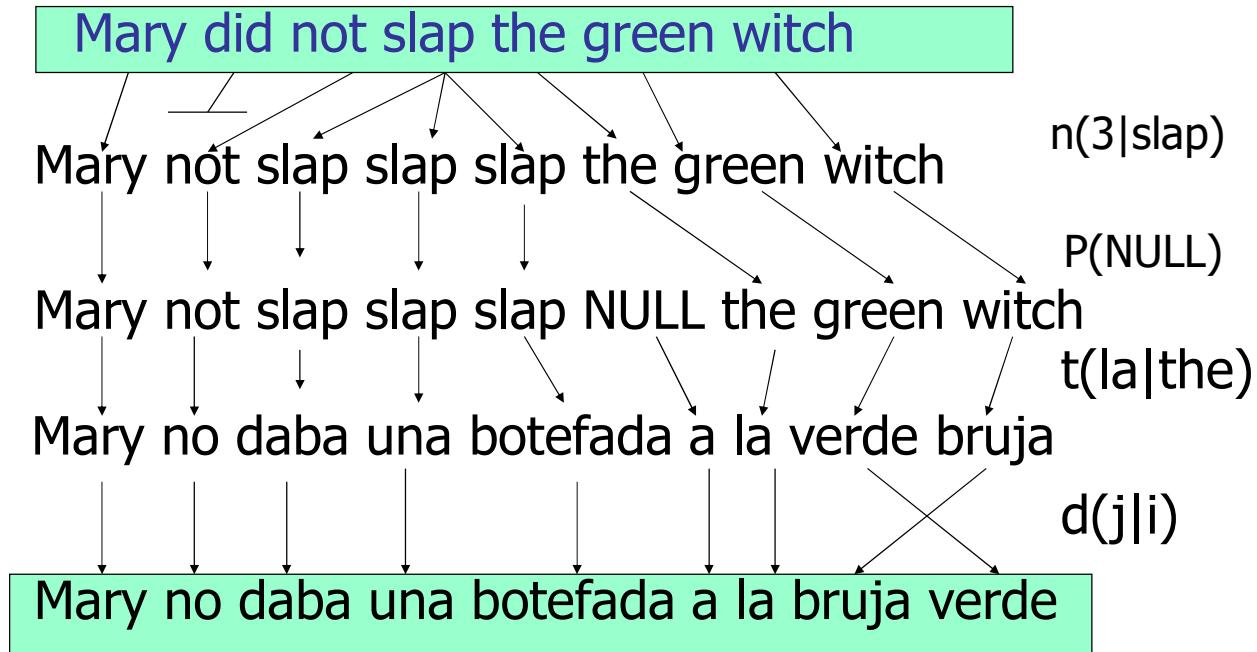


- Re-estimate using the forward-backward algorithm
- Handling nulls requires some care
- What are we still missing?

Models 3+: Fertility



IBM Models 3/4/5



[from Al-Onaizan and Knight, 1998]



Examples: Translation and Fertility

the

f	$t(f e)$	ϕ	$n(\phi e)$
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l'	0.086		
ce	0.018		
cette	0.011		

not

f	$t(f e)$	ϕ	$n(\phi e)$
ne	0.497	2	0.735
pas	0.442	0	0.154
non	0.029	1	0.107
rien	0.011		

farmers

f	$t(f e)$	ϕ	$n(\phi e)$
agriculteurs	0.442	2	0.731
les	0.418	1	0.228
cultivateurs	0.046	0	0.039
producteurs	0.021		



Example: Idioms

nodding

he is nodding
/ ⊥
il hoche la tête

f	$t(f e)$	ϕ	$n(\phi e)$
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		



Example: Morphology

should

f	$t(f e)$	ϕ	$n(\phi e)$
devrait	0.330	1	0.649
devraient	0.123	0	0.336
devrions	0.109	2	0.014
faudrait	0.073		
faut	0.058		
doit	0.058		
aurait	0.041		
doivent	0.024		
devons	0.017		
devrais	0.013		

Word Alignment with IBM Models

- IBM Models create a **many-to-one** mapping
 - words are aligned using an alignment function
 - a function may return the same value for different input
(one-to-many mapping)
 - a function can not return multiple values for one input
(no many-to-one mapping)
- Real word alignments have **many-to-many** mappings

Symmetrization

- Run IBM Model training in both directions
 - two sets of word alignment points
- Intersection: high precision alignment points
- Union: high recall alignment points
- Refinement methods explore the sets between intersection and union

Example

english to spanish

		bofetada		bruja	
Mary			a	la	verde
did					
not					
slap					
the					
green					
witch					

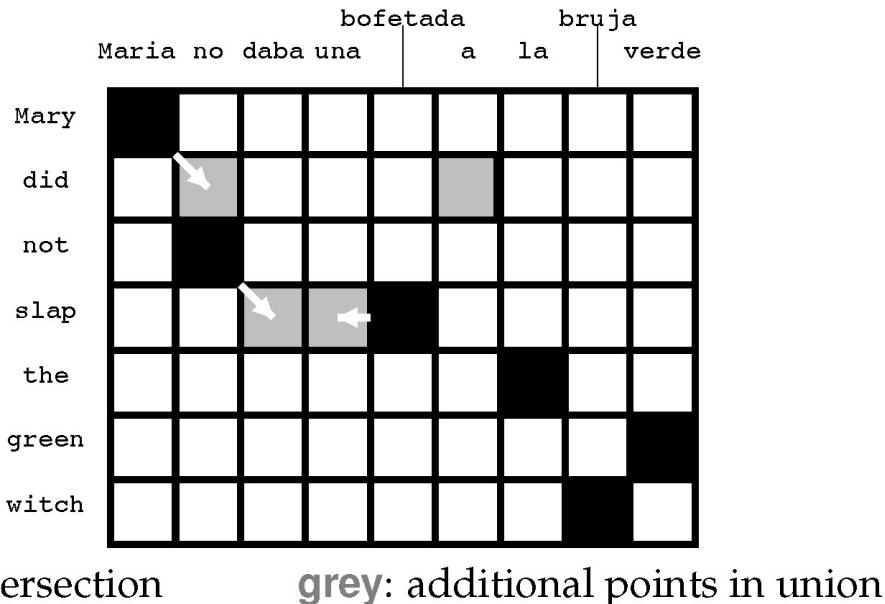
spanish to english

		bofetada		bruja	
Mary			a	la	verde
did					
not					
slap					
the					
green					
witch					

intersection

		bofetada		bruja	
Mary			a	la	verde
did					
not					
slap					
the					
green					
witch					

Growing Heuristics



- Add alignment points from union based on heuristics:
 - directly/diagonally neighboring points
 - finally, add alignments that connect unaligned words in source and/or target
- Popular method: grow-diag-final-and

[Back to Word Alignment](#)