Algorithms for NLP

Machine Translation

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley

Machine Translation

Machine Translation: Examples

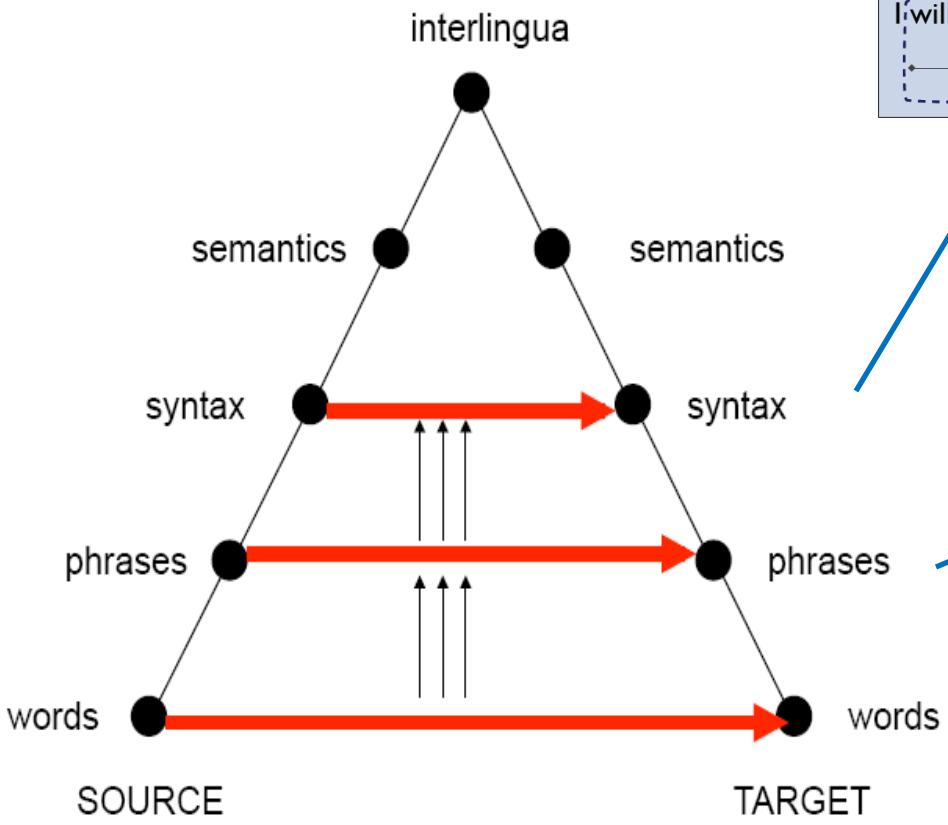
Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

Levels of Transfer



Yo lo haré mañana
I will do it tomorrow
NP
VP

P(VP: MD(will), VP: VB(do), PRN(it), NP | VP: lo haré NP) = 0.8

Yo lo haré mañana
I will do it tomorrow

English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2

Yo lo haré mañana
I will do it tomorrow

English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

Word-Level MT: Examples

la politique de la haine .	*(Foreign Original)*
politics of hate .	(Reference Translation)
the policy of the hatred .	(IBM4+N-grams+Stack)
nous avons signé le protocole .	*(Foreign Original)*
we did sign the memorandum of agreement .	(Reference Translation)
we have signed the protocol .	(IBM4+N-grams+Stack)
où était le plan solide ?	*(Foreign Original)*
but where was the solid plan ?	(Reference Translation)
where was the economic base ?	(IBM4+N-grams+Stack)

Phrasal MT: Examples

Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

Interrogé sur la chaîne CBS dimanche, le président a toutefois clairement précisé que le gouvernement ne preterait pas d'argent sans de fortes contreparties. *"Il faudra faire des sacrifices à tous les niveaux"*, a-t-il prévenu. *"Tout le monde devra se réunir autour de la table et se mettre d'accord sur une restructuration en profondeur"*.

General Motors et Chrysler sont engagés dans des négociations avec le principal syndicat de l'automobile. Les constructeurs souhaitent diminuer leurs cotisations aux caisses de retraites, et accorder en échange des actions aux syndicats. Ils souhaiteraient également négocier des baisses des salaires.

U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February in a restructuring plan based on a total of 22 billion dollars in additional aid .

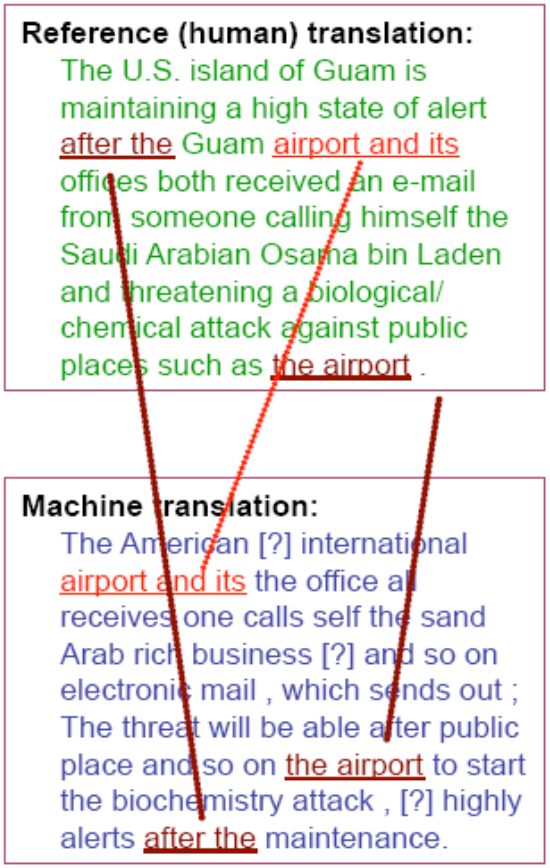
Interviewed on CBS Sunday, the president has clearly stated that the government does not lend money without strong counterparts. *"We must make sacrifices at all levels,"* he warned. *"Everyone should gather around the table and agree on a profound restructuring. "*

General Motors and Chrysler are engaged in negotiations with the major union of the car. Manufacturers wishing to reduce their contributions to pension funds, and give in exchange for the shares to trade unions. They would also negotiate lower wages.

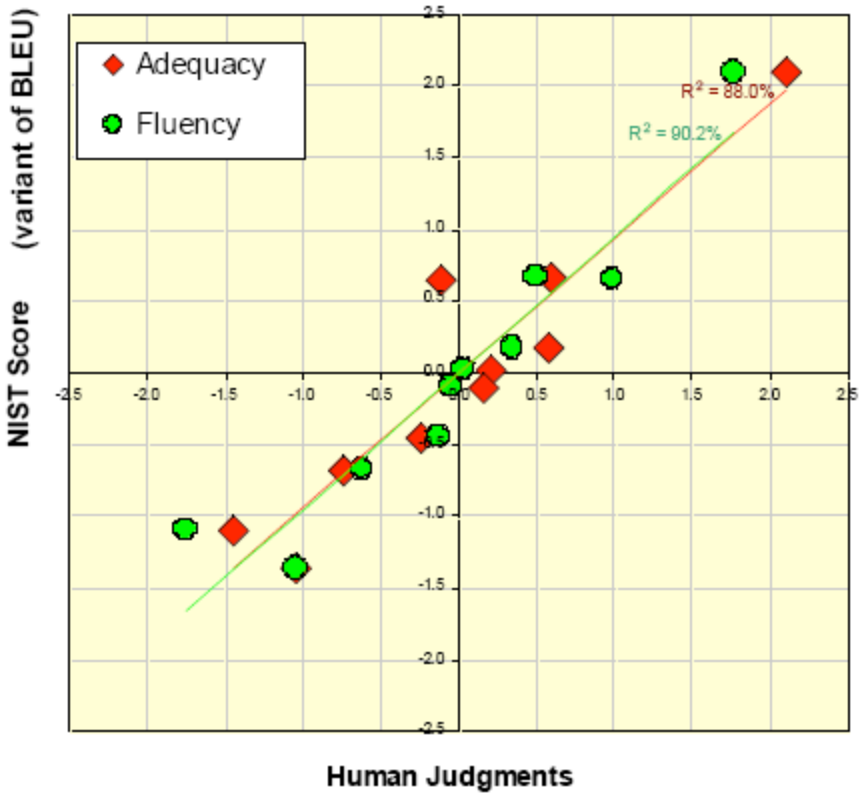
Metrics

MT: Evaluation

- Human evaluations: subject measures, fluency/adequacy
- Automatic measures: n-gram match to references
  - NIST measure: n-gram recall (worked poorly)
  - BLEU: n-gram precision (no one really likes it, but everyone uses it)
  - Lots more: TER, HTER, METEOR, …
- BLEU:
  - P1 = unigram precision
  - P2, P3, P4 = bi-, tri-, 4-gram precision
  - Weighted geometric mean of P1-4
  - Brevity penalty (why?)
  - Somewhat hard to game…
  - Magnitude only meaningful on same language, corpus, number of references, probably only within system types…



Automatic Metrics Work (?)



slide from G. Doddington (NIST)

Systems Overview

Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

Yo lo haré mañana I will do it tomorrow	Hasta pronto See you soon	Hasta pronto See you around

Machine translation system:

Yo lo haré pronto

Novel Sentence

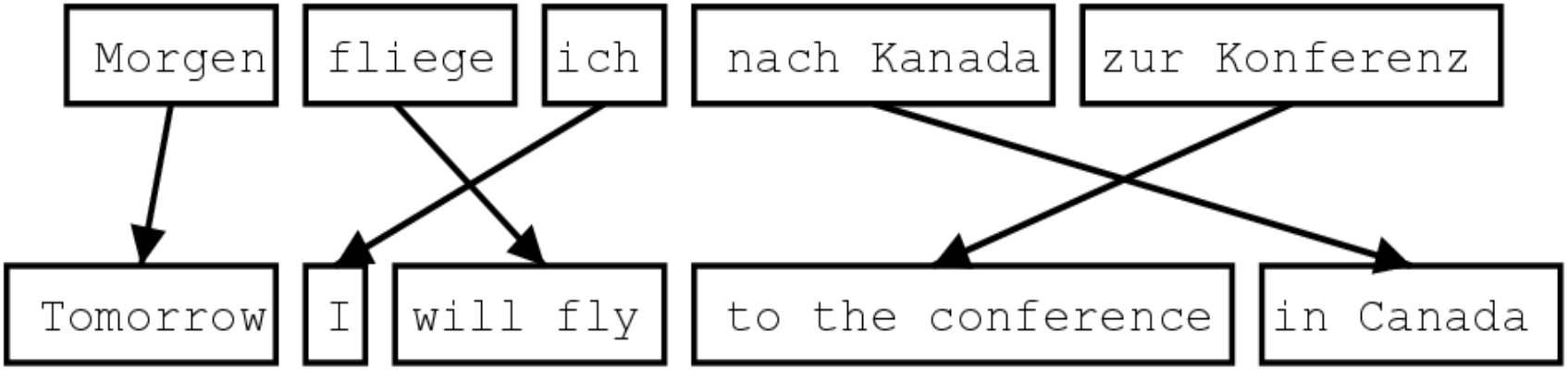
Model of translation

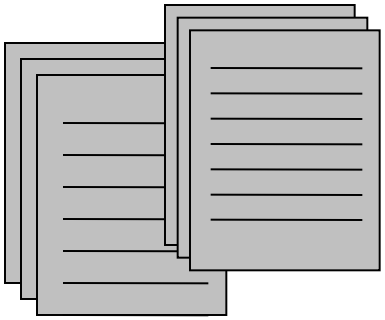
I will do it soon

I will do it around

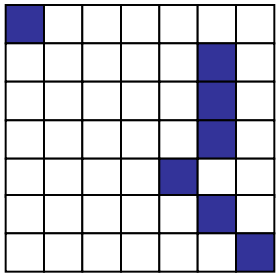
See you tomorrow

Phrase-Based System Overview





Sentence-aligned corpus



Word alignments

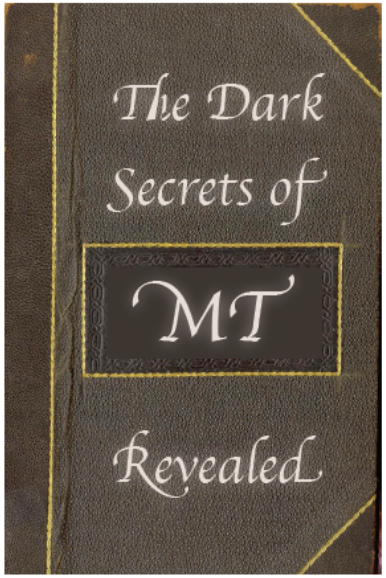
cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
…

Phrase table (translation model)

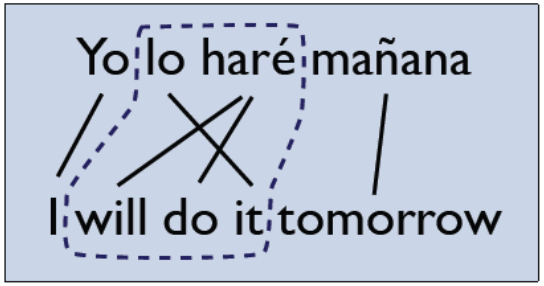
Many slides and examples from Philipp Koehn or John DeNero

Word Alignment

Word Alignment



1. *Align words with a probabilistic model*
2. *Infer presence of larger structures from this alignment*
3. *Translate with the larger structures*

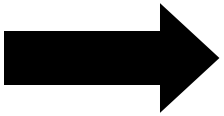


Word Alignment

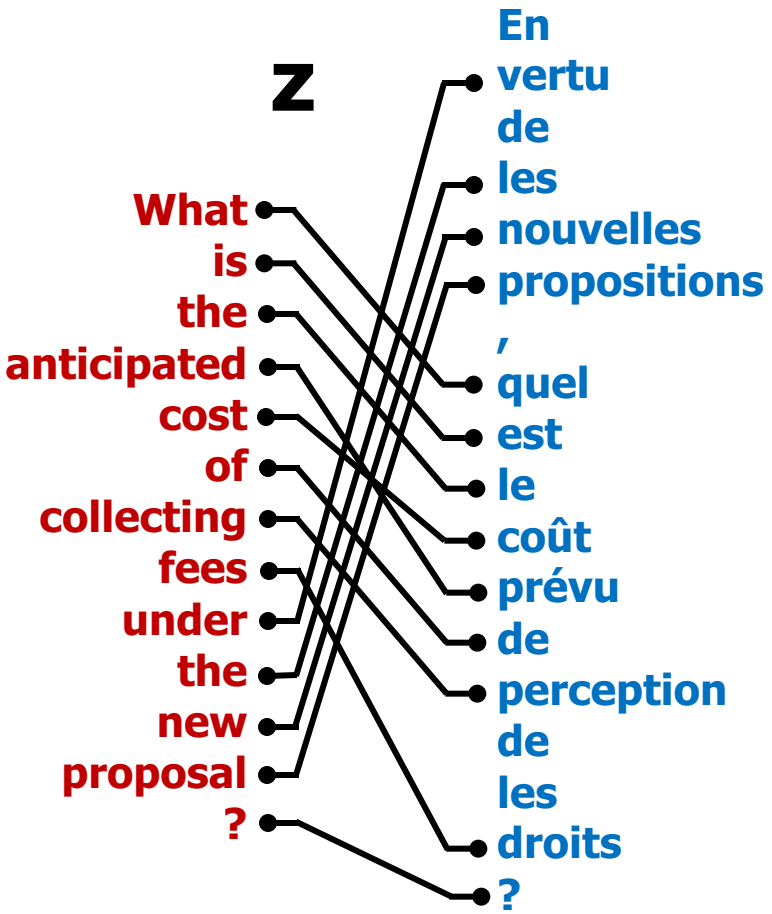
x

What is the anticipated cost of collecting fees under the new proposal?

En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?



z

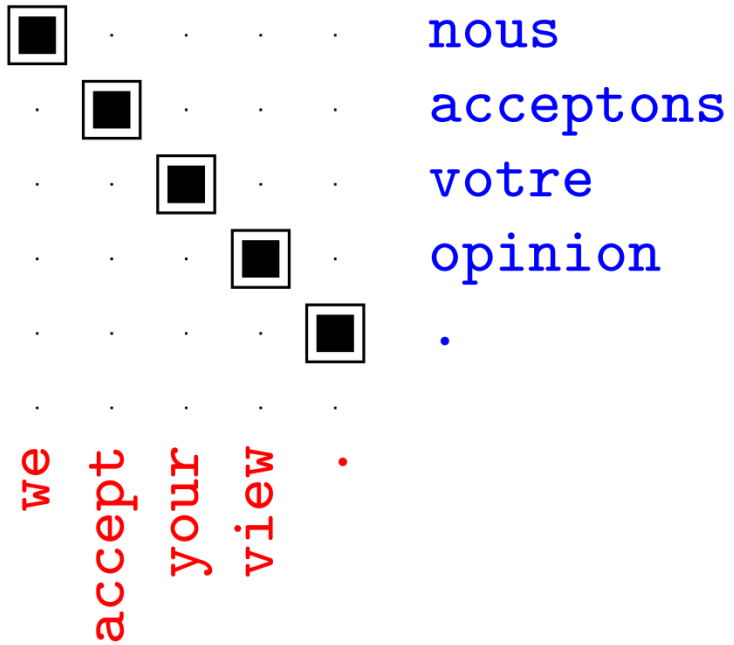


Unsupervised Word Alignment

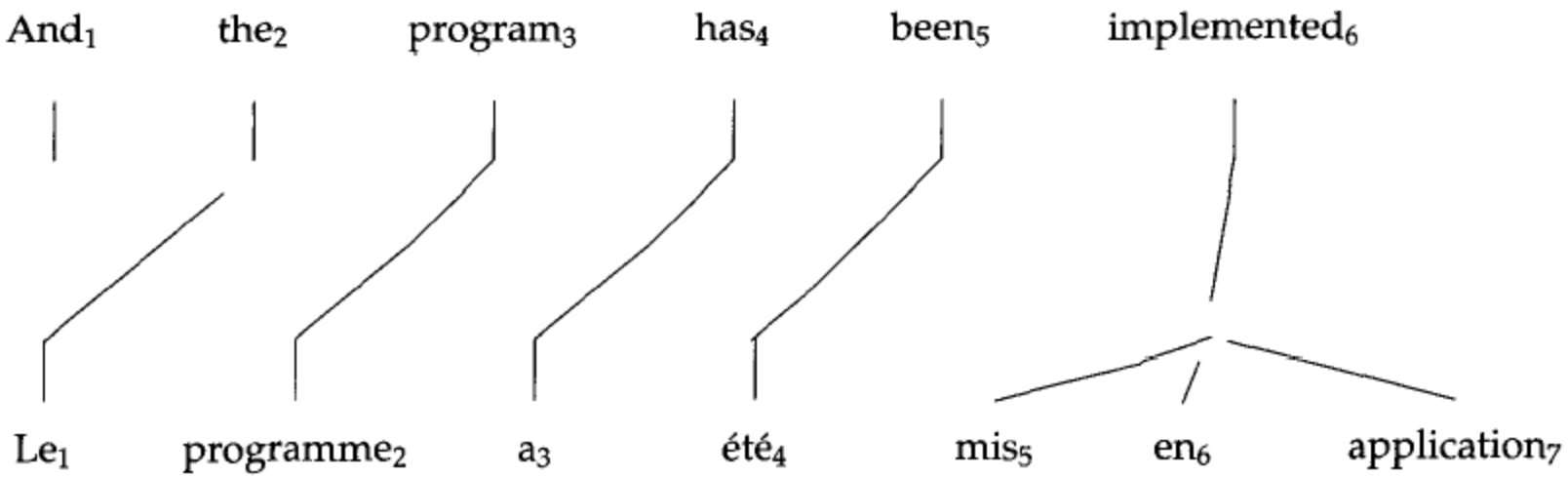
- Input: a ***bitext***: pairs of translated sentences

> **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



1-to-Many Alignments



Evaluating Models

- How do we measure quality of a word-to-word model?
  - Method 1: use in an end-to-end translation system
    - Hard to measure translation quality
    - Option: human judges
    - Option: reference translations (NIST, BLEU)
    - Option: combinations (HTER)
    - Actually, no one uses word-to-word models alone as TMs
  - Method 2: measure quality of the alignments produced
    - Easy to measure
    - Hard to know what the gold alignments should be
    - Often does not correlate well with translation quality (like perplexity in LMs)

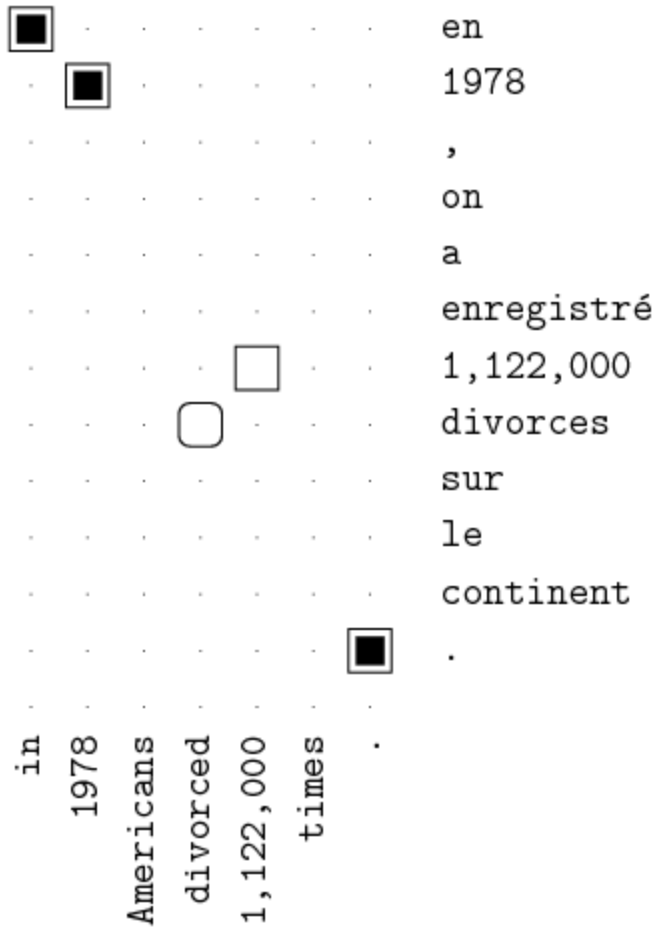
Alignment Error Rate

- Alignment Error Rate

= Sure

= Possible

= Predicted



$$AER(A,S,P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

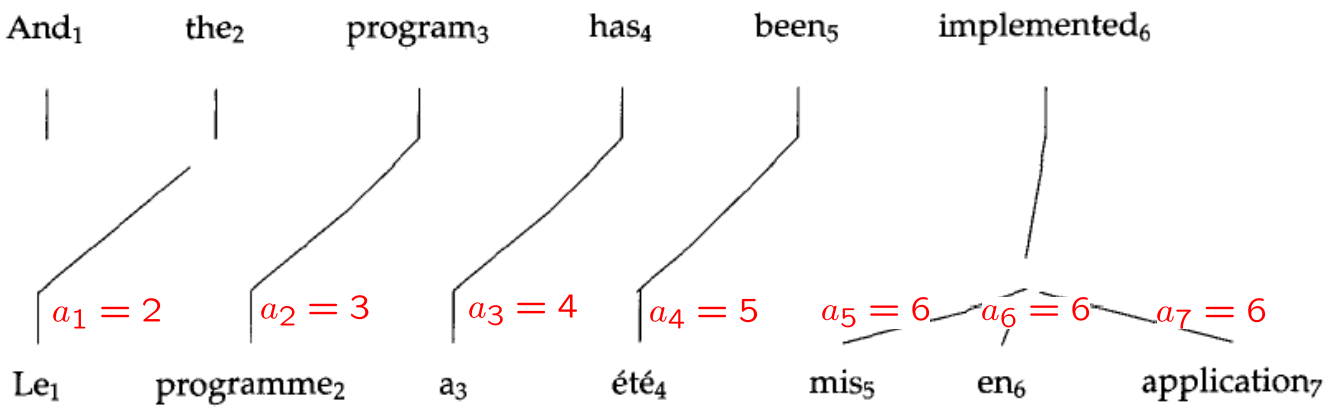
$$= \left(1 - \frac{3+3}{3+4}\right) = \frac{1}{7}$$

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.

$a = a_1 \ldots a_J$

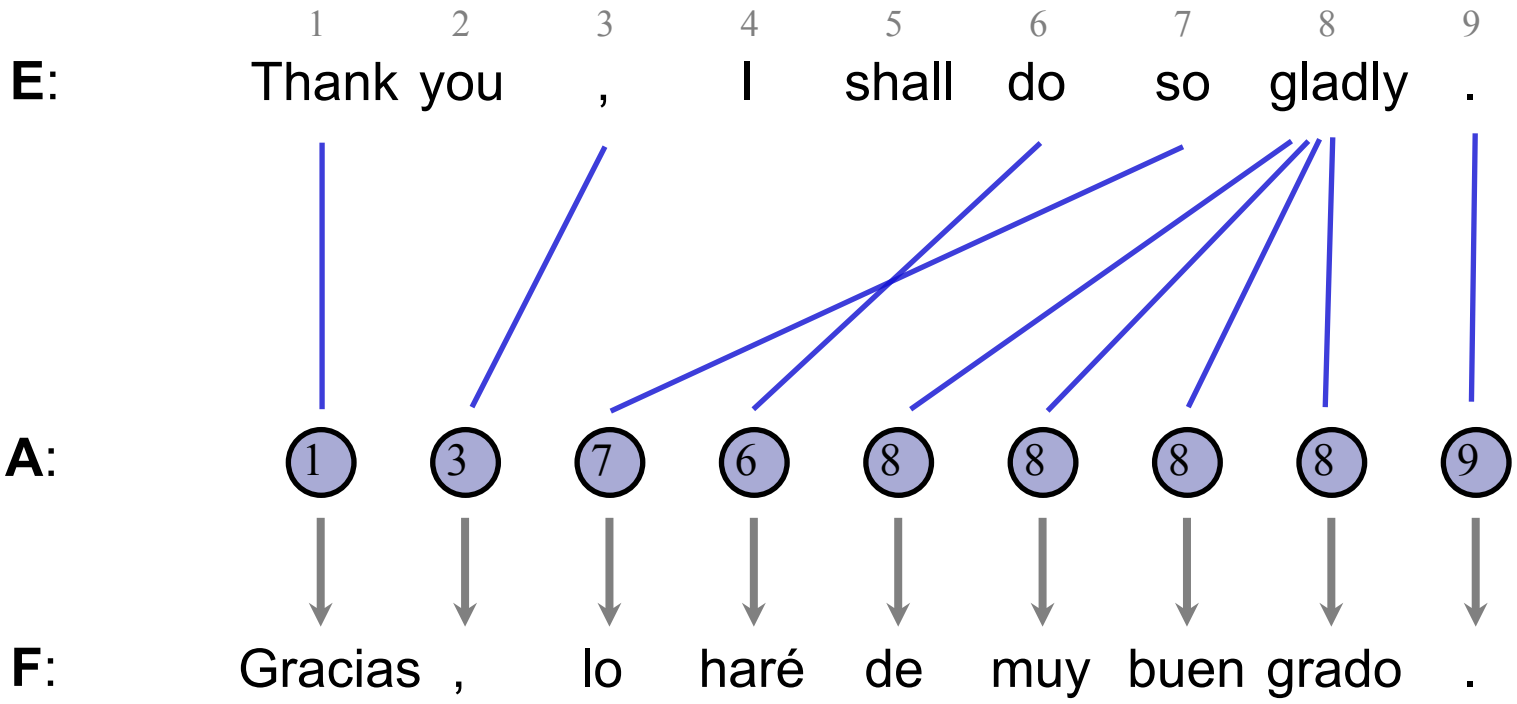


$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$

$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

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

IBM Models 1/2

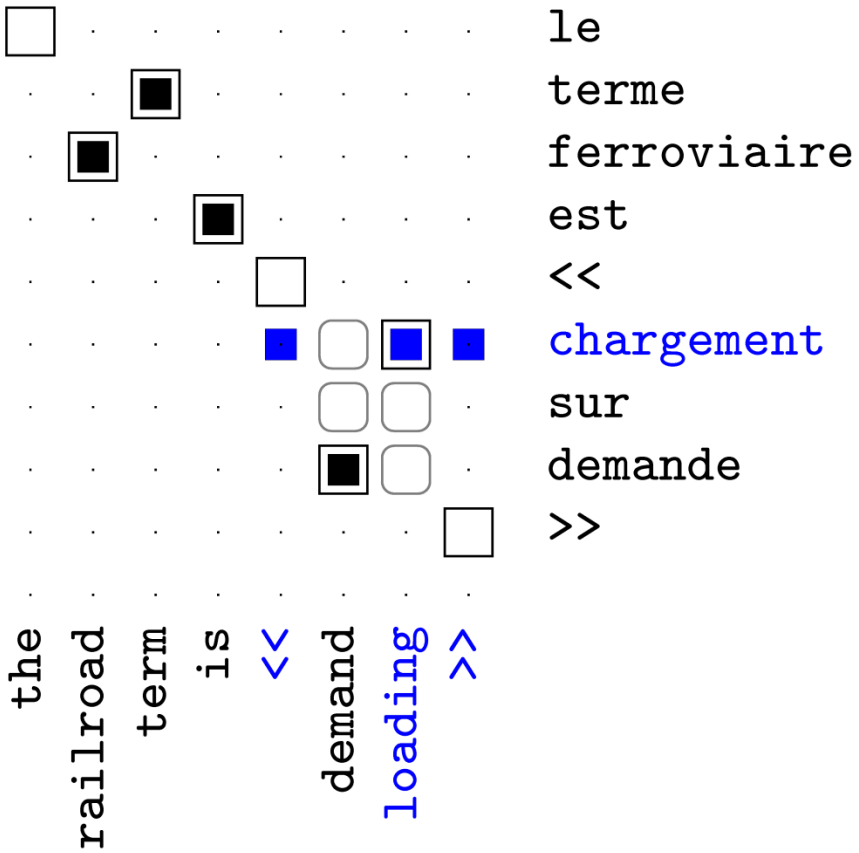


Model Parameters

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

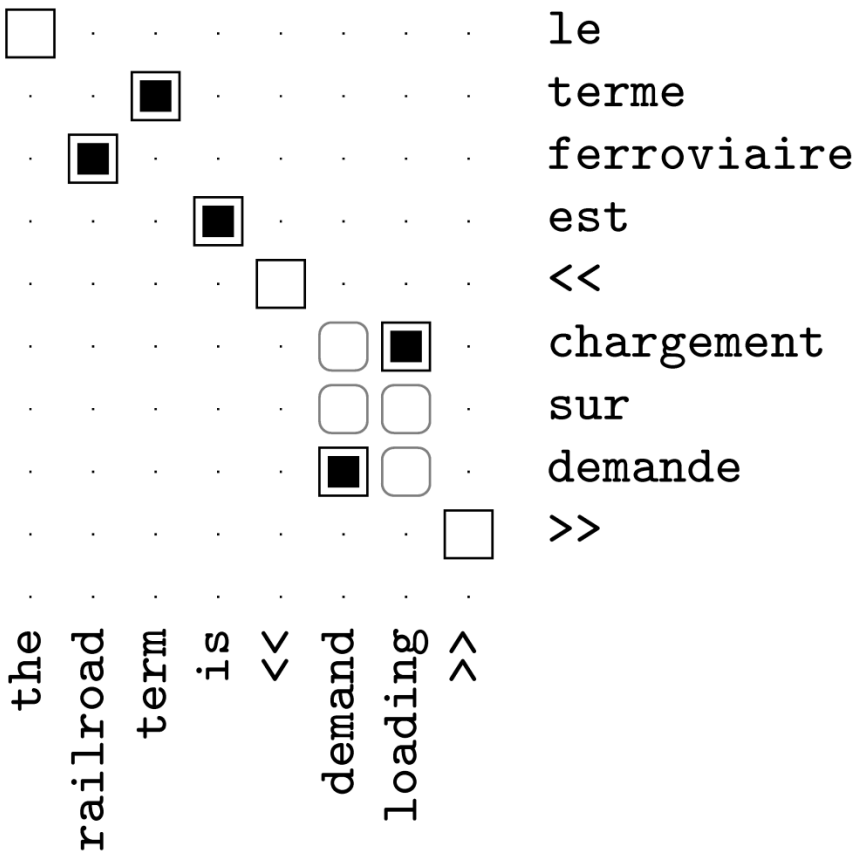
Problems with Model 1

- 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



Intersected Model 1

- Post-intersection: standard practice to train models in each direction then intersect their predictions [Och and Ney, 03]
- Second model is basically a filter on the first
  - Precision jumps, recall drops
  - End up not guessing hard alignments



Model	P/R	AER
Model 1 E→F	82/58	30.6
Model 1 F→E	85/58	28.7
Model 1 AND	96/46	34.8

Joint Training?

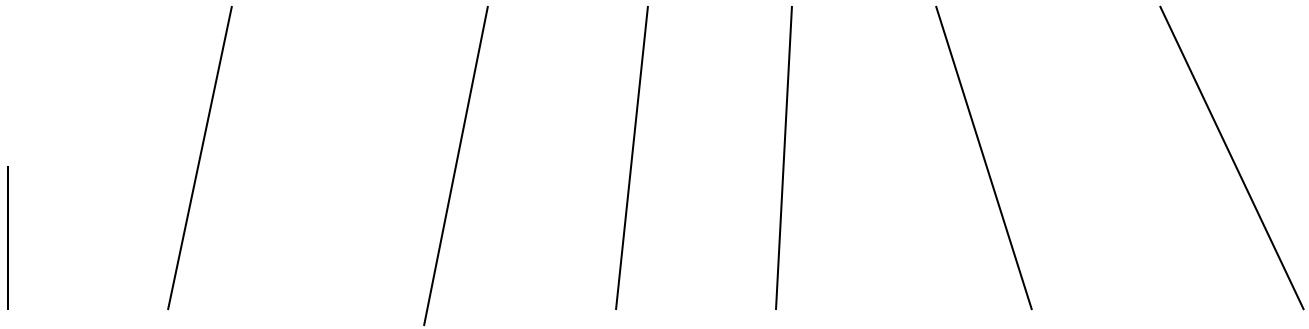
- Overall:
  - Similar high precision to post-intersection
  - But recall is much higher
  - More confident about positing non-null alignments

Model	P/R	AER
Model 1 E→F	82/58	30.6
Model 1 F→E	85/58	28.7
Model 1 AND	96/46	34.8
Model 1 INT	93/69	19.5

IBM Model 2: Global Monotonicity

Monotonic Translation

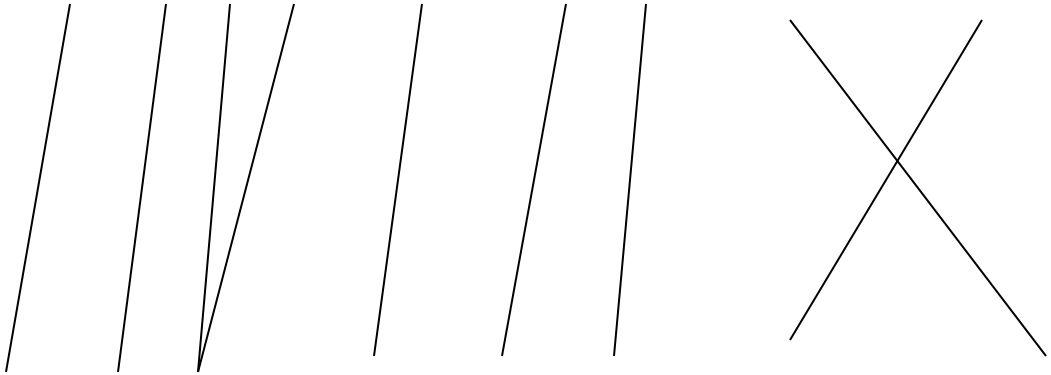
Japan shaken by two new quakes



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)

$$P(f, a|e) = \prod_j P(a_j = i|j, I, J) P(f_j|e_i)$$

$$P(dist = i - j\frac{I}{J})$$

$$\frac{1}{Z} e^{-\alpha(i - j\frac{I}{J})}$$

- Other schemes for biasing alignments towards the diagonal:
  - Relative vs absolute alignment
  - Asymmetric distances
  - Learning a full multinomial over distances

EM for Models 1/2

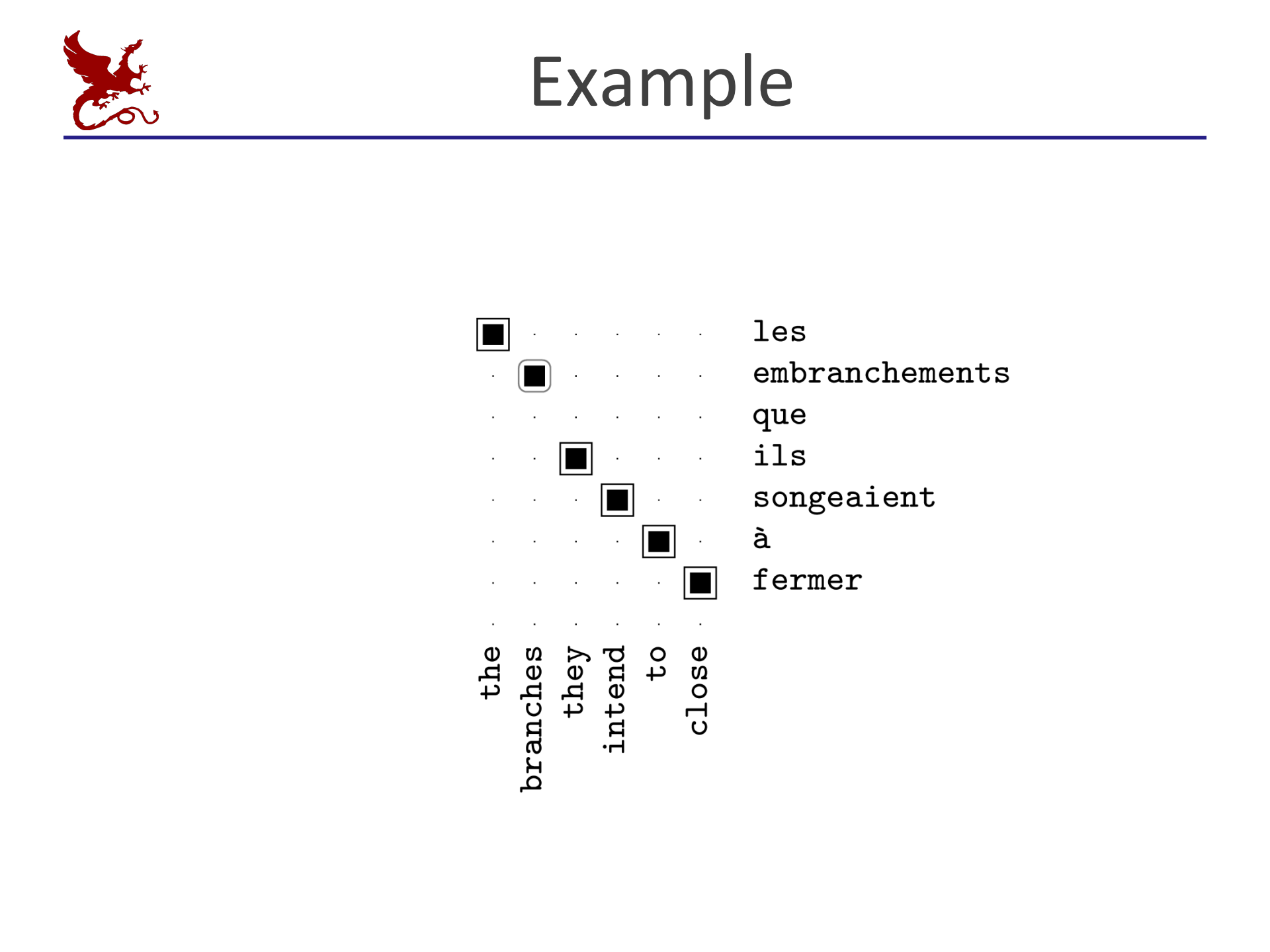
- Model 1 Parameters:
  Translation probabilities (1+2)
  Distortion parameters (2 only)

$$P(f_j|e_i)$$
$$P(a_j = i|j, I, J)$$

- Start with $P(f_j|e_i)$ uniform, including $P(f_j|null)$
- For each sentence:
  - For each French position j
    - Calculate posterior over English positions

$$P(a_j = i|f, e) = \frac{P(a_j = i|j, I, J)P(f_j|e_i)}{\sum_{i'} P(a_j = i'|j, I, J)P(f_j|e'_i)}$$

    - (or just use best single alignment)
    - Increment count of word f_j with word e_i by these amounts
    - Also re-estimate distortion probabilities for model 2
- Iterate until convergence



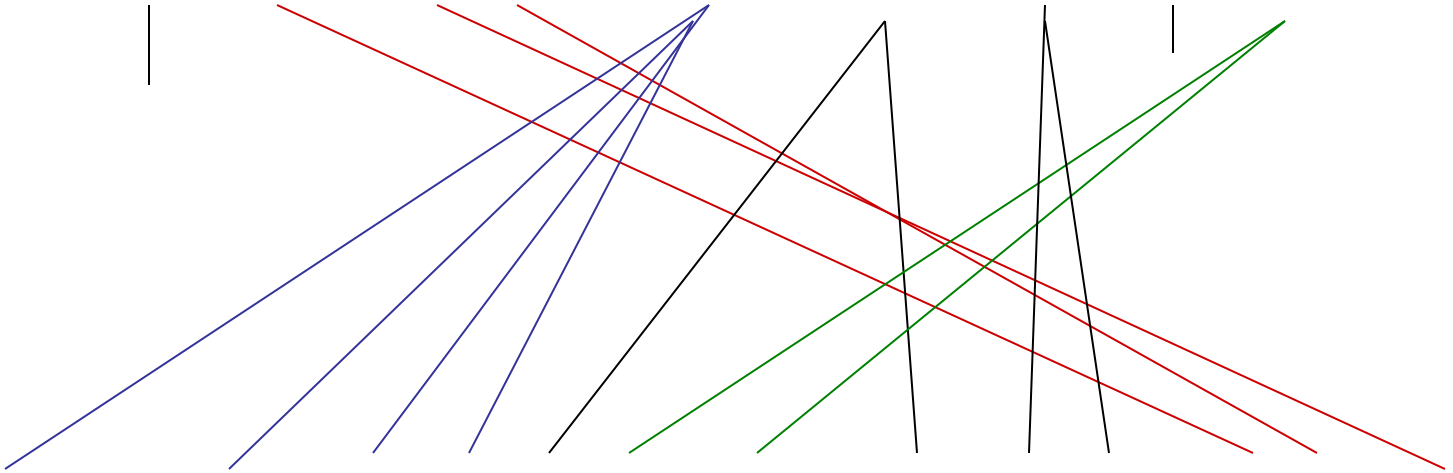
Example

les
embranchements
que
ils
songeaient
à
fermer
the
branches
they
intend
to
close

HMM Model: Local Monotonicity

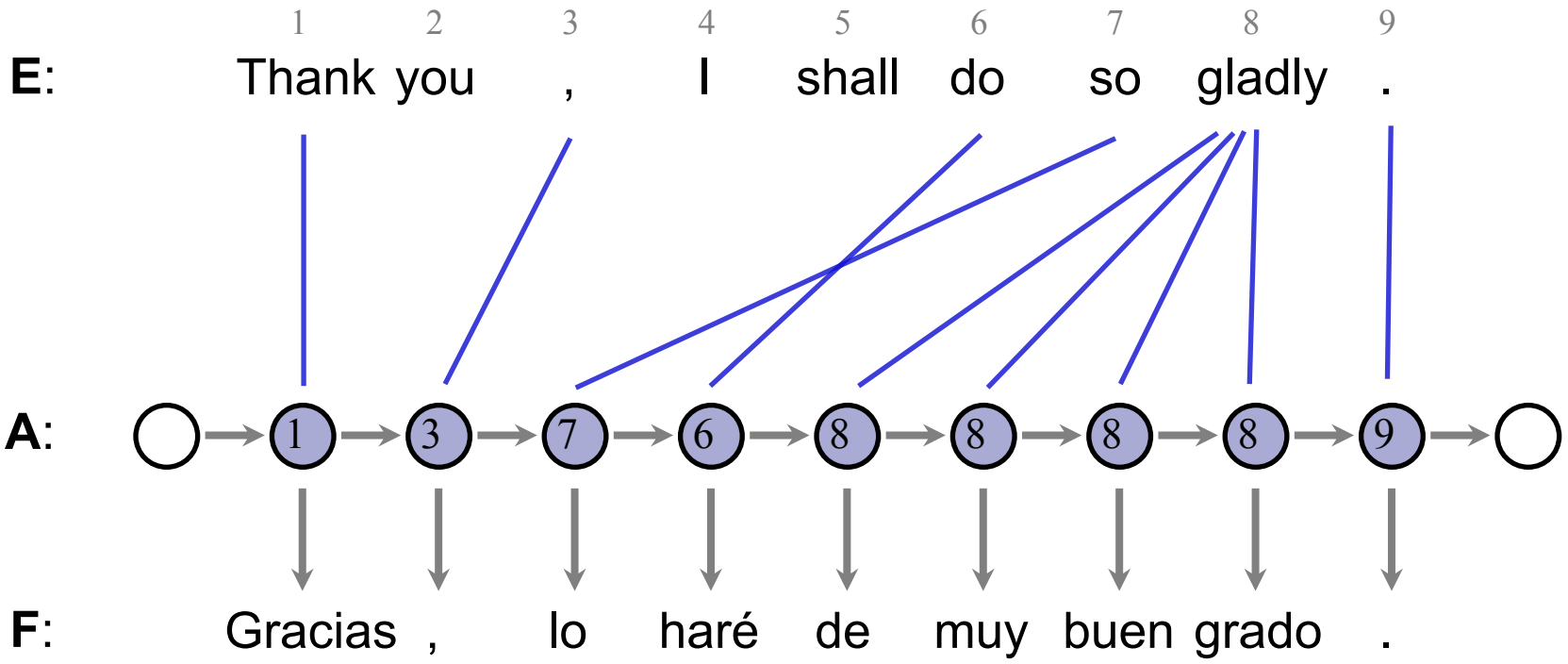
Phrase Movement

On Tuesday Nov. 4, earthquakes rocked Japan once again



Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.

The HMM Model



Model Parameters

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

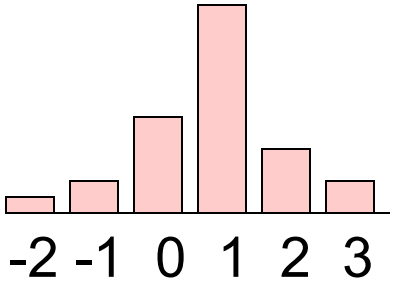
The HMM Model

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

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

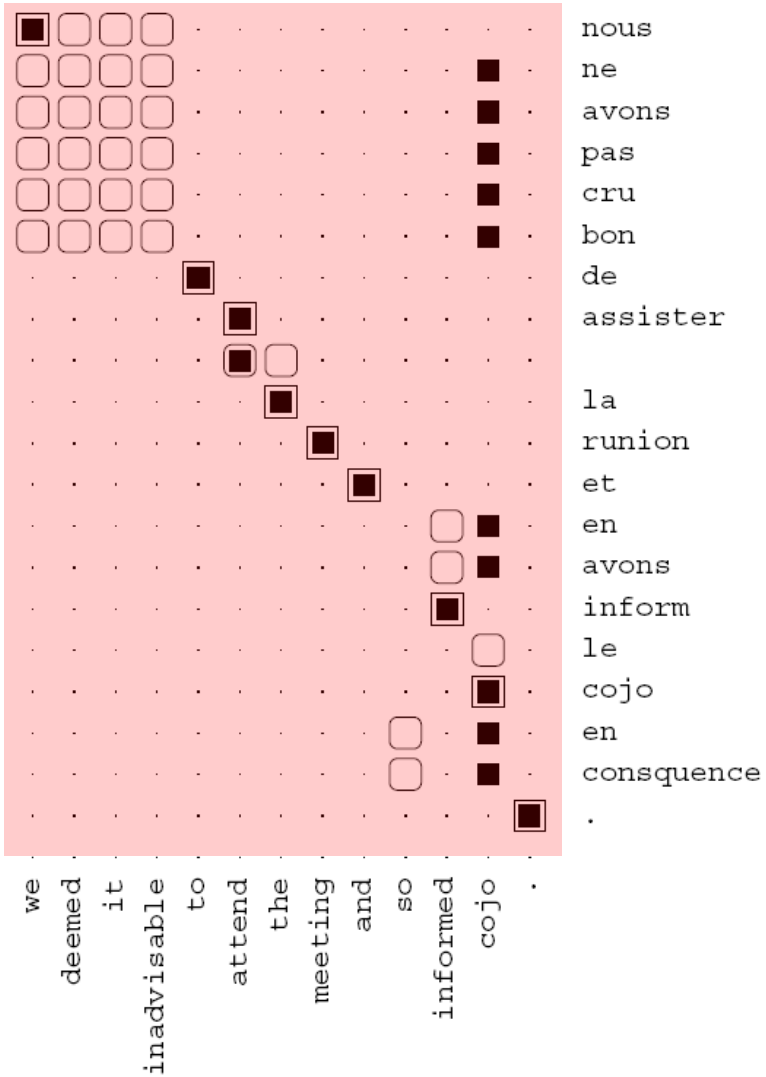
$$P(f, a|e) = \prod_j P(a_j|a_{j-1})P(f_j|e_i)$$

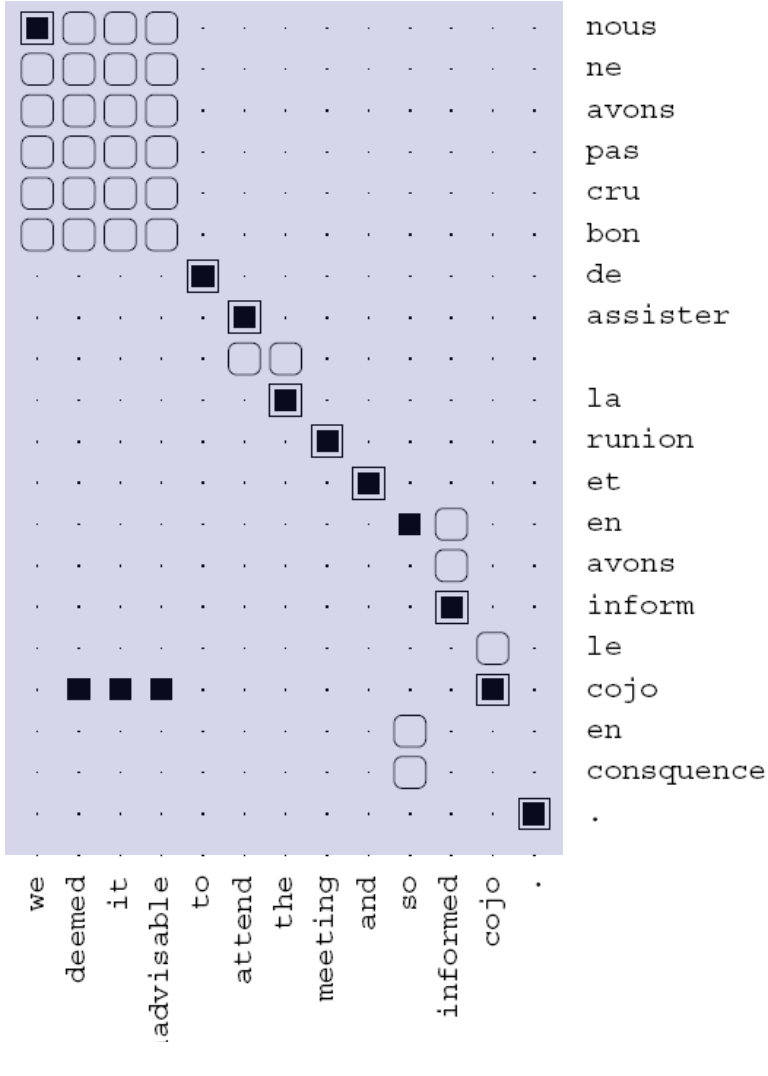
$$P(a_j - a_{j-1})$$



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

HMM Examples



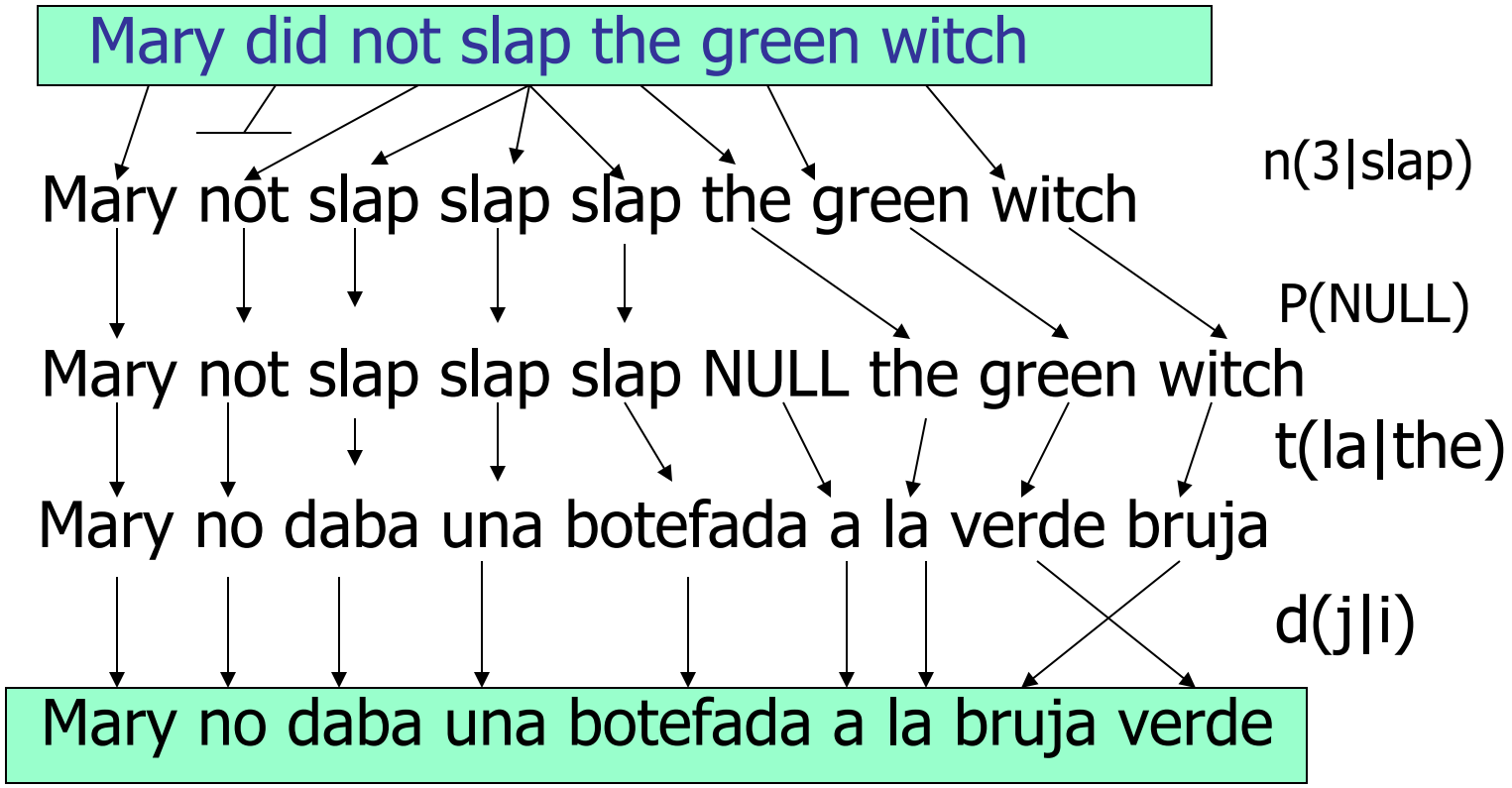


AER for HMMs

Model	AER
Model 1 INT	19.5
HMM E→F	11.4
HMM F→E	10.8
HMM AND	7.1
HMM INT	4.7
GIZA M4 AND	6.9

Models 3, 4, and 5: Fertility

IBM Models 3/4/5



[from Al-Onaizan and Knight, 1998]

Examples: Translation and Fertility

the

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
le	0.497	1	0.746
la	0.207	0	0.254
les	0.155		
l′	0.086		
ce	0.018		
cette	0.011		

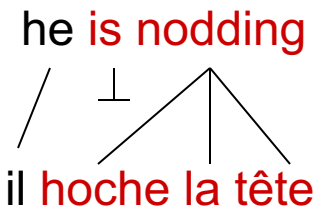
not

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

farmers

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

Example: Idioms



nodding

f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
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

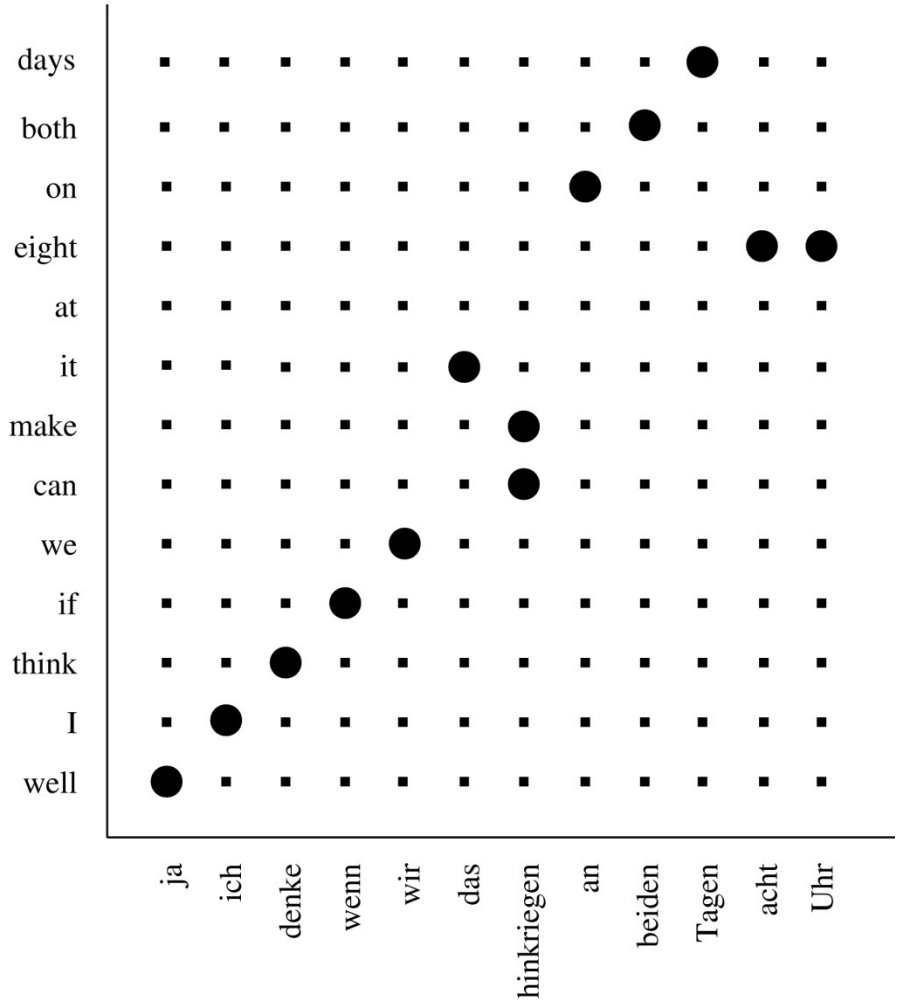
f	$t(f \mid e)$	ϕ	$n(\phi \mid e)$
devrait	0.330	1	0.649
devraient	0.123	0	0.336
devrions	0.109	2	0.014
faudrait	0.073		
faut	0.058		
doit	0.058		
aurait	0.041		
doivent	0.024		
devons	0.017		
devrais	0.013		

Some Results

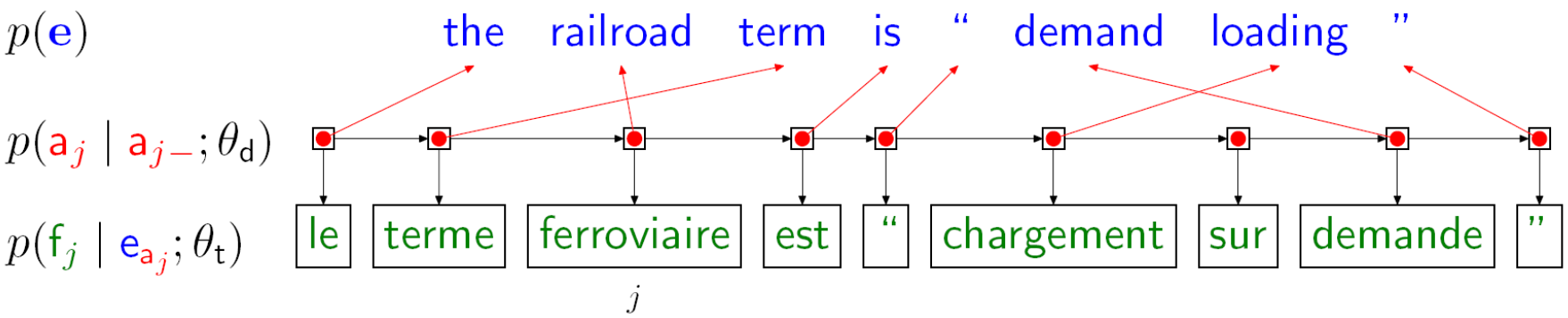
- [Och and Ney 03]

Model	Training scheme	0.5K	8K	128K	1.47M
Dice		50.9	43.4	39.6	38.9
Dice+C		46.3	37.6	35.0	34.0
Model 1	1^5	40.6	33.6	28.6	25.9
Model 2	$1^5 2^5$	46.7	29.3	22.0	19.5
HMM	$1^5 H^5$	26.3	23.3	15.0	10.8
Model 3	$1^5 2^5 3^3$	43.6	27.5	20.5	18.0
	$1^5 H^5 3^3$	27.5	22.5	16.6	13.2
Model 4	$1^5 2^5 3^3 4^3$	41.7	25.1	17.3	14.1
	$1^5 H^5 3^3 4^3$	26.1	20.2	13.1	9.4
	$1^5 H^5 4^3$	26.3	21.8	13.3	9.3
Model 5	$1^5 H^5 4^3 5^3$	26.5	21.5	13.7	9.6
	$1^5 H^5 3^3 4^3 5^3$	26.5	20.4	13.4	9.4
Model 6	$1^5 H^5 4^3 6^3$	26.0	21.6	12.8	8.8
	$1^5 H^5 3^3 4^3 6^3$	25.9	20.3	12.5	8.7

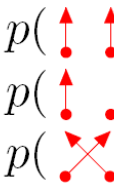
Phrase Movement



The HMM Model



Distortion θ_d

$p($  $) = 0.6$

$p($ $) = 0.2$

$p($ $) = \mathbf{0.1}$

$\ldots$

Translation θ_t

$p($ the $\rightarrow$ le $) = 0.53$

$p($ the $\rightarrow$ la $) = 0.24$

$p($ **railroad** $\rightarrow$ **ferroviaire** $) = \mathbf{0.19}$

$p($ NULL $\rightarrow$ le $) = 0.12$

$\ldots$