CSE 517
Natural Language Processing
Winter 2017

Machine Translation

Yejin Choi

Slides from Dan Klein, Luke Zettlemoyer, Dan Jurafsky, Ray Mooney

Translation: Codebreaking?

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 (1955:18, quoting a letter he wrote in 1947)

Brief History of NLP

- Mid 1950's – mid 1960's: Birth of NLP and Linguistics
  - At first, people thought MT would be easy! Researchers predicted that "machine translation" can be solved in 3 years or so.
- Mid 1960's – Mid 1970's: A Dark Era
  - People started believing that machine translation is impossible.
- 1970's and early 1980's – Slow Revival of NLP
  - Small toy problems, linguistic heavy, weak empirical evaluation
- Late 1980's and 1990's – Statistical Revolution!
  - By this time, the computing power increased substantially .
  - Data-driven, statistical approaches with simple representation.

➔ "*Whenever I fire a linguist, our MT performance improves.*" (Jelinek,1988)

- 2000's – Statistics Powered by Linguistic Insights
  - More complex statistical models & richer linguistic representations.

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.

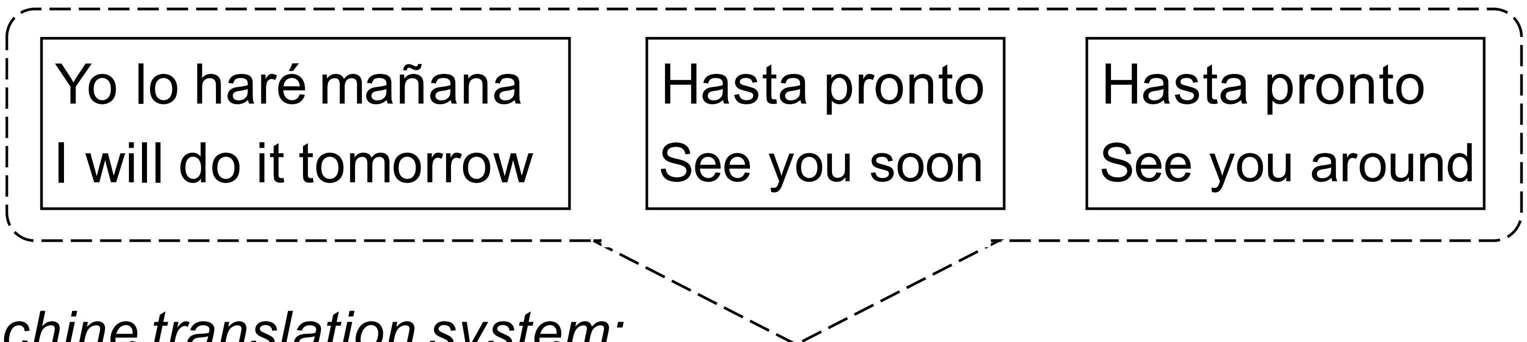
Atlanta, taken the killer of the palace of Justice

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.

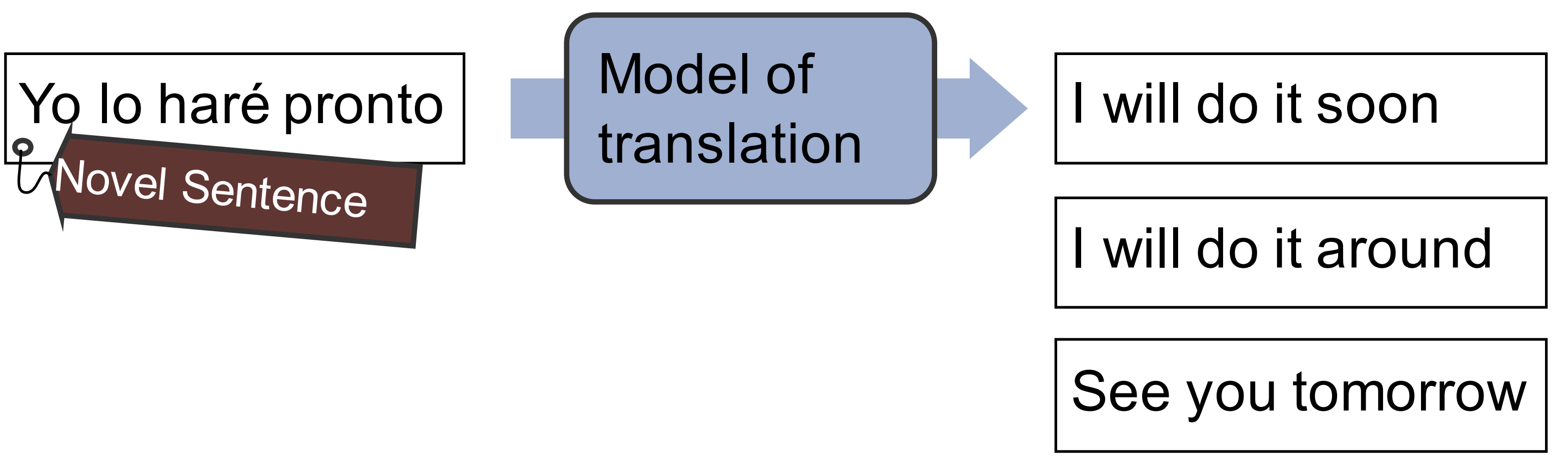
Corpus-Based MT

Modeling correspondences between languages

Sentence-aligned parallel corpus:

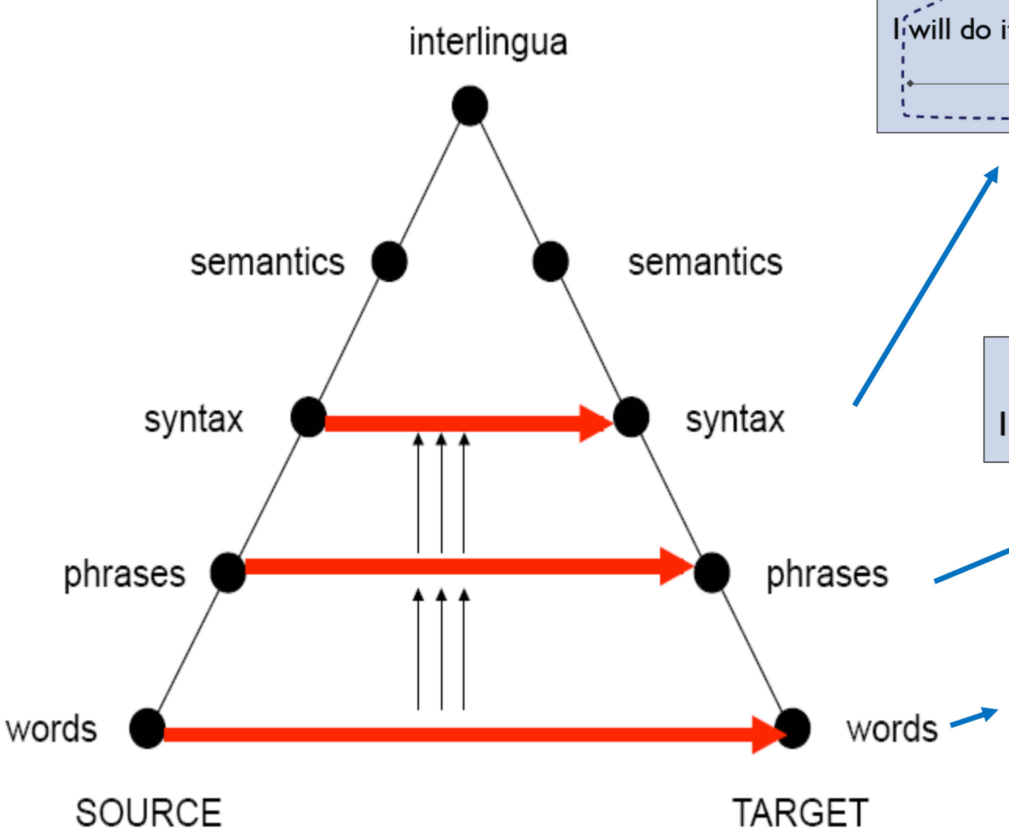


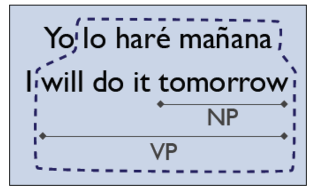
Machine translation system:

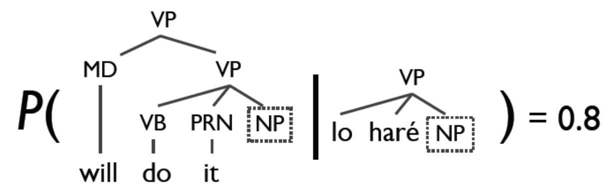


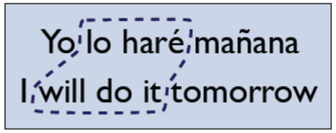
Levels of Transfer

“Vauquois Triangle”

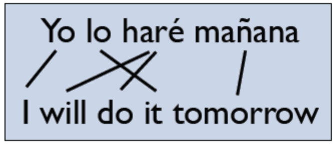








English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



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

General Approaches

- Rule-based approaches
  - Expert system-like rewrite systems
  - Interlingua methods (analyze and generate)
  - Lexicons come from humans
  - Can be very fast, and can accumulate a lot of knowledge over time (e.g. **Systran**)
- Statistical approaches
  - Word-to-word translation
  - Phrase-based translation
  - Syntax-based translation (tree-to-tree, tree-to-string)
  - Trained on parallel corpora
  - Usually noisy-channel (at least in spirit)

Translation is hard!

zi 自 zhu 助 zhong 终 duan 端

self help terminal device

help oneself terminating machine

(ATM, “self-service terminal”)

Examples from Liang Huang

Translation is hard!

Examples from Liang Huang

Translation is hard!

Examples from Liang Huang

Translation is hard!

Examples from Liang Huang

Translation is hard!

Examples from Liang Huang

or even…

Examples from Liang Huang

Human Evaluation

Madame la présidente, votre présidence de cette institution a été marquante.
Mrs Fontaine, your presidency of this institution has been outstanding.
Madam President, president of this house has been discoveries.
Madam President, your presidency of this institution has been impressive.

Je vais maintenant m'exprimer brièvement en irlandais.
I shall now speak briefly in Irish .
I will now speak briefly in Ireland .
I will now speak briefly in Irish .

Nous trouvons en vous un président tel que nous le souhaitions.
We think that you are the type of president that we want.
We are in you a president as the wanted.
We are in you a president as we the wanted.

Evaluation Questions:

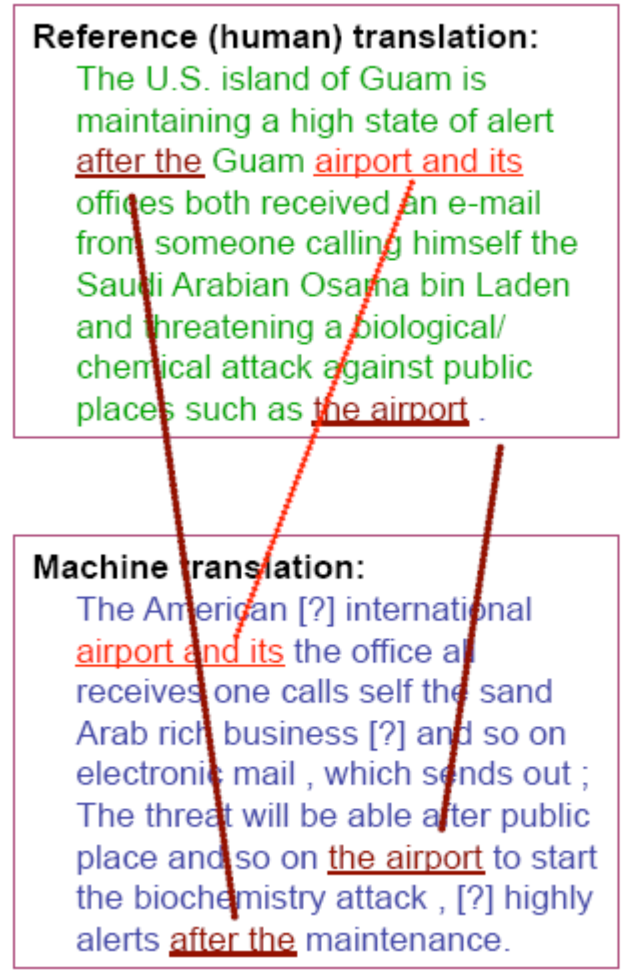
- Are translations fluent/grammatical?
- Are they adequate (you understand the meaning)?

MT: Automatic Evaluation

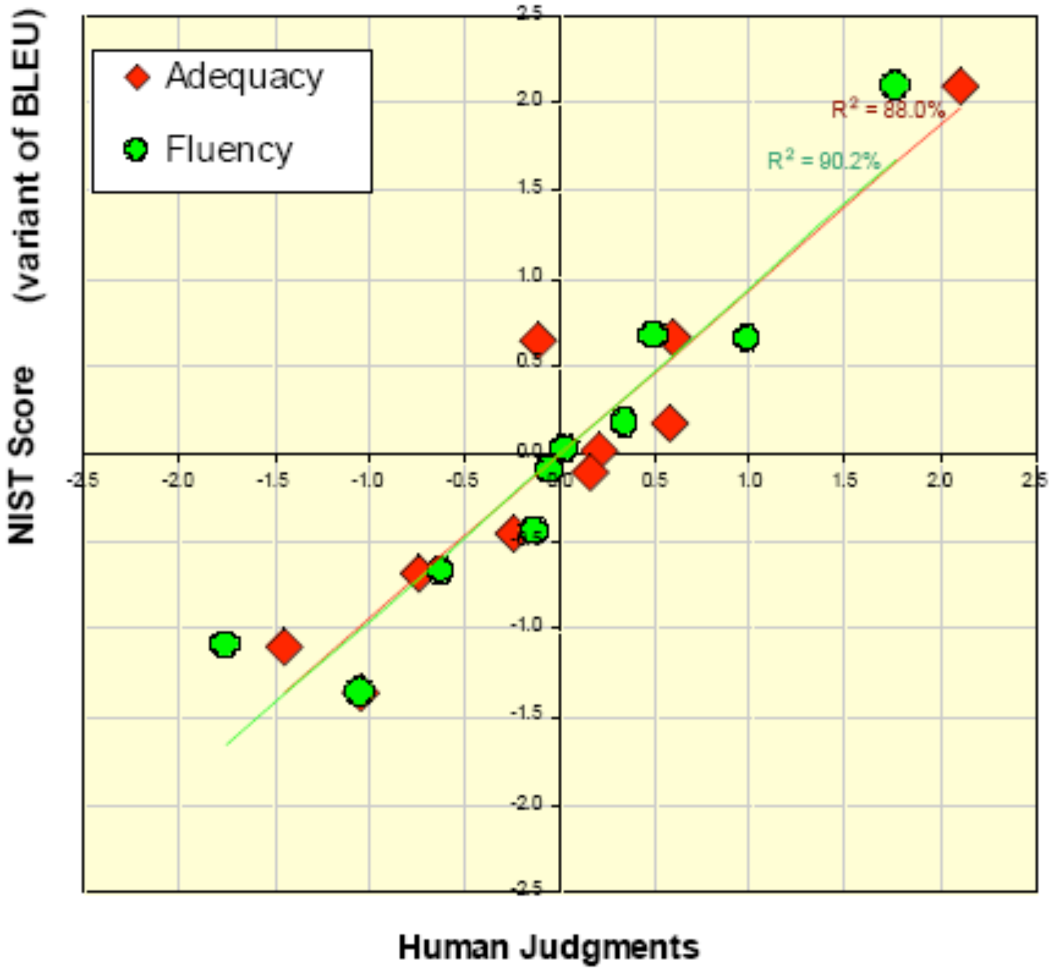
- Human evaluations: subject measures, fluency/adequacy

- 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)

- 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…

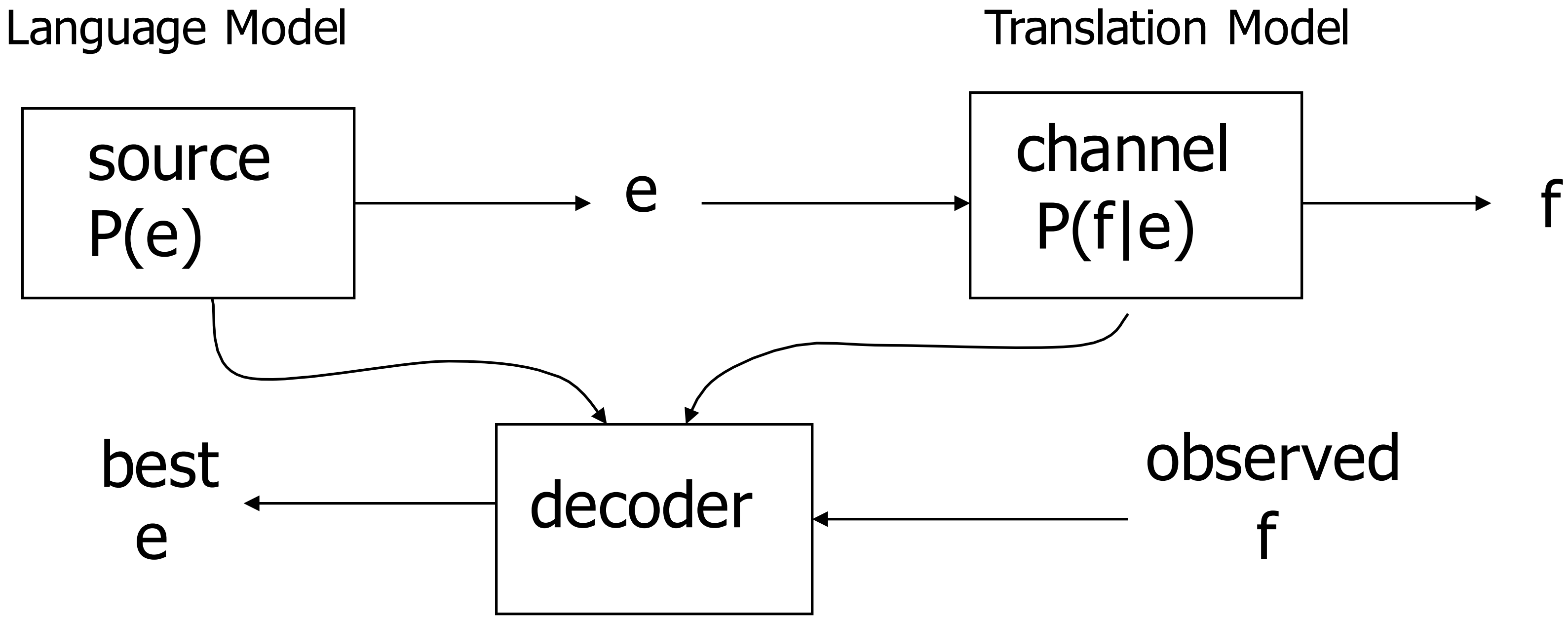


Automatic Metrics Work (?)



slide from G. Doddington (NIST)

MT System Components – Noisy Channel Model



$$\underset{e}{\operatorname{argmax}}\ P(e|f) = \underset{e}{\operatorname{argmax}}\ P(f|e)P(e)$$

Part I – Word Alignment Models

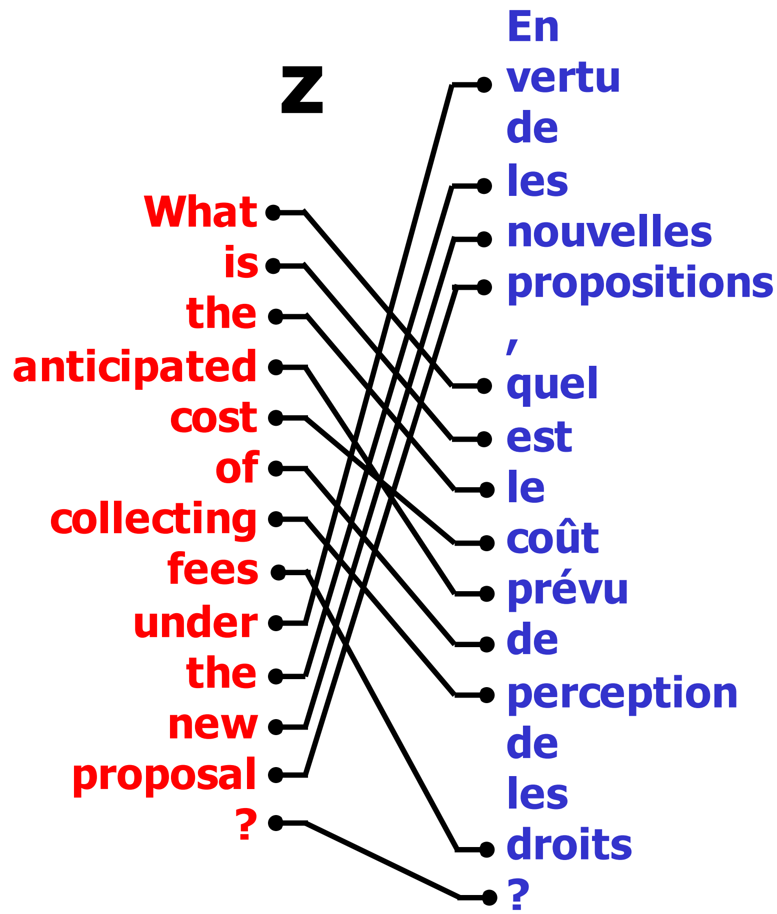
Word Alignment

x

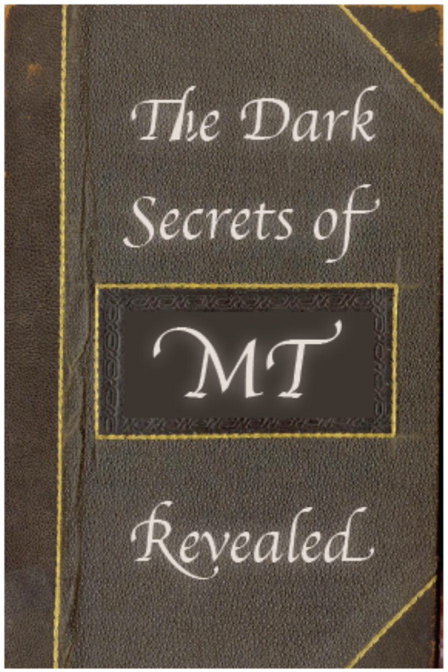
What is the anticipated cost of collecting fees under the new proposal?

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

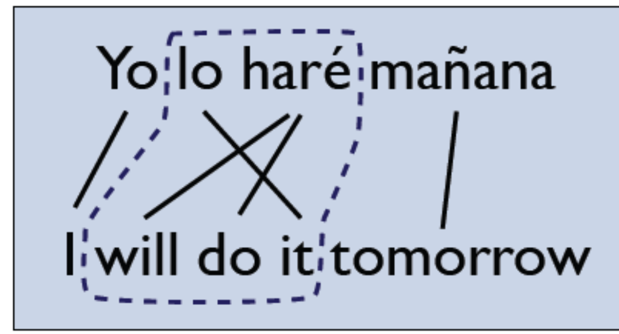
z



Word Alignment



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



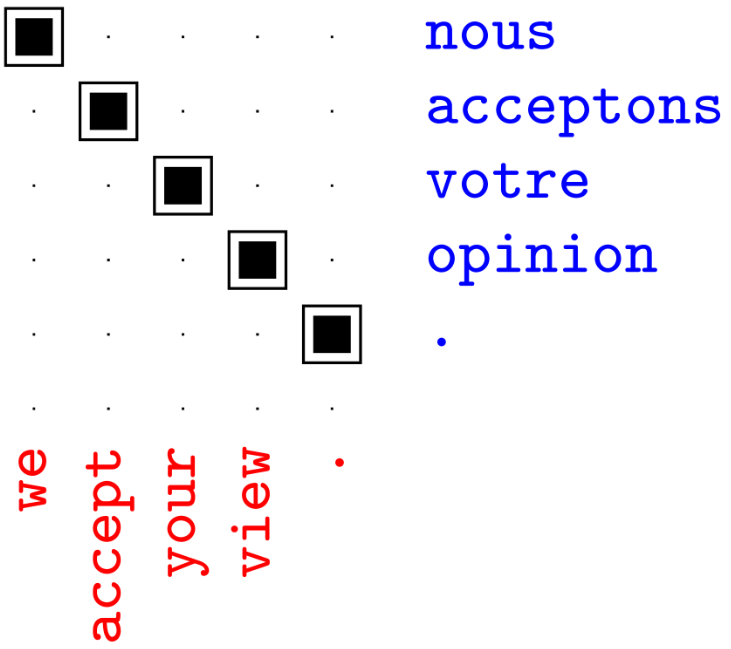
Unsupervised Word Alignment

- 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



The IBM Translation Models

[Brown et al 1993]

The Mathematics of Statistical Machine Translation: Parameter Estimation

Peter F. Brown*
IBM T.J. Watson Research Center

Stephen A. Della Pietra*
IBM T.J. Watson Research Center

Vincent J. Della Pietra*
IBM T.J. Watson Research Center

Robert L. Mercer*
IBM T.J. Watson Research Center

We describe a series of five statistical models of the translation process and give algorithms for estimating the parameters of these models given a set of pairs of sentences that are translations of one another. We define a concept of word-by-word alignment between such pairs of sentences. For any given pair of such sentences each of our models assigns a probability to each of the possible word-by-word alignments. We give an algorithm for seeking the most probable of these alignments. Although the algorithm is suboptimal, the alignment thus obtained accounts well for the word-by-word relationships in the pair of sentences. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. Accordingly, we have restricted our work to these two languages; but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus.

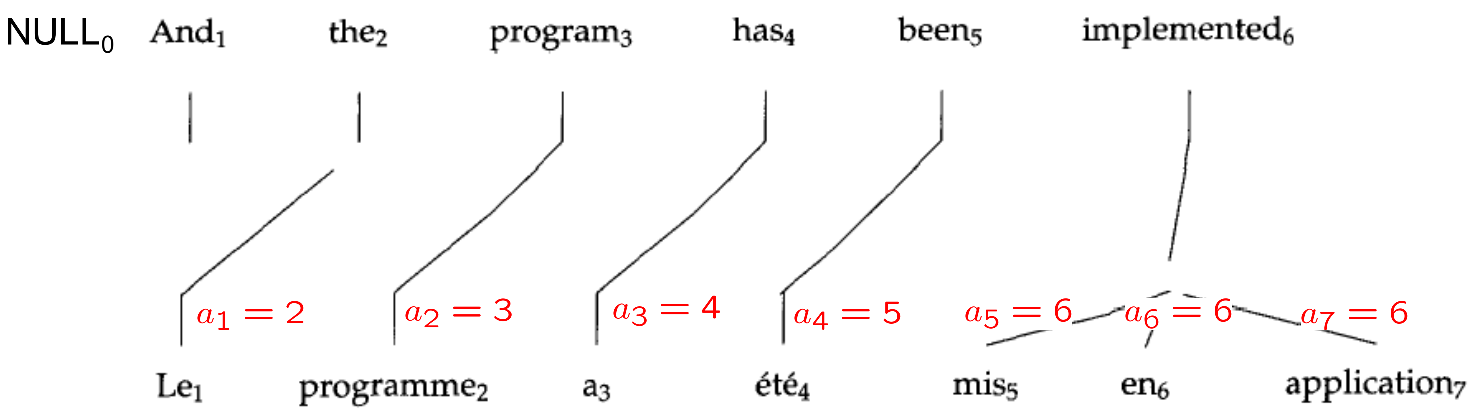
IBM Model 1 (Brown 93)

- Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer
- *The mathematics of statistical machine translation: Parameter estimation*. In: *Computational Linguistics* 19 (2), 1993.
- 3667 citations.

Vincent (left) and Stephen Della Pietra

IBM Model 1 (Brown 93)

- Model parameters: $t(f|e) := p('e' \text{ is translated into } 'f'|e)$
- A (hidden) alignment vector $(a_1, ..., a_m)$ where $a_i = j$ means 'i'th target word is translated from 'j'th source word.
- Include a "null" word on the source side
- This alignment vector defines 1-to-many mappings. (why?)



$$p(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})$$

Uniform alignment model!

$$= \prod_{i=1}^{m} \frac{1}{l+1} t(f_i | e_{a_i})$$

IBM Model 1: Learning

- If given data with alignment $\{(e_1 \ldots e_l, a_1 \ldots a_m, f_1 \ldots f_m)_k | k=1..n\}$

$$t_{ML}(f|e) = \frac{c(e,f)}{c(e)} \quad \text{where} \quad \begin{array}{l} \delta(k,i,j) = 1 \text{ if } a_i^{(k)} = j, \; 0 \text{ otherwise} \\ c(e,f) = \sum_k \sum_{i \text{ s.t. } e_i = e} \sum_{j \text{ s.t. } f_j = f} \delta(k,i,j) \end{array}$$

- In practice, no such data available at large scale.
- Thus, learn the translation model parameters while keeping alignment as latent variables, using EM,
  - Repeatedly re-compute the expected counts:

$$\delta(k,i,j) = \frac{t(f_i^{(k)}|e_j^{(k)})}{\sum_{j'} t(f_i^{(k)}|e_{j'}^{(k)})}$$

- Basic idea: compute expected source for each word, update co-occurrence statistics, repeat

Sample EM Trace for Alignment
(IBM Model 1 with no NULL Generation)

Training Corpus

green house	the house
casa verde	la casa

Translation Probabilities

	verde	casa	la
green	1/3	1/3	1/3
house	1/3	1/3	1/3
the	1/3	1/3	1/3

Assume uniform initial probabilities

Compute Alignment Probabilities P(A, F | E)

green house casa verde	green house casa verde	the house la casa	the house la casa
1/3 X 1/3 = 1/9	1/3 X 1/3 = 1/9	1/3 X 1/3 = 1/9	1/3 X 1/3 = 1/9

Normalize to get P(A | F, E)

$$\frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2} \qquad \frac{1/9}{2/9} = \frac{1}{2}$$

Example cont.

green house casa verde	green house casa verde	the house la casa	the house la casa
1/2	**1/2**	**1/2**	**1/2**

Compute weighted translation counts

	verde	casa	la
green	1/2	1/2	0
house	1/2	1/2 + 1/2	1/2
the	0	1/2	1/2

Normalize rows to sum to one to estimate P(f | e)

	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Example cont.

Translation Probabilities

	verde	casa	la
green	1/2	1/2	0
house	1/4	1/2	1/4
the	0	1/2	1/2

Recompute Alignment Probabilities P(A, F | E)

green house	**green house**	**the house**	**the house**
casa verde	**casa verde**	**la casa**	**la casa**
1/2 X 1/4=1/8	**1/2 X 1/2=1/4**	**1/2 X 1/2=1/4**	**1/2 X 1/4=1/8**

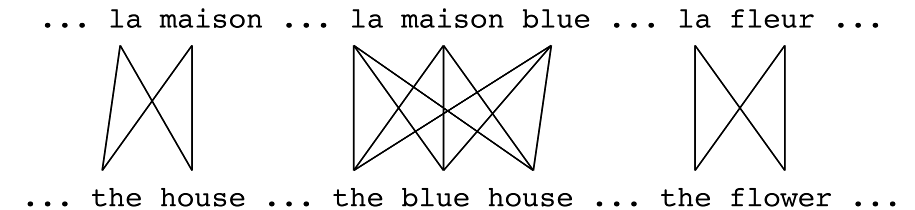
Normalize to get P(A | F, E)

$$\frac{1/8}{3/8} = \frac{1}{3} \qquad \frac{1/4}{3/8} = \frac{2}{3} \qquad \frac{1/4}{3/8} = \frac{2}{3} \qquad \frac{1/8}{3/8} = \frac{1}{3}$$

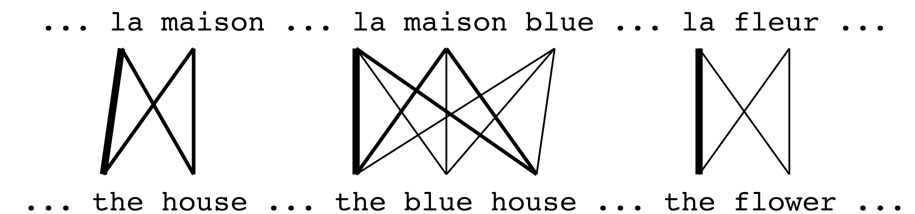
Continue EM iterations until translation parameters converge

IBM Model 1 - EM intuition

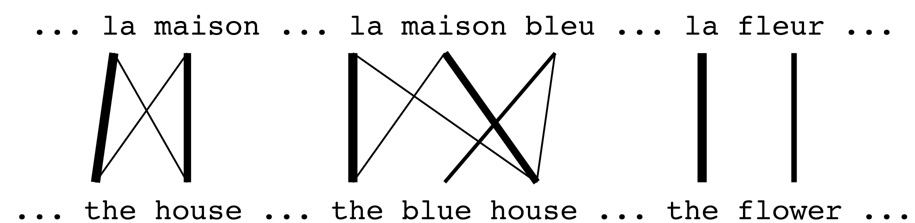
Step 1



Step 2

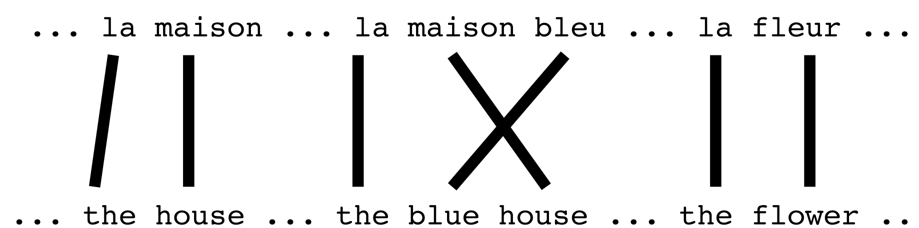


Step 3



...

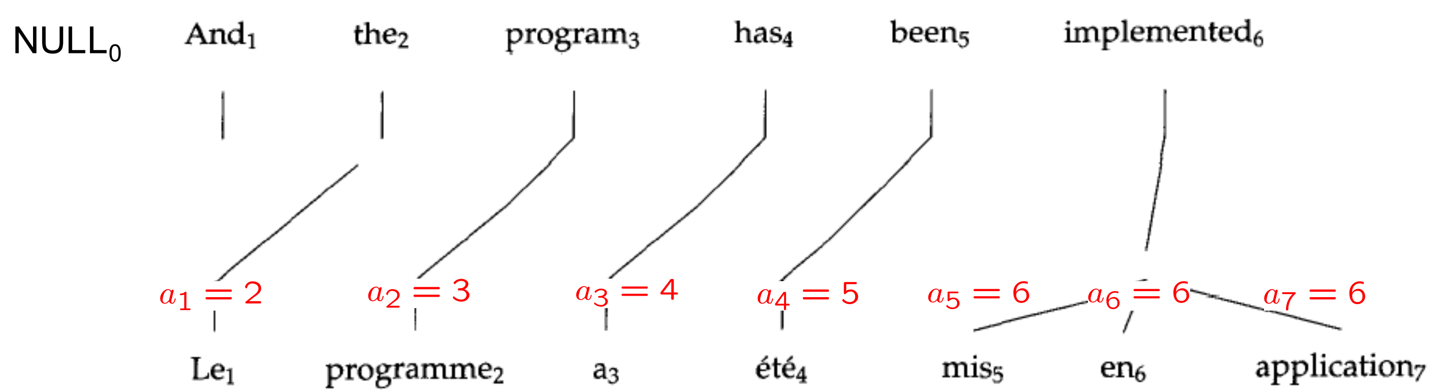
Step N



Example from Philipp Koehn

IBM Model 1: Inference

- Model parameters: $t(f|e) := p('e' \text{ is translated into } 'f'|e)$
- A (hidden) alignment vector $(a_1, \ldots, a_m)$ where $a_i = j$ means 'i'th target word is translated from 'j'th source word.



$$p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i|i,l,m)\, t(f_i|e_{a_i})$$

Uniform alignment model!

$$= \prod_{i=1}^{m} \frac{1}{l+1}\, t(f_i|e_{a_i})$$

- Inference: Find the best alignment a given (f,e) pairs. Is this hard?

Evaluating Alignments

- 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

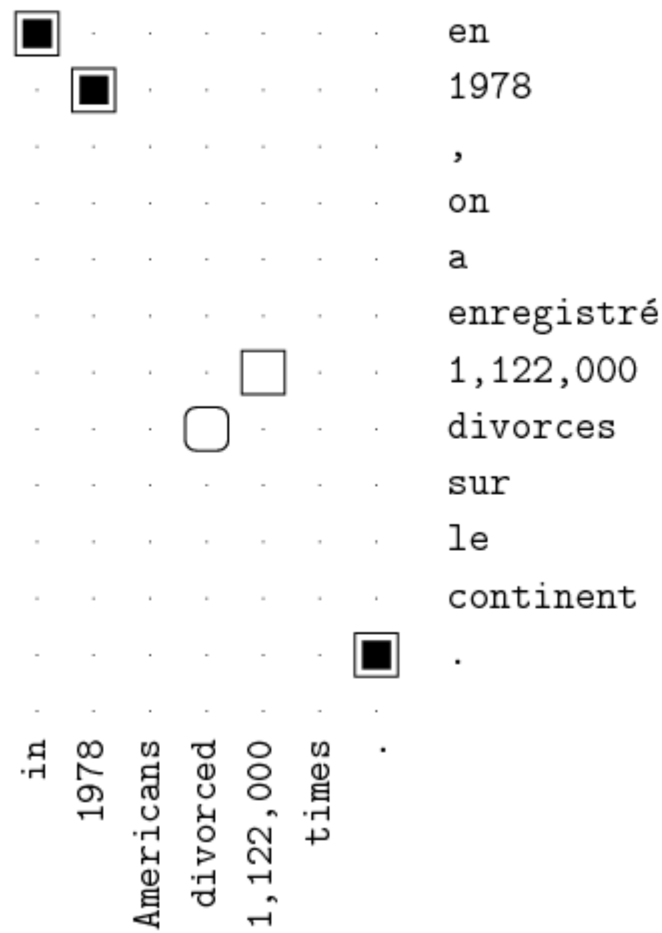
◯ = Possible

■ = Predicted

- A := predicted alignments
- S := sure alignments
- P := possible alignments
  (including sure alignments)

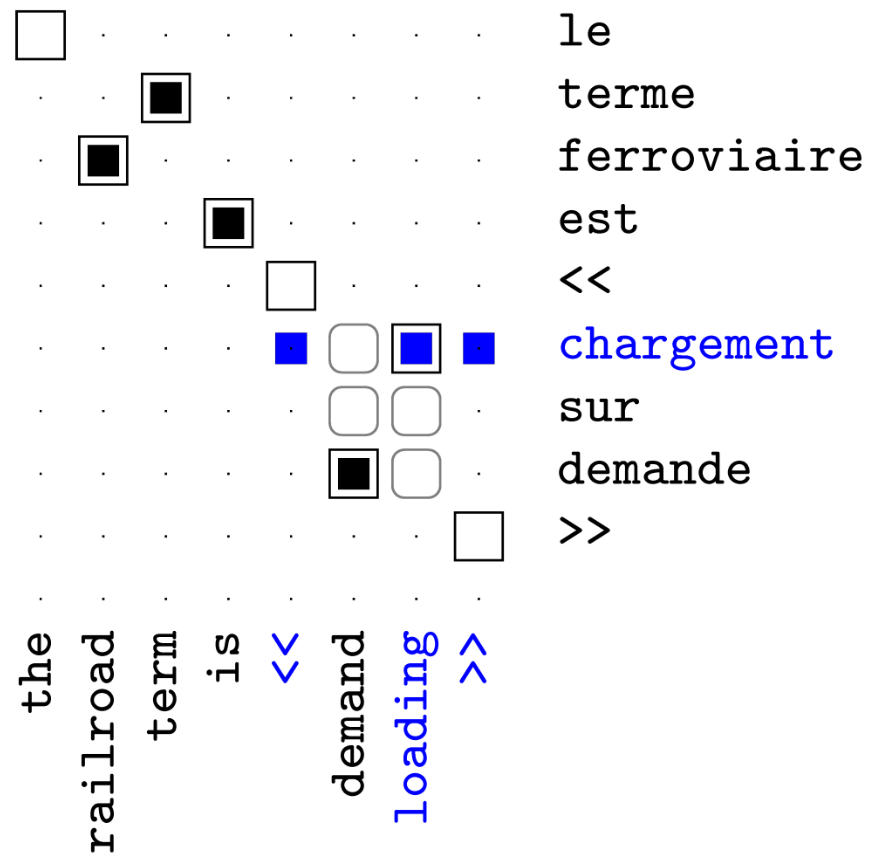
$$AER(A,S,P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

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



Problems with Model 1

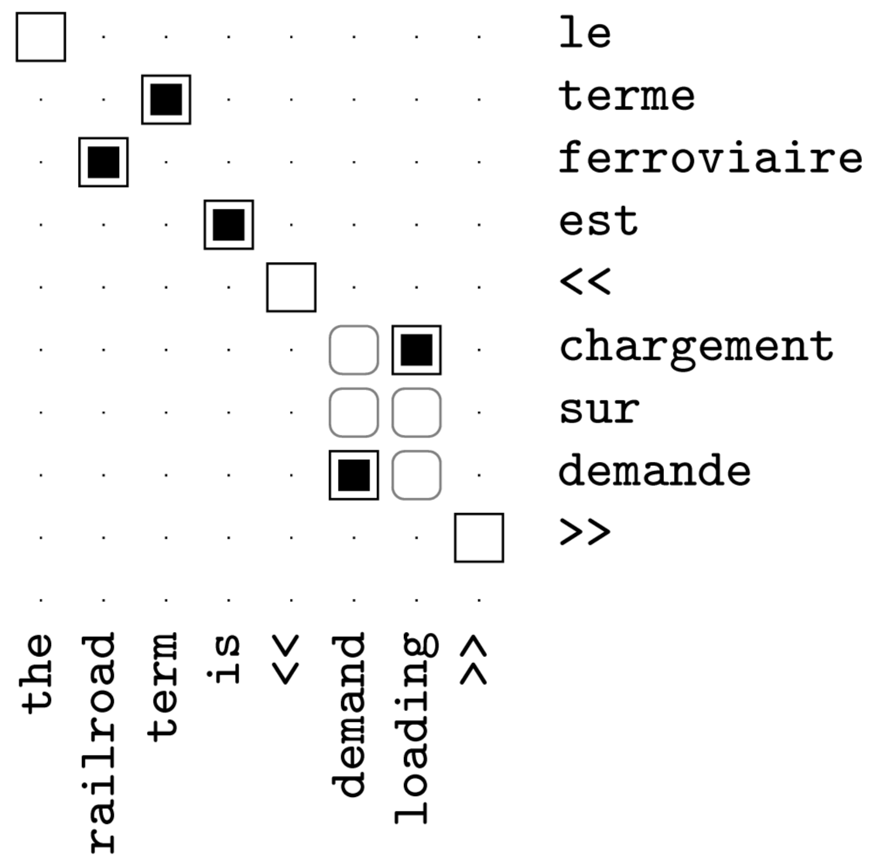
- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
  - Training data: 1.1M sentences of French-English text, Canadian Hansards
  - Evaluation metric: alignment error Rate (AER)
  - Evaluation data: 447 hand-aligned sentences



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

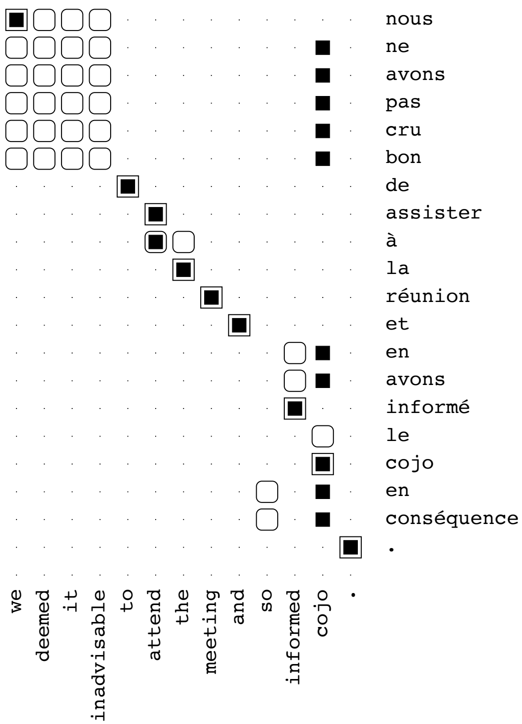


Joint Training?

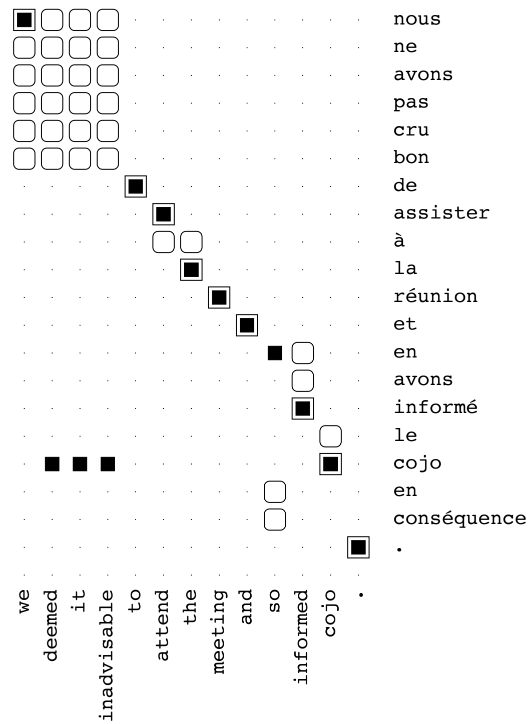
- "Alignment by agreement" (Liang et al, 2006)
  - Similar high precision to post-intersection
  - But recall is much higher
  - More confident about positing non-null alignments

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

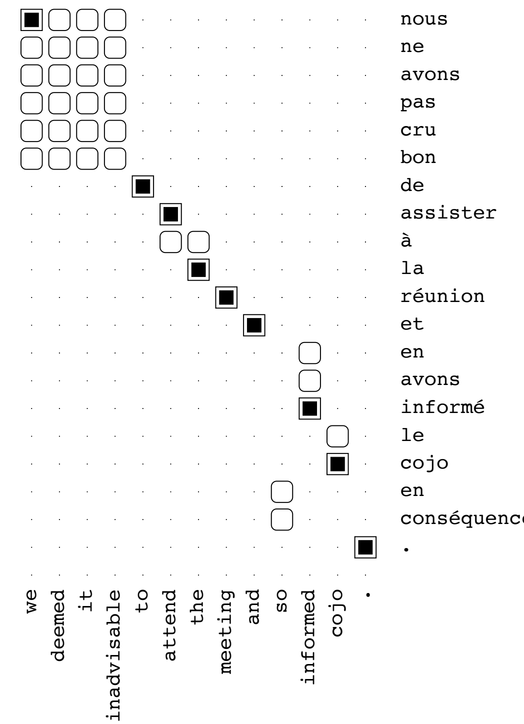
Independent Training



E→F: 84.2/92.0/13.0

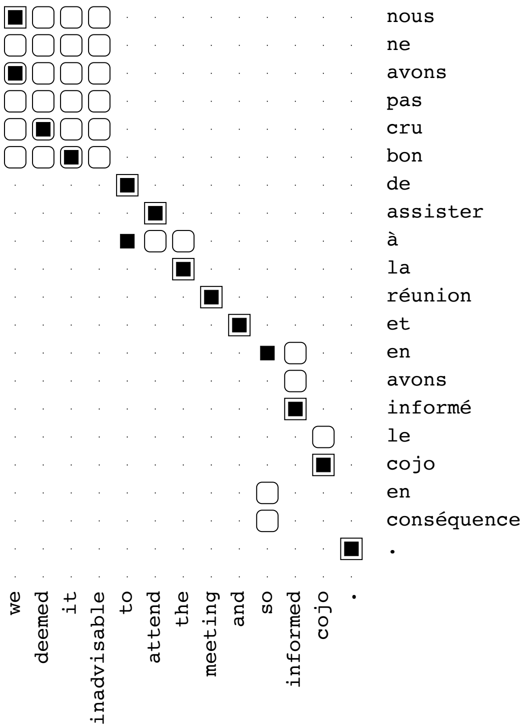


F→E: 86.9/91.1/11.5

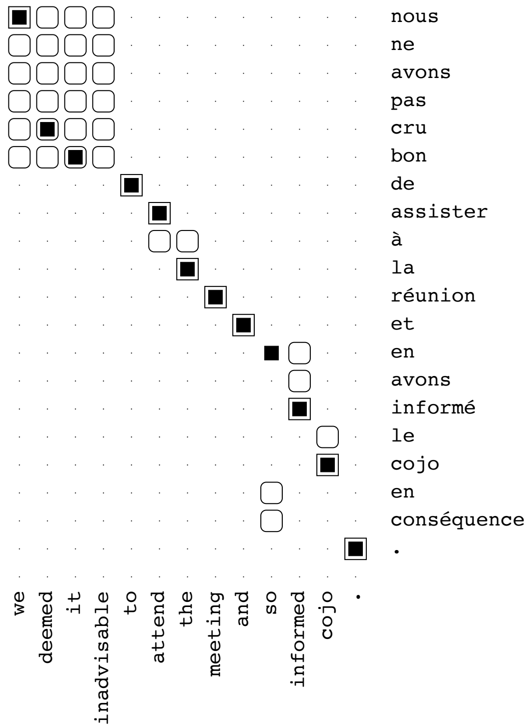


Intersection: 97.0/86.9/7.6

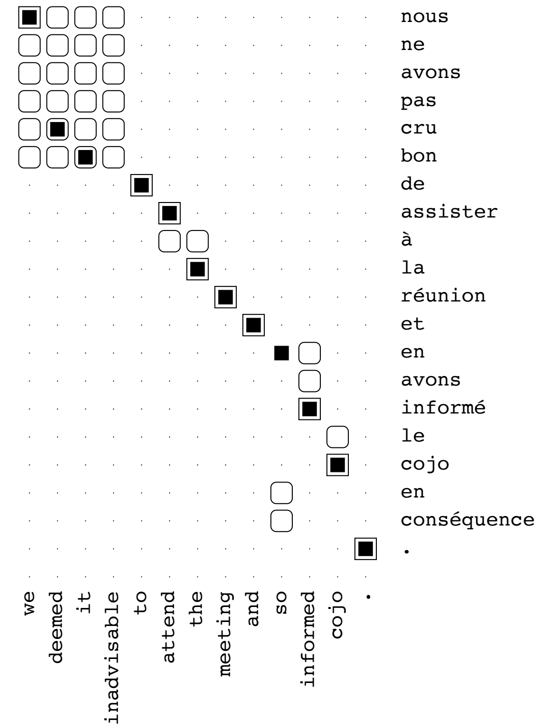
Joint Training



E→F: 89.9/93.6/8.7



F→E: 92.2/93.5/7.3



Intersection: 96.5/91.4/5.7

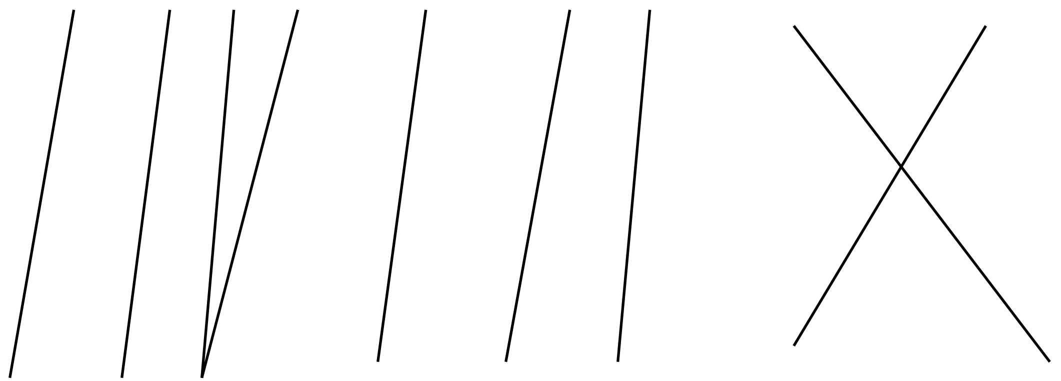
Monotonic Translation

Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes

Local Order Change

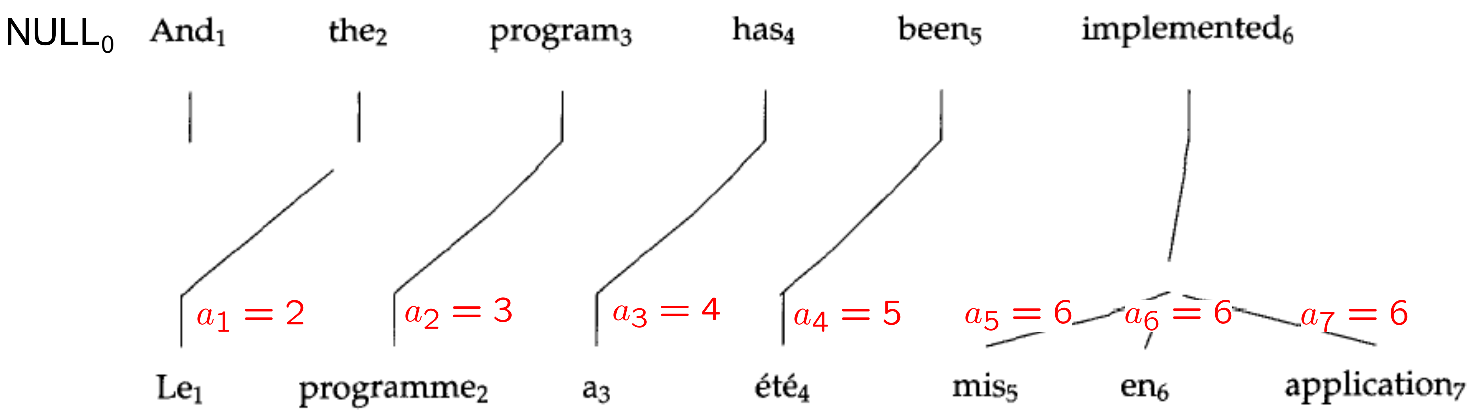
Japan is at the junction of four tectonic plates



Le Japon est au confluent de quatre plaques tectoniques

IBM Model 2 (Brown 93)

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



$$p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})$$

- Same decomposition as Model 1, but we will use a multi-nomial distribution for q!

IBM Model 2: Learning

- Given data $\{(e_1 \ldots e_l, a_1 \ldots a_m, f_1 \ldots f_m)_k | k=1..n\}$ where

$$\delta(k,i,j) = 1 \text{ if } a_i^{(k)} = j, \ 0 \text{ otherwise}$$

$$t_{ML}(f|e) = \frac{c(e,f)}{c(e)} \quad q_{ML}(j|i,l,m) = \frac{c(j|i,l,m)}{c(i,l,m)}$$

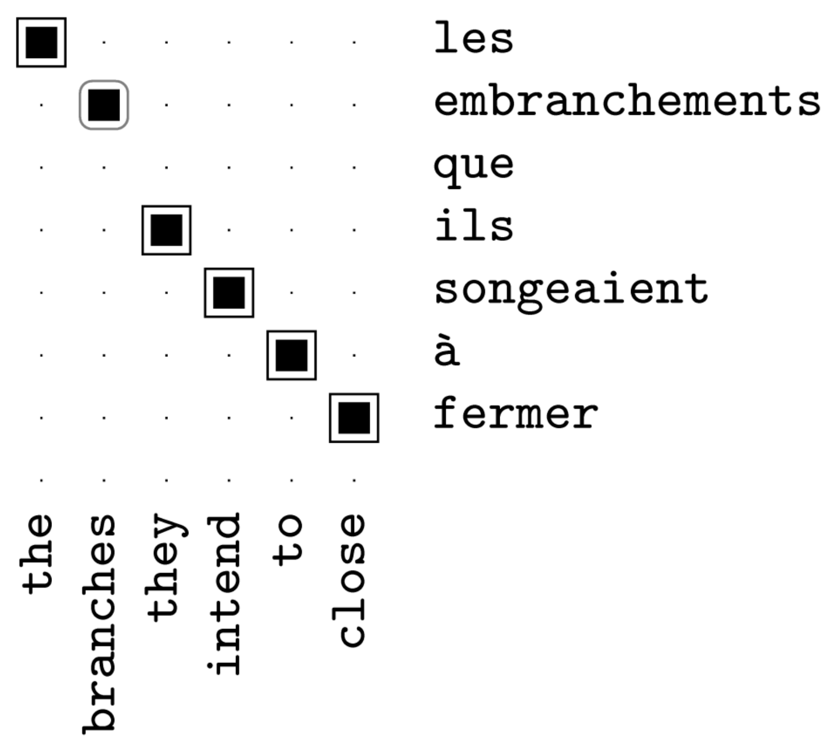
$$c(e,f) = \sum_k \sum_{i \text{ s.t. } e_i = e} \sum_{j \text{ s.t. } f_j = f} \delta(k,i,j)$$

- Better approach: re-estimated generative models with EM,
  - Repeatedly compute counts, using redefined deltas:

$$\delta(k,i,j) = \frac{q(j|i,l_k,m_k)t(f_i^{(k)}|e_j^{(k)})}{\sum_{j'} q(j'|i,l_k,m_k)t(f_i^{(k)}|e_{j'}^{(k)})}$$

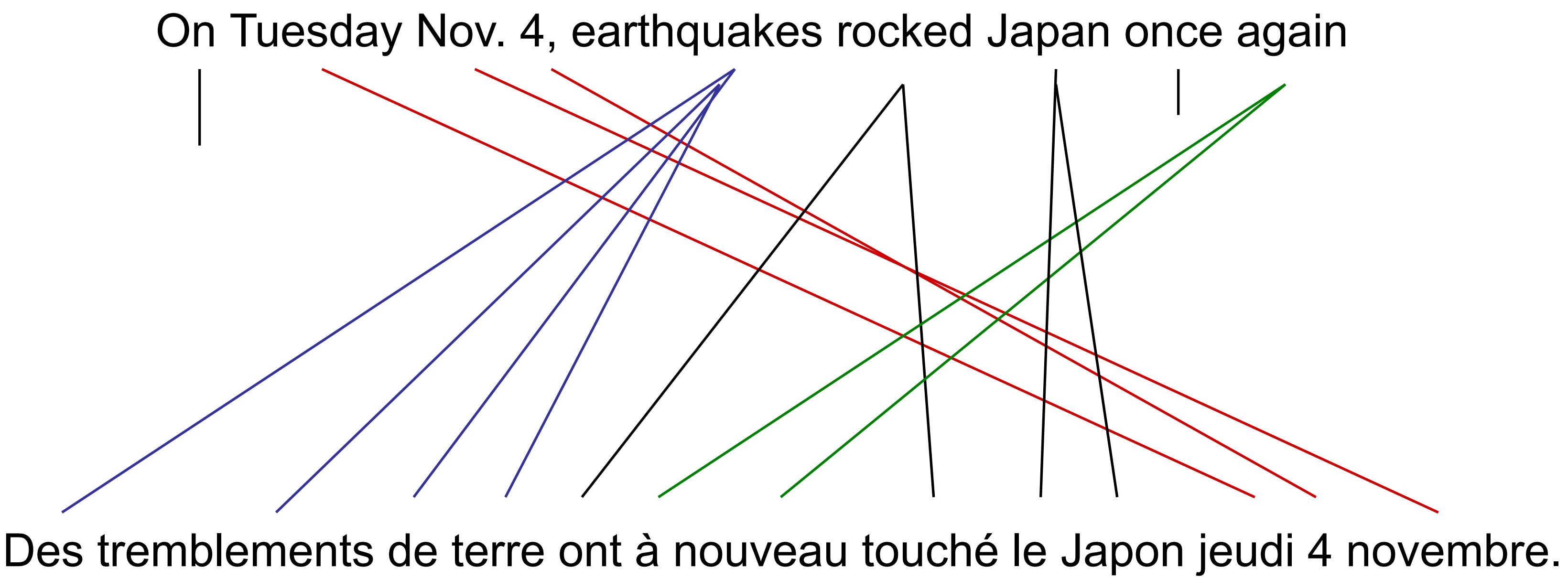
- Basic idea: compute expected source for each word, update co-occurrence statistics, repeat
- Q: What about inference? Is it hard?

Example

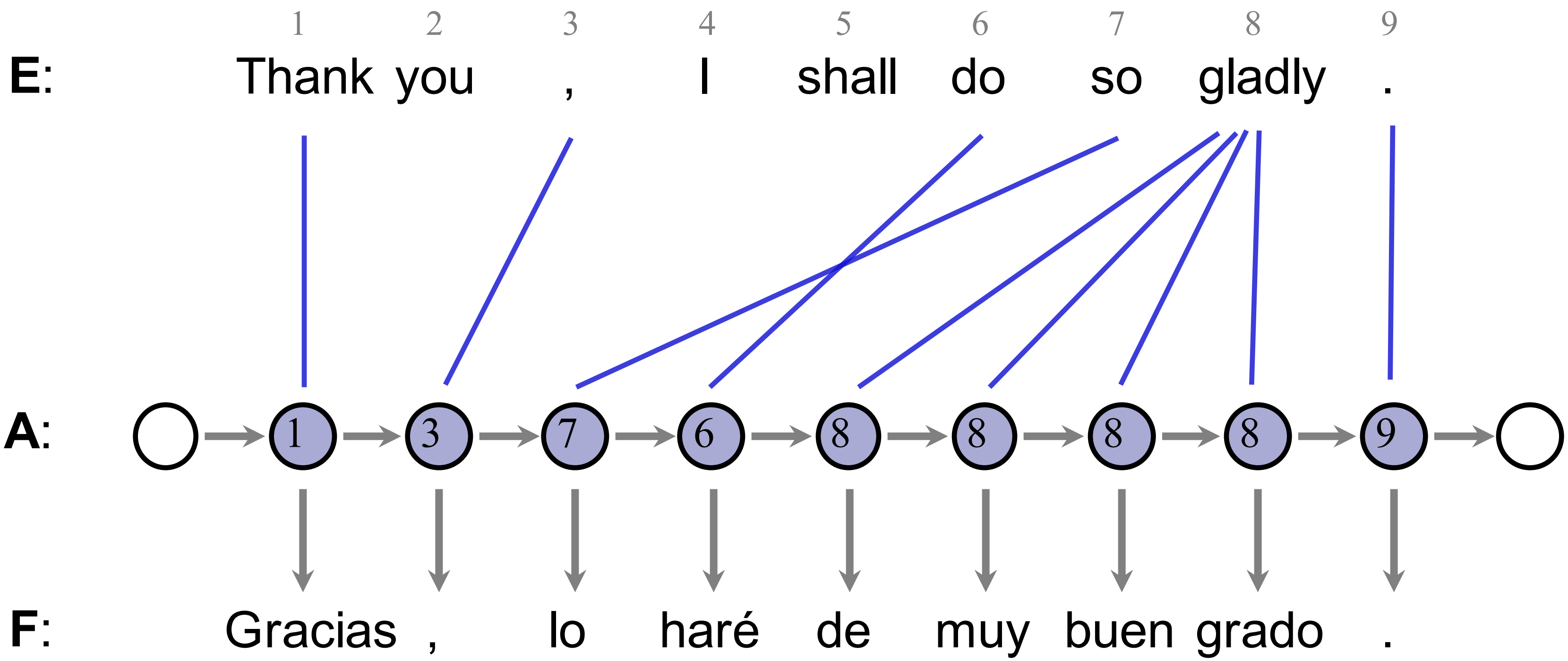


les
embranchements
que
ils
songeaient
à
fermer
the
branches
they
intend
to
close

Phrase Movement



The HMM Model



Model Parameters

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

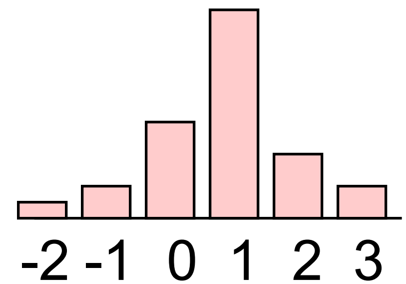
The HMM Model

- Model 2 can learn complex alignments
- 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

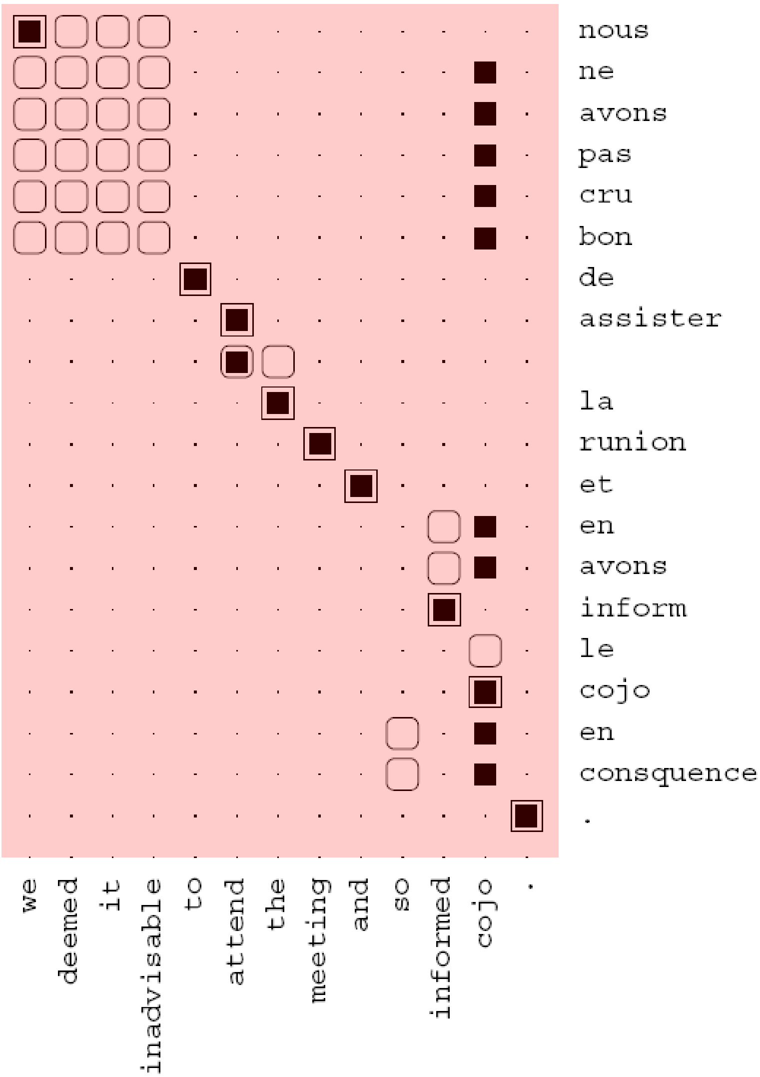
$$P(f, a|e) = \prod_j P(a_j|a_{j-1}) P(f_j|e_i)$$

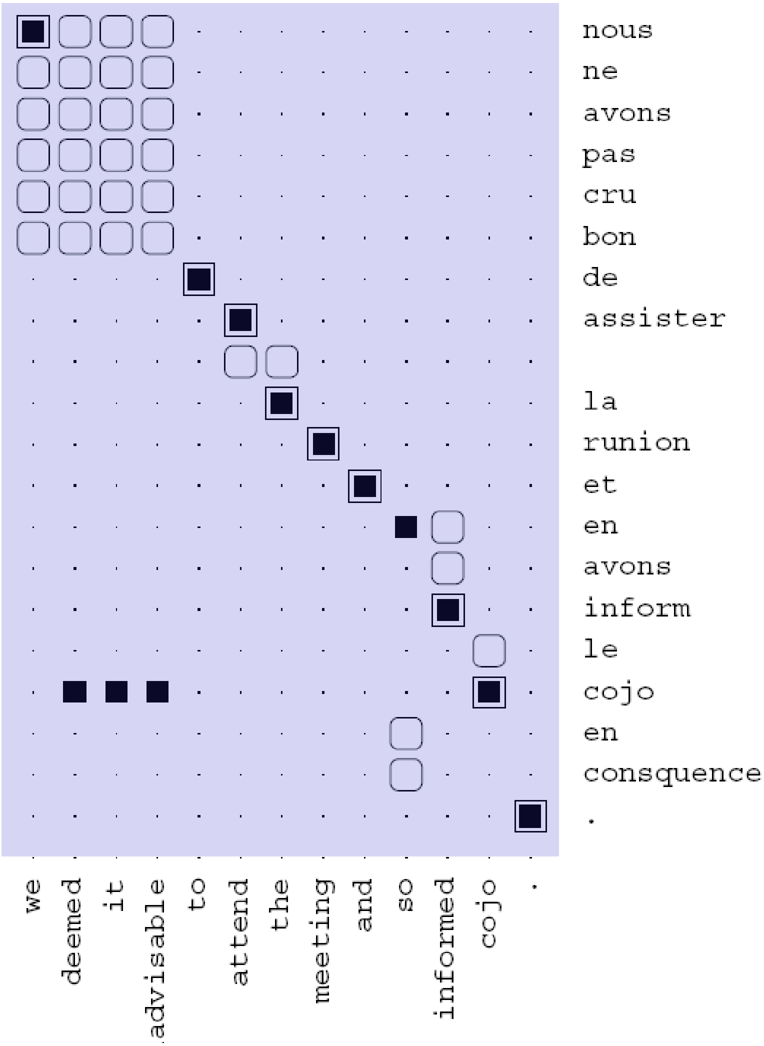
$$P(a_j - a_{j-1})$$



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

HMM Examples

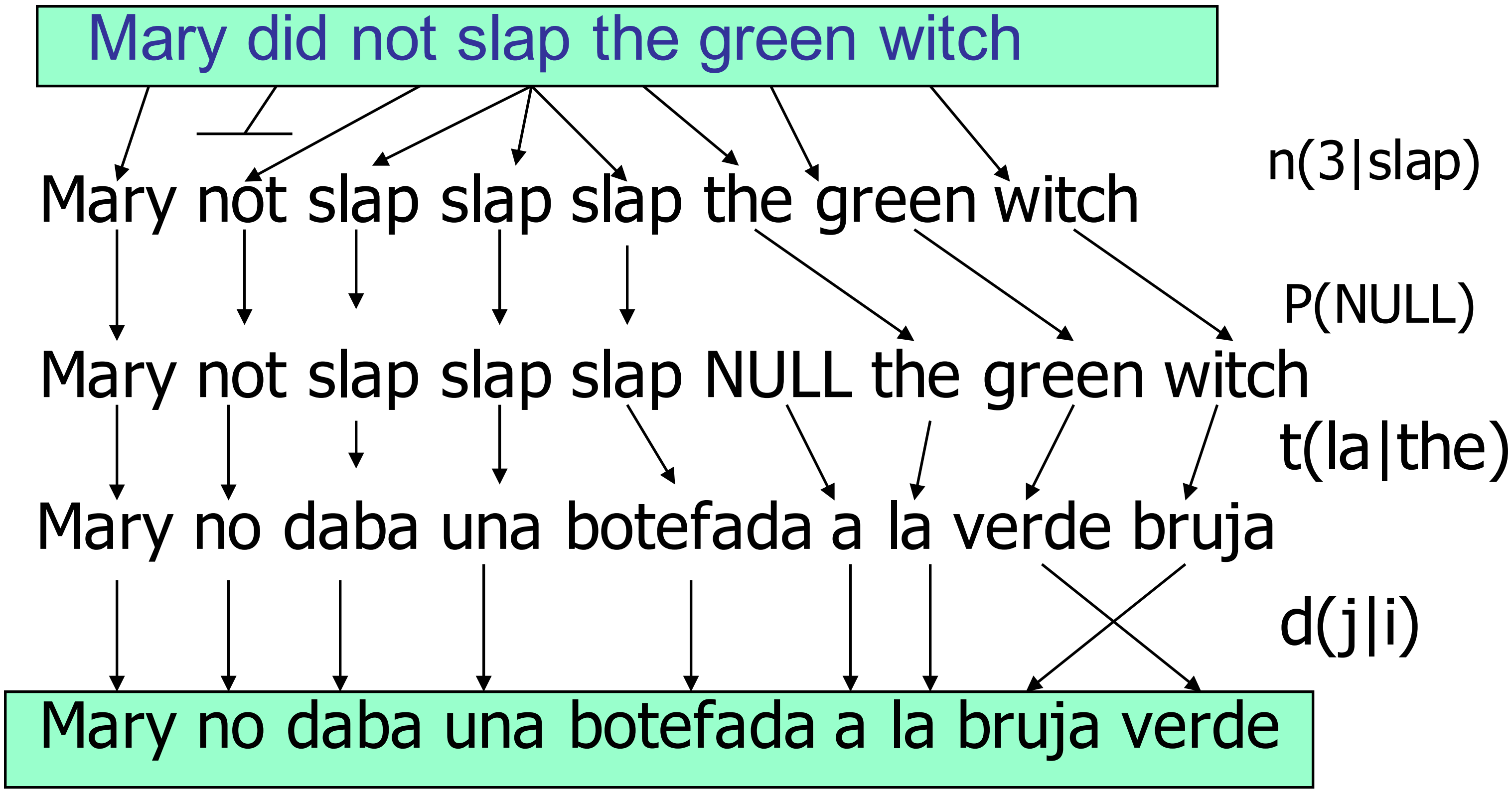




AER for HMMs

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

IBM Models 3/4/5



[from Al-Onaizan and Knight, 1998]

Overview of Alignment Models

Table 1
Overview of the alignment models.

Model	Alignment model	Fertility model	E-step	Deficient
Model 1	uniform	no	exact	no
Model 2	zero-order	no	exact	no
HMM	first-order	no	exact	no
Model 3	zero-order	yes	approximative	yes
Model 4	first-order	yes	approximative	yes
Model 5	first-order	yes	approximative	no
Model 6	first-order	yes	approximative	yes

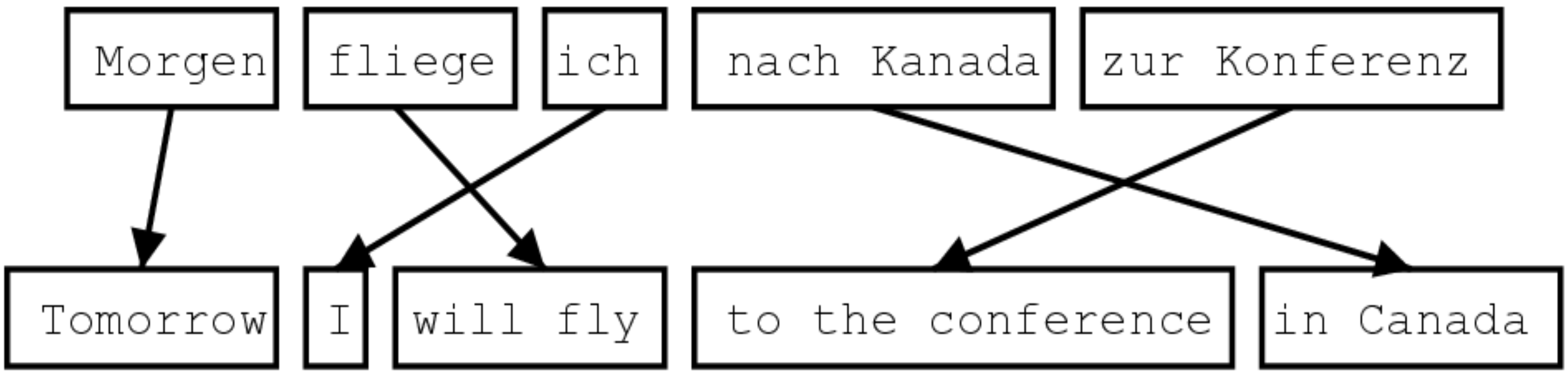
Some Results

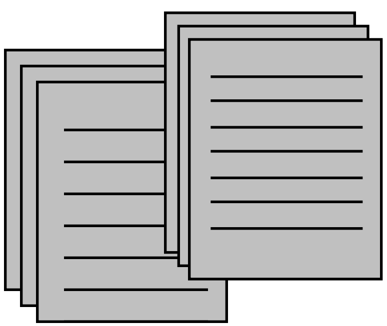
- [Och and Ney 03]

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

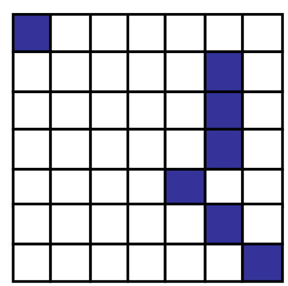
Part II - Phrase Translation Model

Phrase-Based Systems





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)

Phrase Translation Tables

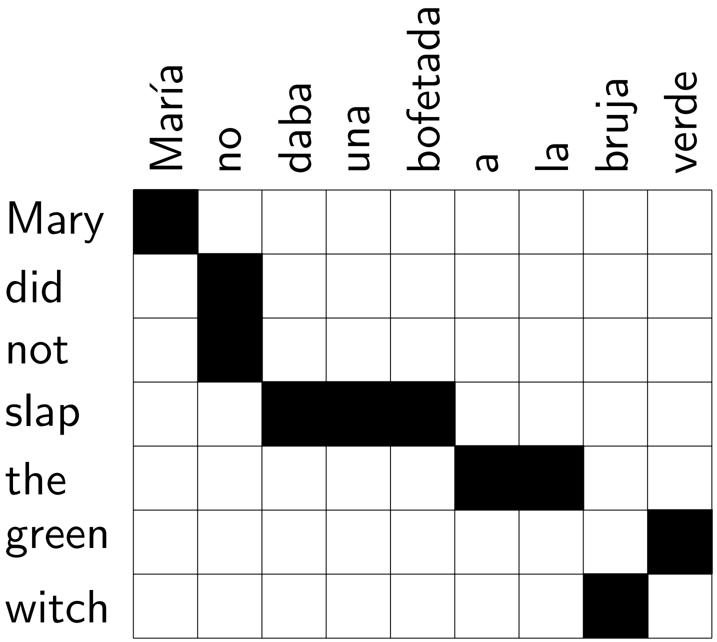
- Defines the space of possible translations
  - each entry has an associated "probability"
- One learned example, for "den Vorschlag" from Europarl data

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

  - This table is noisy, has errors, and the entries do not necessarily match our linguistic intuitions about consistency….

Extracting Phrases

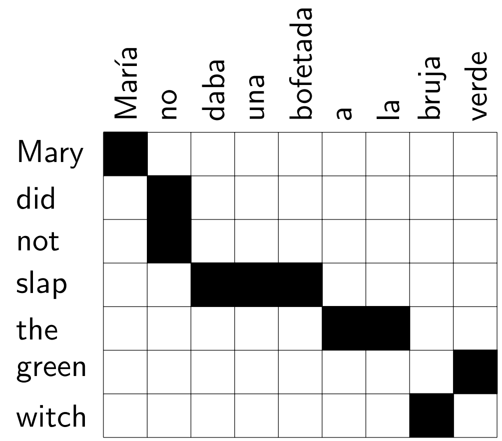
- We will use word alignments to find phrases

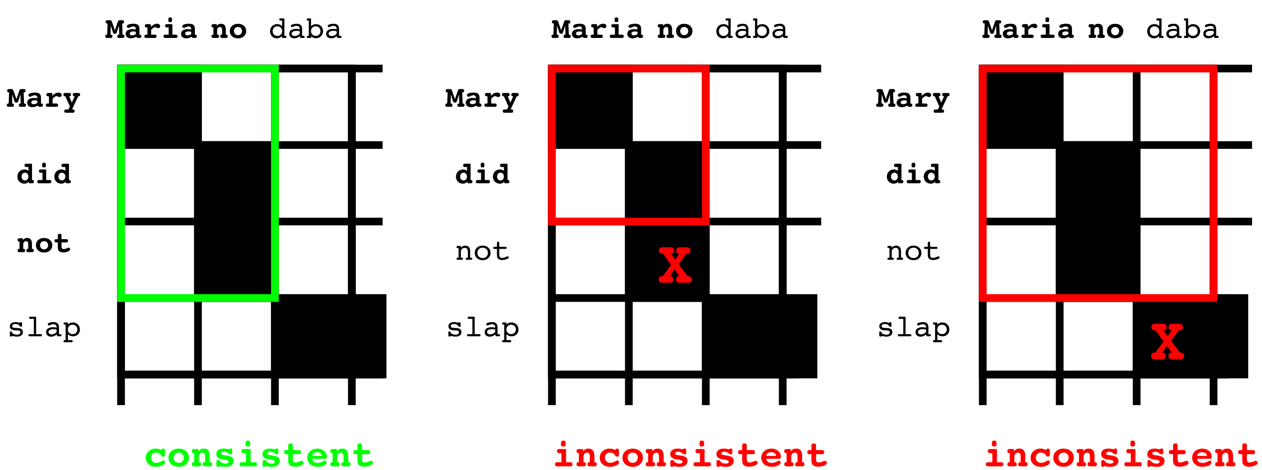


- Question: what is the best set of phrases?

Extracting Phrases

- Phrase alignment must
  - Contain at least one alignment edge
  - Contain all alignments for phrase pair





- Extract all such phrase pairs!

Phrase Pair Extraction Example

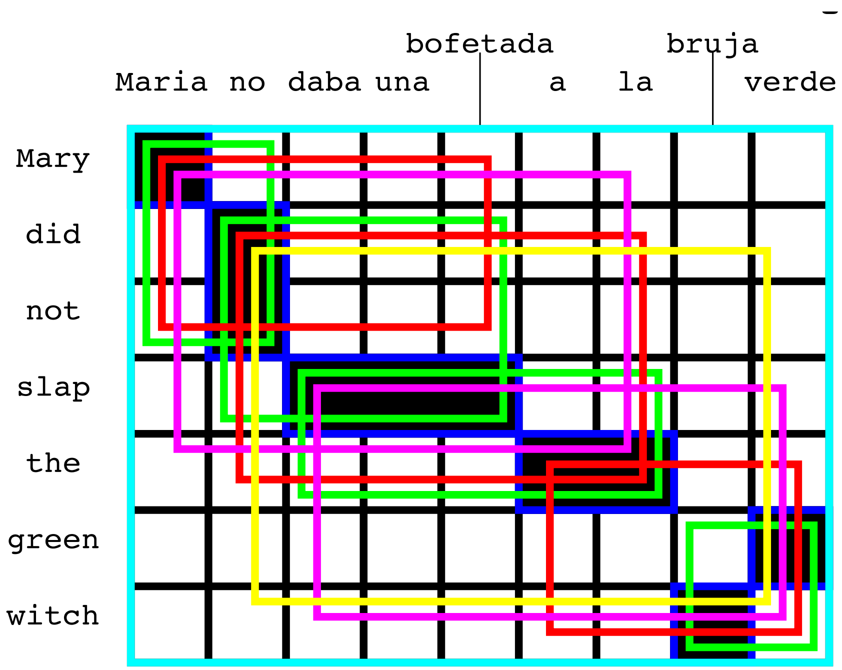
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

(Maria no daba una bofetada, Mary did not slap), (no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

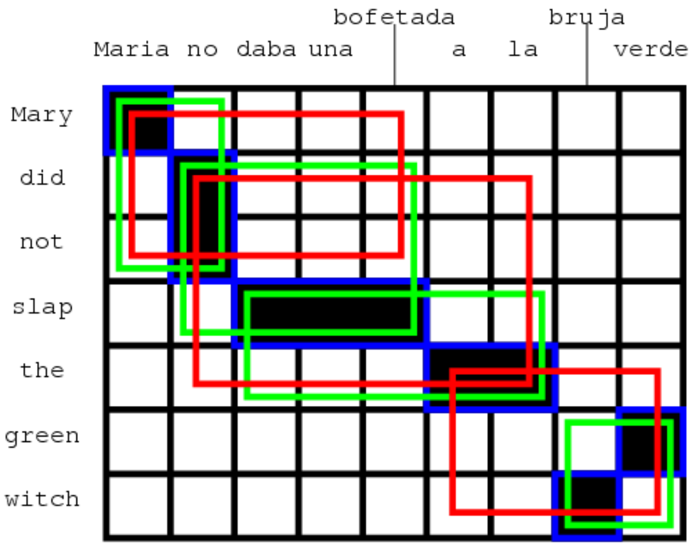
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)

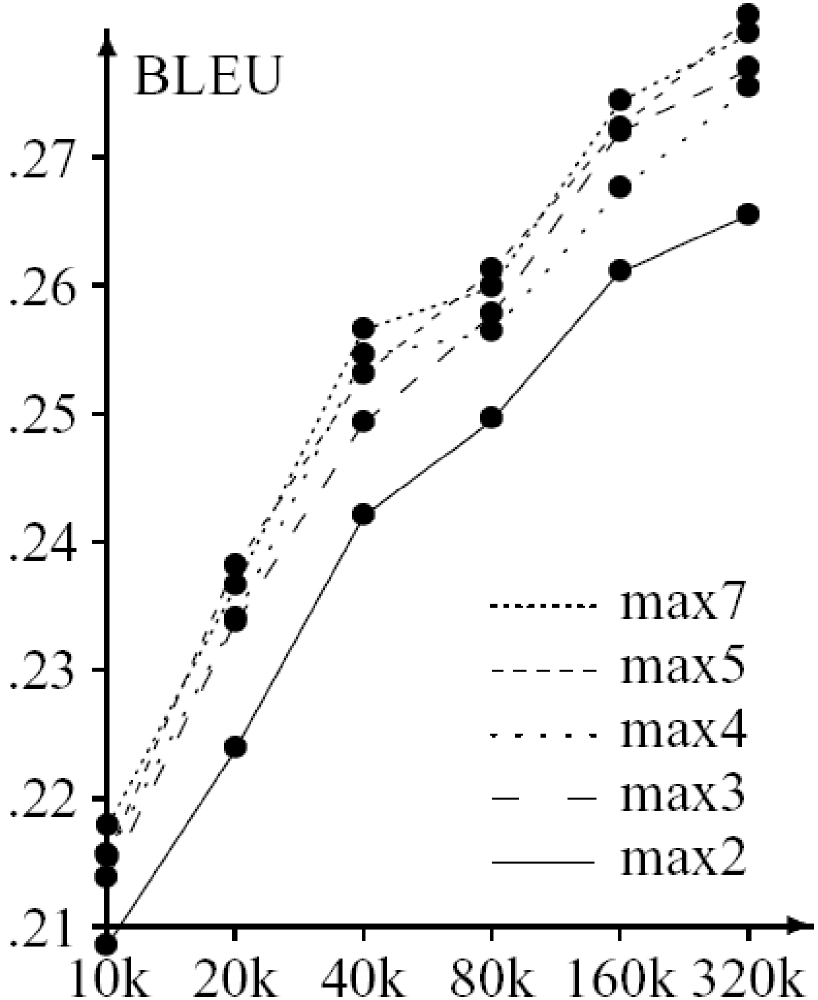
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)



Phrase Size

- Phrases do help
  - But they don't need to be long
  - Why should this be?





Why not Learn Phrases w/ EM?

EM Training of the Phrase Model

- We presented a heuristic set-up to build phrase translation table (word alignment, phrase extraction, phrase scoring)

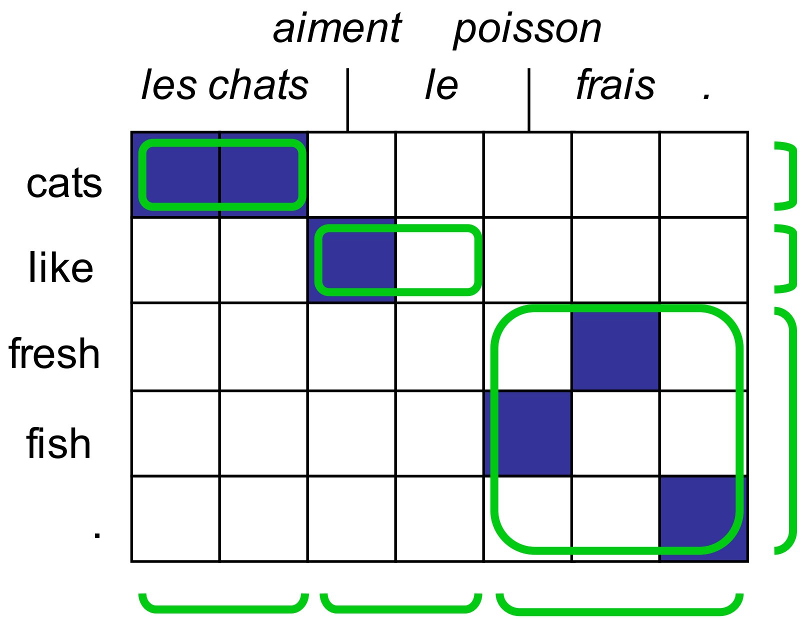
- Alternative: align phrase pairs directly with EM algorithm
  - initialization: uniform model, all $\phi(\bar{e}, \bar{f})$ are the same
  - expectation step:
    - estimate likelihood of all possible phrase alignments for all sentence pairs
  - maximization step:
    - collect counts for phrase pairs $(\bar{e}, \bar{f})$, weighted by alignment probability
    - update phrase translation probabilties $p(\bar{e}, \bar{f})$

- However: method easily overfits (learns very large phrase pairs, spanning entire sentences)

Phrase Scoring

$$g(f, e) = \log \frac{c(e, f)}{c(e)}$$

$$g(\text{les chats}, \text{cats}) = \log \frac{c(\text{cats}, \text{les chats})}{c(\text{cats})}$$



- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - … and others
- Seems not to work well, for a variety of partially understood reasons
- Main issue: big chunks get all the weight, obvious priors don’t help
  - Though, [DeNero et al 08]

Part III - Decoding

Phrase-Based Translation

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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.

Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

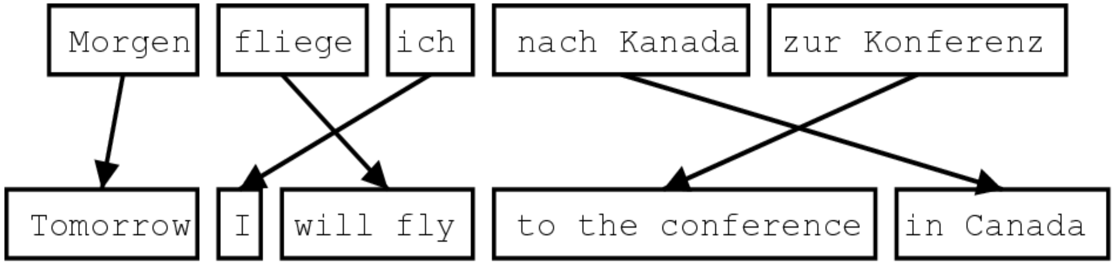
Phrase-Based Translation

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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Scoring:



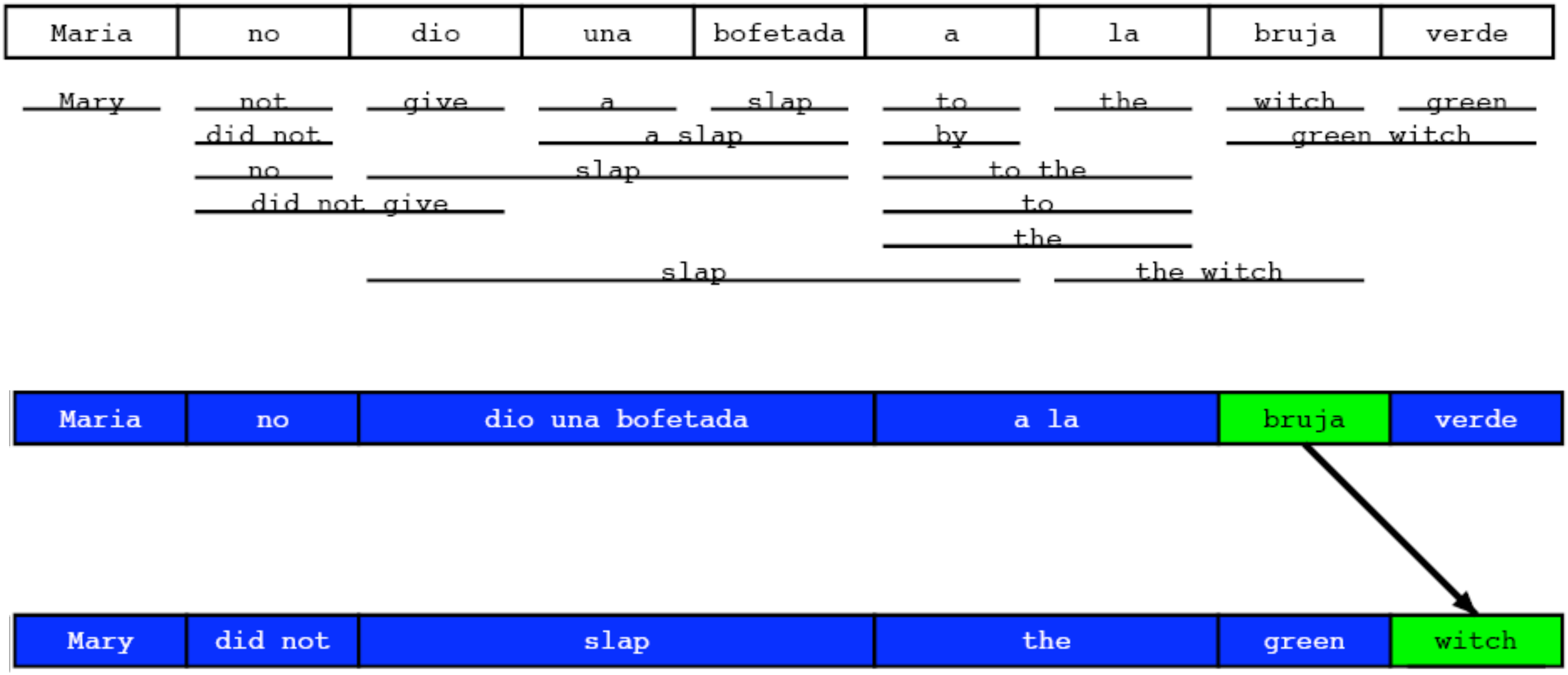
- Basic approach, sum up phrase translation scores and a language model
  - Define $y = p_1 p_2 \ldots p_L$ to be a translation with phrase pairs p_i
  - Define e(y) be the output English sentence in y
  - Let h() be the log probability under a tri-gram language model
  - Let g() be a phrase pair score (from last slide)
  - Then, the full translation score is:

$$f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k)$$

- Goal, compute the best translation

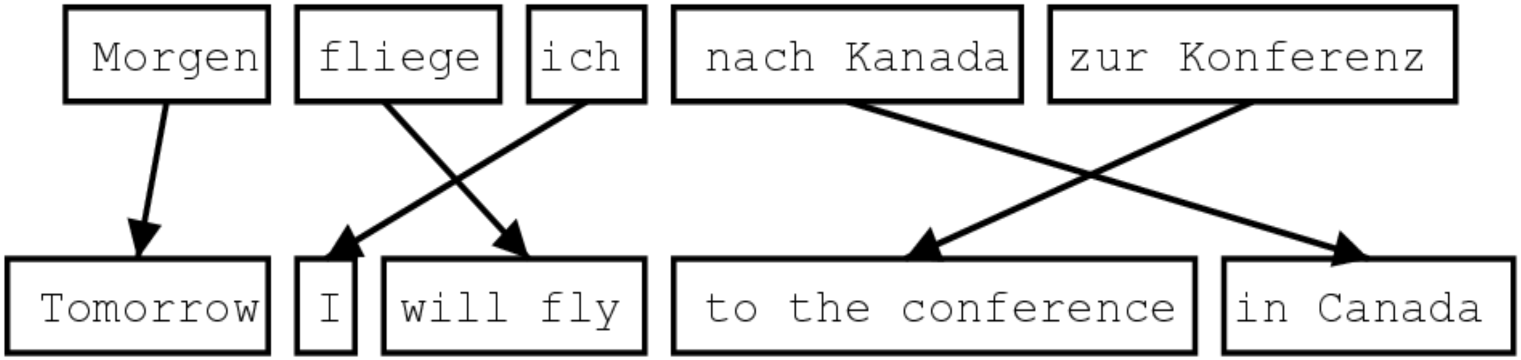
$$y^*(x) = \arg\max_{y \in \mathcal{Y}(x)} f(y)$$

The Pharaoh Decoder



- Scores at each step include LM and TM

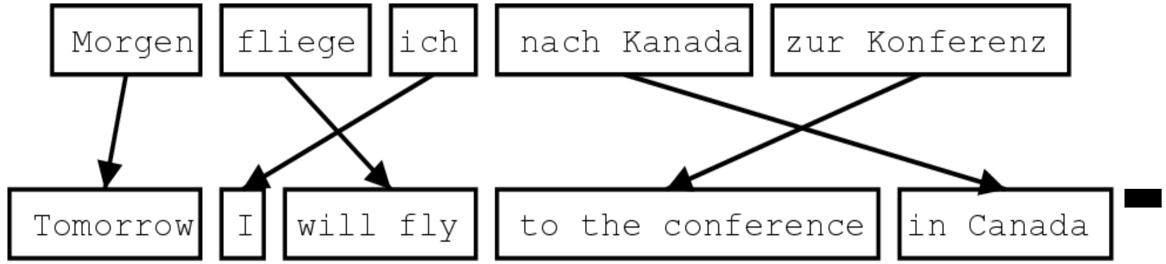
The Pharaoh Decoder



Space of possible translations

- Phrase table constrains possible translations
- Output sentence is built left to right
  - but source phrases can match any part of sentence
- Each source word can only be translated once
- Each source word must be translated

Scoring:



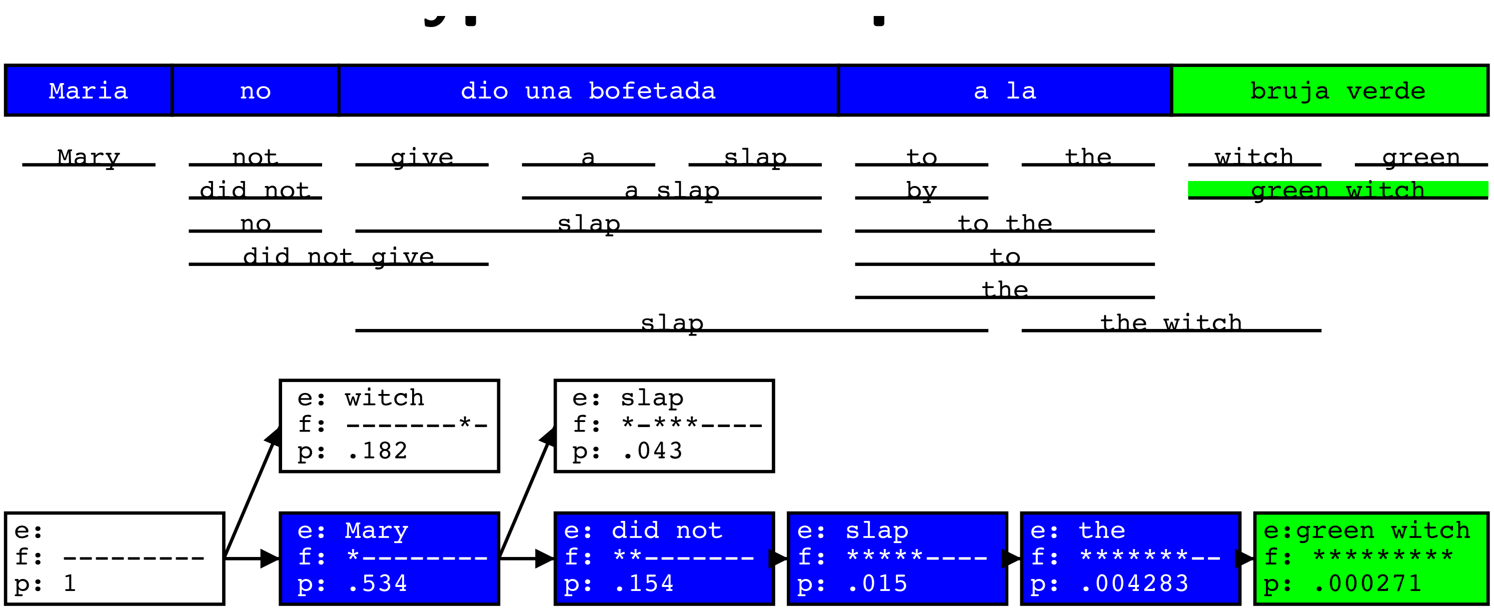
- In practice, much like for alignment models, also include a distortion penalty
  - Define y = $p_1 p_2 \ldots p_L$ to be a translation with phrase pairs p_i
  - Let $s(p_i)$ be the start position of the foreign phrase
  - Let $t(p_i)$ be the end position of the foreign phrase
  - Define η to be the distortion score (usually negative!)
  - Then, we can define a score *with distortion penalty*:

$$f(y) = h(e(y)) + \sum_{k=1}^{L} g(p_k) + \sum_{k=1}^{L-1} \eta \times |t(p_k) + 1 - s(p_{k+1})|$$

- Goal, compute the best translation

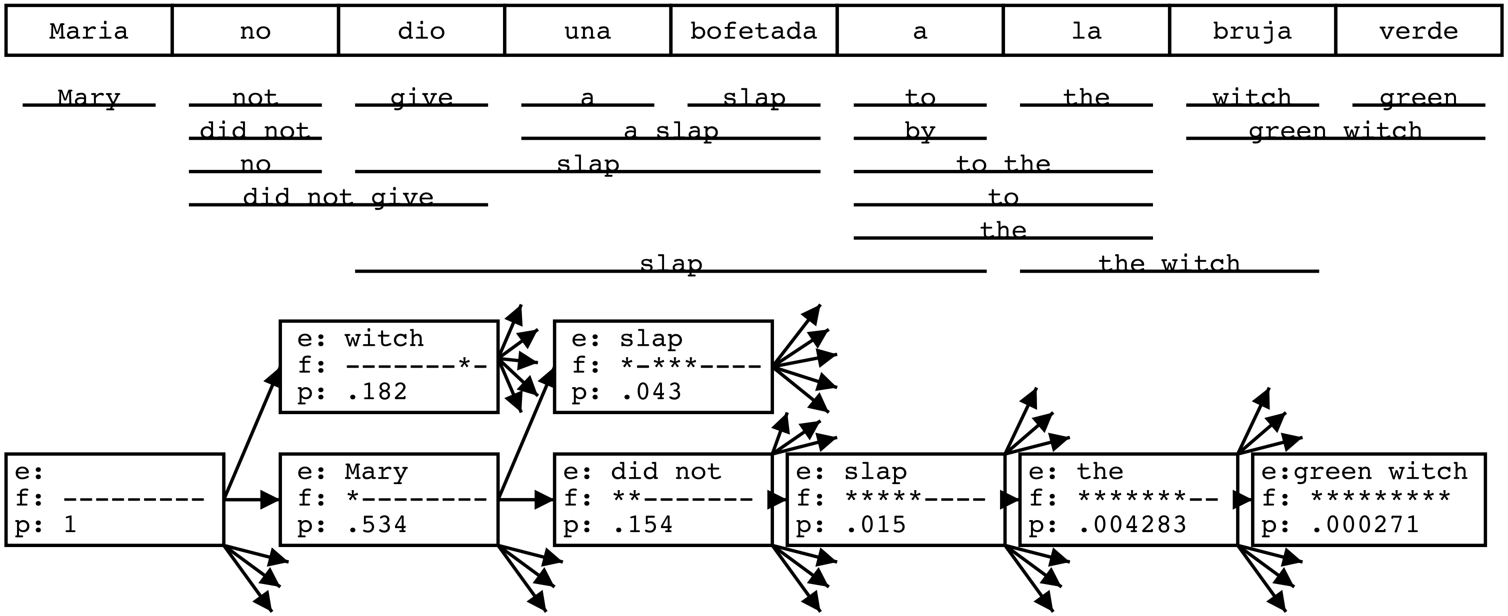
$$y^*(x) = \arg \max_{y \in \mathcal{Y}(x)} f(y)$$

Hypothesis Expansion



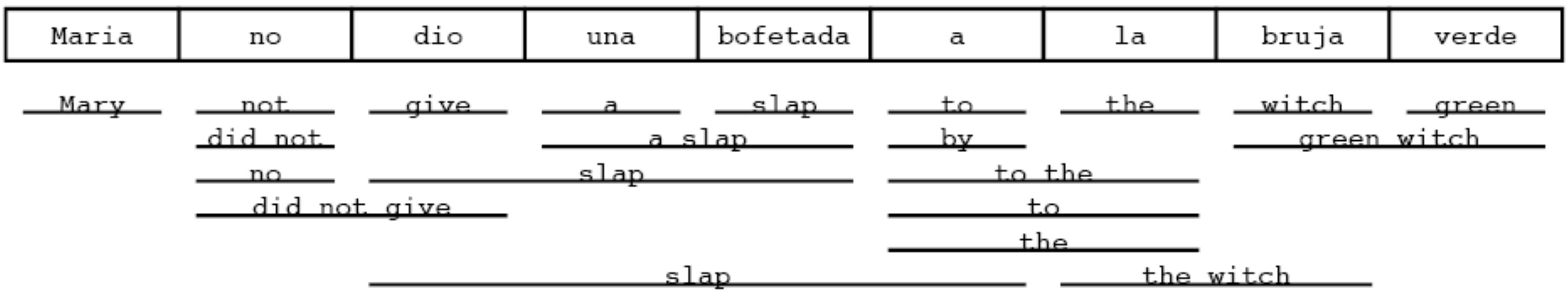
- ... until all foreign words *covered*
  - find *best hypothesis* that covers all foreign words
  - *backtrack* to read off translation

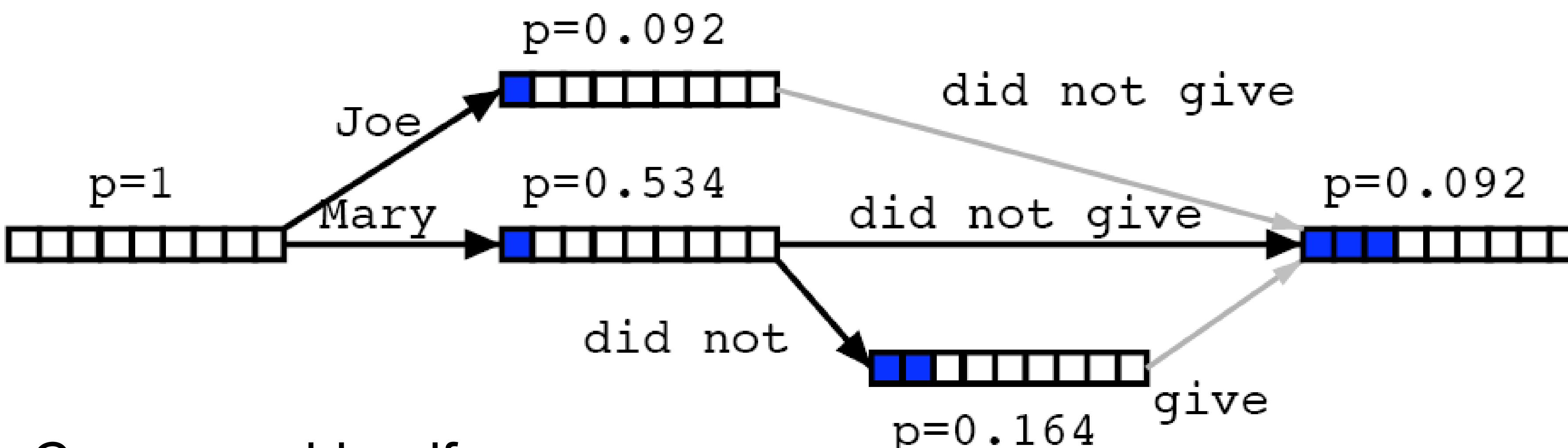
Hypothesis Explosion!



- Q: How much time to find the best translation?
  - Exponentially many translations, in length of source sentence
  - NP-hard, just like for word translation models
  - So, we will use approximate search techniques!

Hypothesis Lattices





Can recombine if:

- Last two English words match
- Foreign word coverage vectors match

Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where q_0 is initial state with no words translated

For i=0 … n-1 [where n in input sentence length]

•For each state $q \in beam(Q)$ and phrase $p \in ph(q)$

1. q'=next(q,p) [compute the new state]
2. Add(Q,q',q,p) [add the new state to the beam]

Notes:

•ph(q): set of phrases that can be added to partial translation in state q

•next(q,p): updates the translation in q and records which words have been translated from input

•Add(Q,q',q,p): updates beam, q' is added to Q if it is in the top-n overall highest scoring partial translations

Decoder Pseudocode

Initialization: Set beam $Q=\{q_0\}$ where q_0 is initial state with no words translated

For i=0 … n-1 [where n in input sentence length]

- For each state $q \in \text{beam}(Q)$ and phrase $p \in \text{ph}(q)$
  1. q'=next(q,p) [compute the new state]
  2. Add(Q,q',q,p) [add the new state to the beam]

Possible State Representations:

- Full: q = (e, b, α), e.g. ("Joe did not give," 11000000, 0.092)
  - e is the partial English sentence
  - b is a bit vector recorded which source words are translated
  - α is score of translation so far

Decoder Pseudocode

Initialization: Set beam Q={q_0} where q_0 is initial state with no words translated

For i=0 … n-1 [where n in input sentence length]

•For each state q∈beam(Q) and phrase p∈ph(q)

1. q'=next(q,p) [compute the new state]
2. Add(Q,q',q,p) [add the new state to the beam]

Possible State Representations:

•Full: q = (e, b, α), e.g. ("Joe did not give," 11000000, 0.092)

•Compact: q = (e_1, e_2, b, r, α) ,

- e.g. ("not," "give," 11000000, 4, 0.092)
- e_1 and e_2 are the last two words of partial translation
- r is the length of the partial translation

•Compact representation is more efficient, but requires back pointers to get the final translation