CSP 517
Natural Language Processing
Winter 2015

Machine Translation: Word Alignment

Yejin Choi

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

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:

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

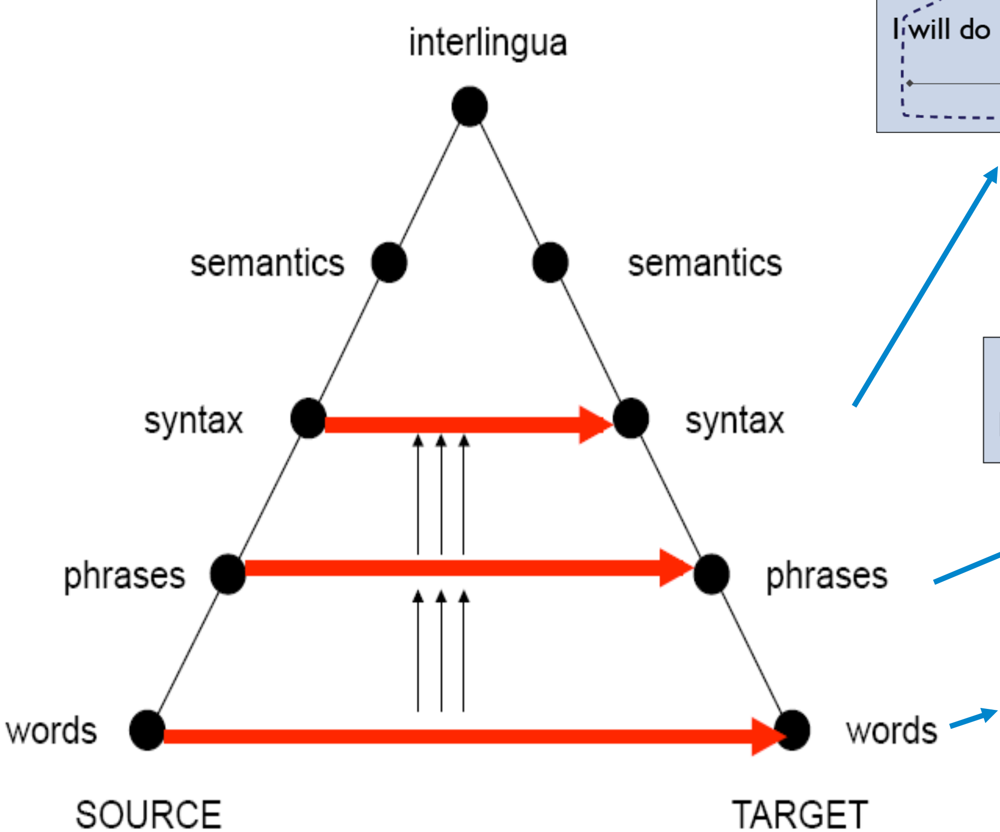
I will do it soon

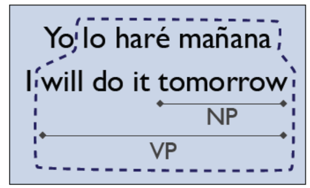
I will do it around

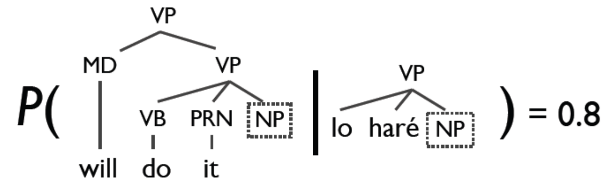
See you tomorrow

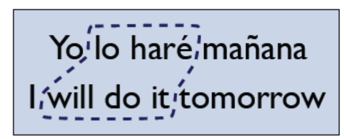
Levels of Transfer

"Vauquois Triangle"

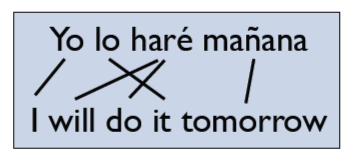








English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2



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

World-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)

Lexical Divergences

- Word to phrases:
  - English `computer science`
  - French `informatique`

- Part of Speech divergences
  - English `She likes to sing`
  - German `Sie singt gerne` [She sings likefully]
  - English `I'm hungry`
  - Spanish `Tengo hambre` [I have hunger]

Examples from Dan Jurafsky

Lexical Divergences: Semantic Specificity

English `brother`

Mandarin `gege` (older brother), `didi` (younger brother)

English `wall`

German `Wand` (inside) `Mauer` (outside)

English `fish`

Spanish `pez` (the creature) `pescado` (fish as food)

Cantonese `ngau`

English `cow` `beef`

Examples from Dan Jurafsky

Predicate Argument divergences

L. Talmy. 1985. Lexicalization patterns: Semantic Structure in Lexical Form.

- English
  The bottle **floated** out.
- Spanish
  La botella salió **flotando**.
  The bottle exited **floating**
- **Satellite-framed languages**:
  - direction of motion is marked on the satellite
  - `Crawl out, float off, jump down, walk over to, run after`
  - Most of Indo-European, Hungarian, Finnish, Chinese
- **Verb-framed languages**:
  - direction of motion is marked on the verb
  - Spanish, French, Arabic, Hebrew, Japanese, Tamil, Polynesian, Mayan, Bantu families

Examples from Dan Jurafsky

Predicate Argument divergences: Heads and Argument swapping

Dorr, Bonnie J., "Machine Translation Divergences: A Formal Description and Proposed Solution," Computational Linguistics, 20:4, 597--633

Heads:

English: X swim across Y

Spanish: X crucar Y nadando

English: I **like** to *eat*

German: Ich *esse* **gern**

English: I'd prefer vanilla

German: Mir wäre Vanille lieber

Arguments:

Spanish: **Y** me gusta

English: I like **Y**

German: **Der Termin** fällt mir ein

English: I forget **the date**

Examples from Dan Jurafsky

Predicate-Argument Divergence Counts

B.Dorr et al. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment

Found divergences in 32% of sentences in UN Spanish/English Corpus

Part of Speech	X tener hambre Y have hunger	98%
Phrase/Light verb	X dar puñaladas a Z X stab Z	83%
Structural	X entrar en Y X enter Y	35%
Heads swap	X cruzar Y nadando X swim across Y	8%
Arguments swap	X gustar a Y Y likes X	6%

Examples from Dan Jurafsky

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)

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…

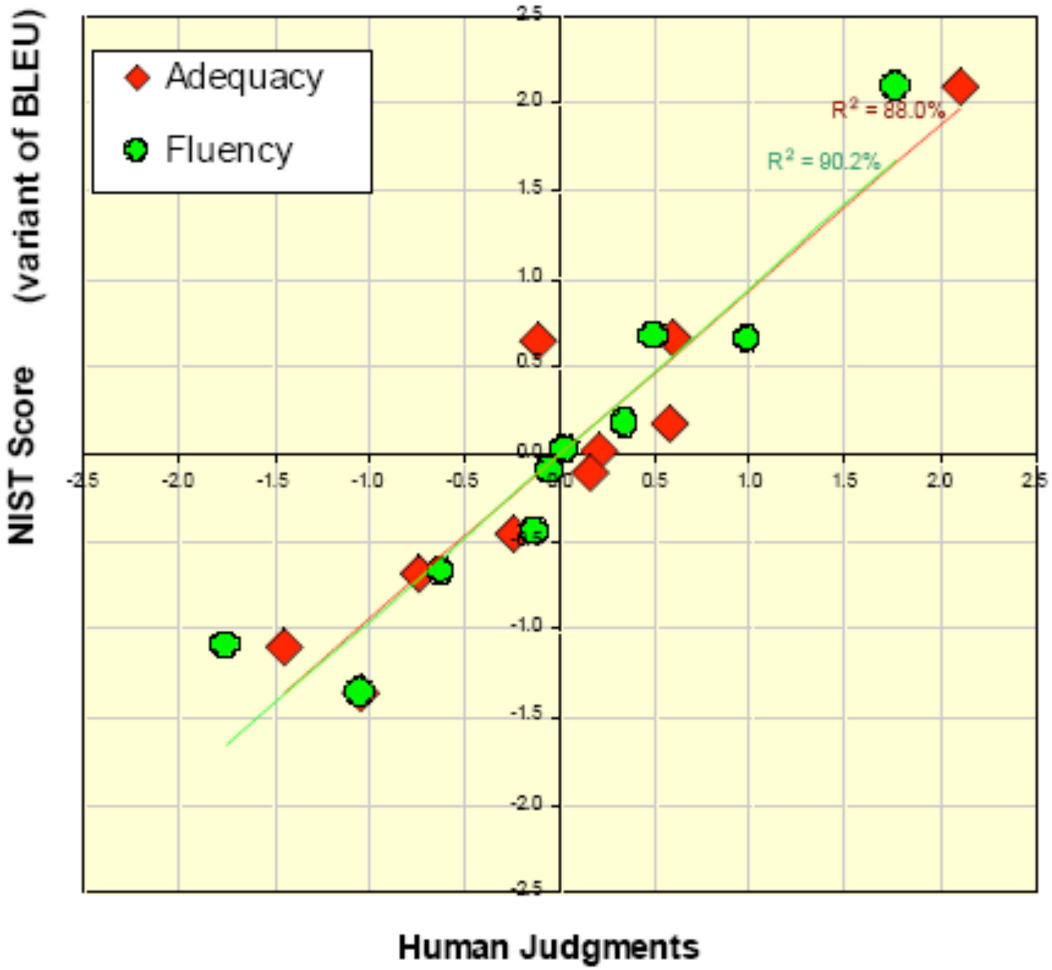
Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:

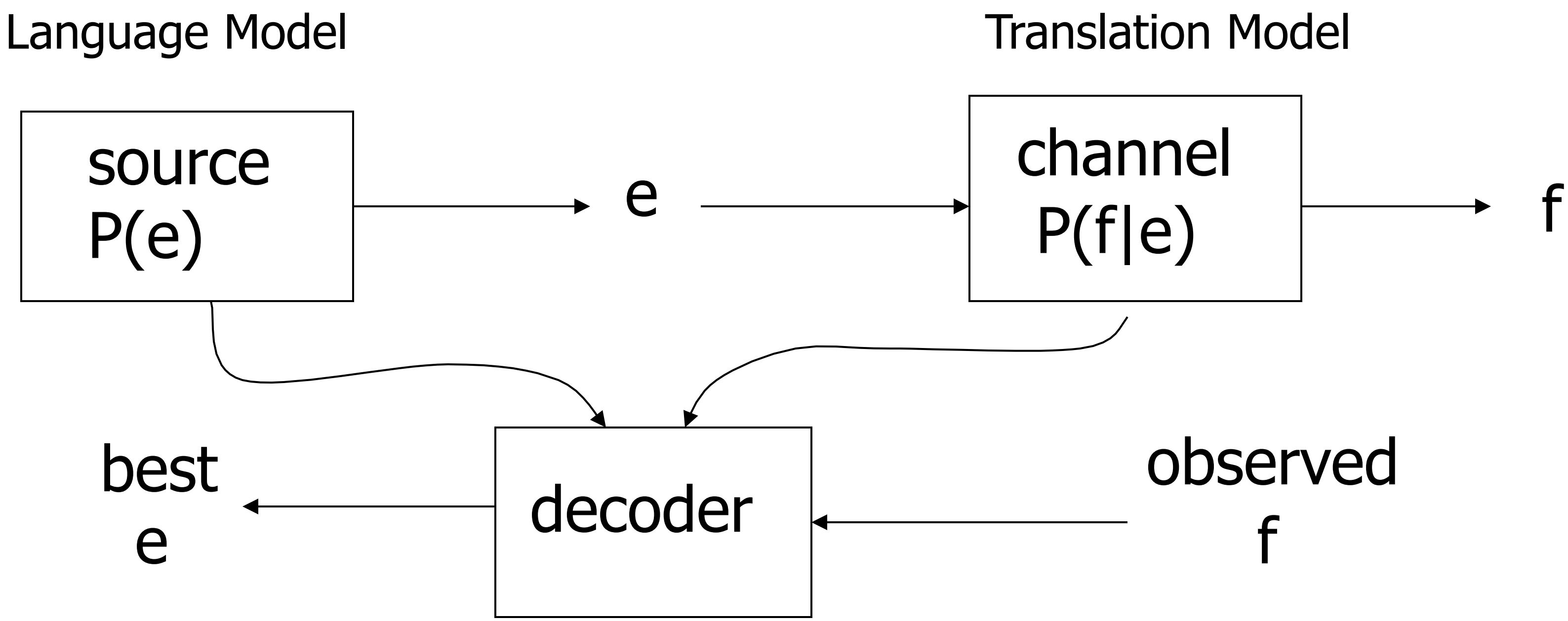
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Automatic Metrics Work (?)



slide from G. Doddington (NIST)

MT System Components



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

Today

- The components of a simple MT system
  - You already know about the LM
  - Word-alignment based TMs
    - IBM models 1 and 2, HMM model
  - A simple decoder

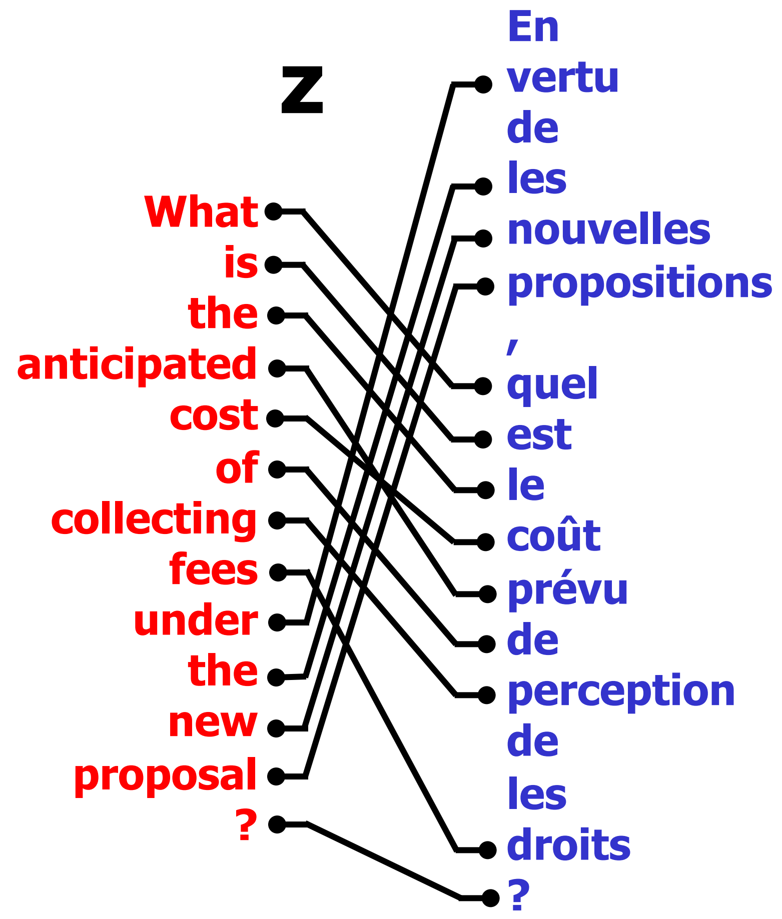
- Next few classes
  - More complex word-level and phrase-level TMs
  - Tree-to-tree and tree-to-string TMs
  - More sophisticated decoders

Word Alignment

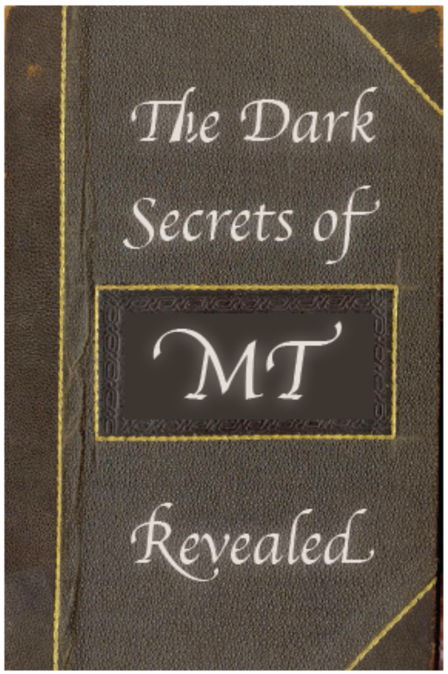
x

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

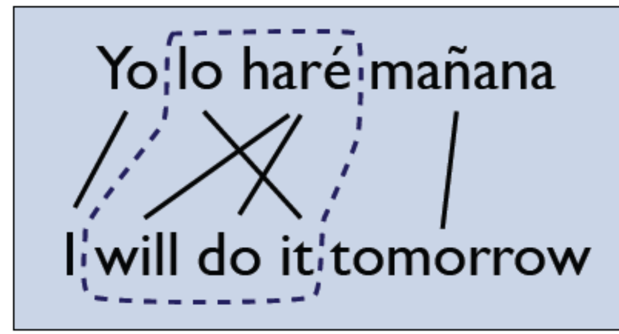
En vertu des nouvelles propositions, quel est le coût prévu de perception des droits?



Word Alignment



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



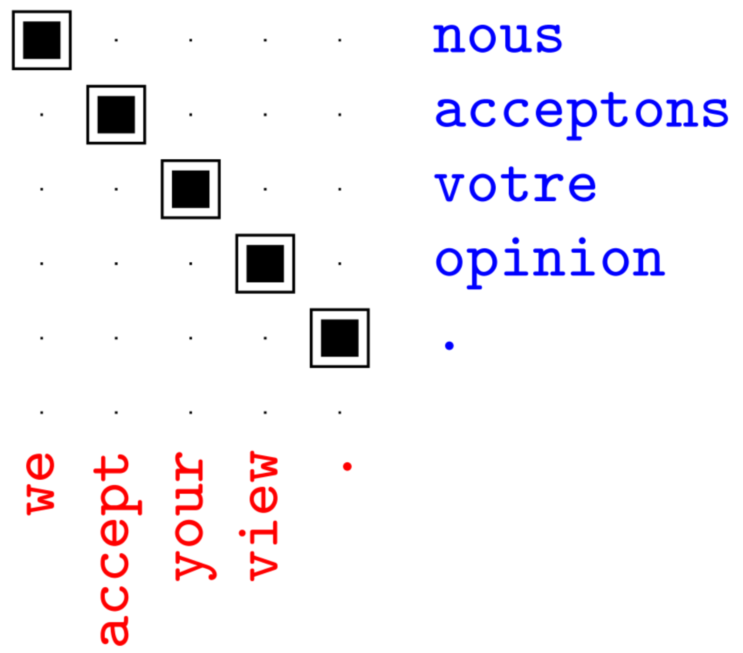
Unsupervised Word Alignment

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

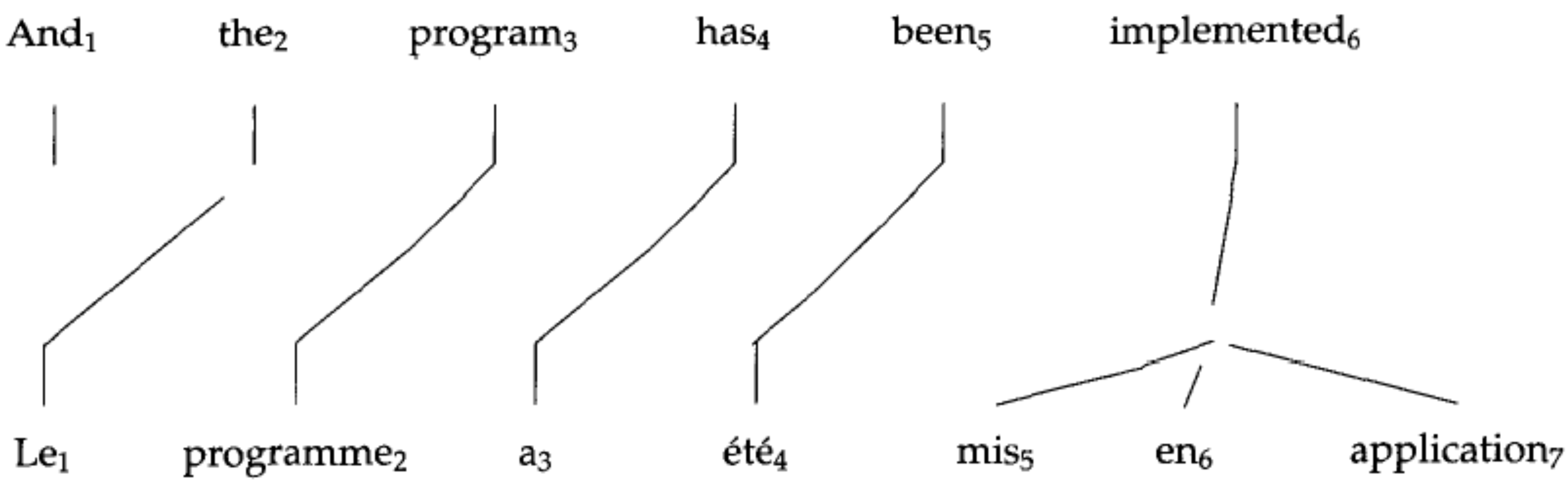
```
nous acceptons votre opinion .

we accept your view .
```

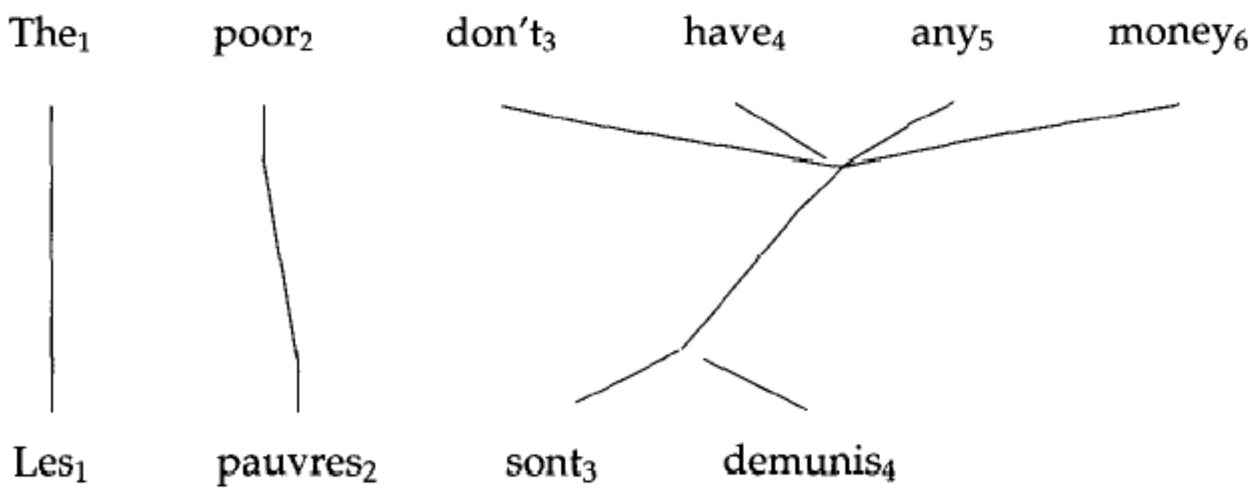
- Output: ***alignments***: pairs of translated words
  - When words have unique sources, can represent as a (forward) alignment function a from French to English positions



1-to-Many Alignments



Many-to-Many Alignments



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.

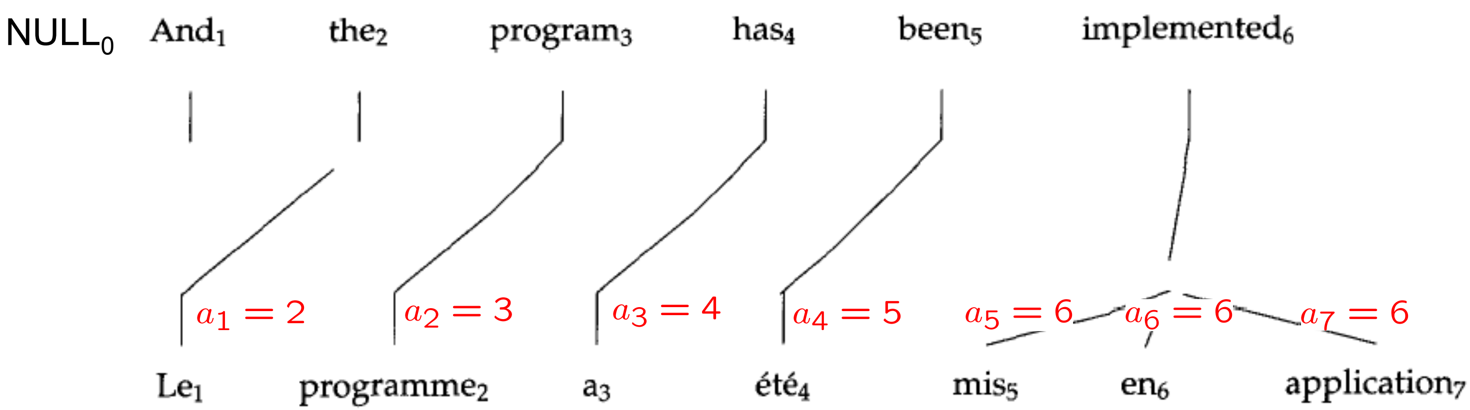
IBM Model 1 (Brown 93)

- 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)

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



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

Uniform alignment model!

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

IBM Model 1: Learning

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

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

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

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

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

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

Training Corpus

green house
casa verde

the house
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
1/3 X 1/3 = 1/9

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

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

the house
la casa
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	**green house**	**the house**	**the house**
casa verde	**casa verde**	**la casa**	**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 casa verde	green house casa verde	the house la casa	the house 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: Example

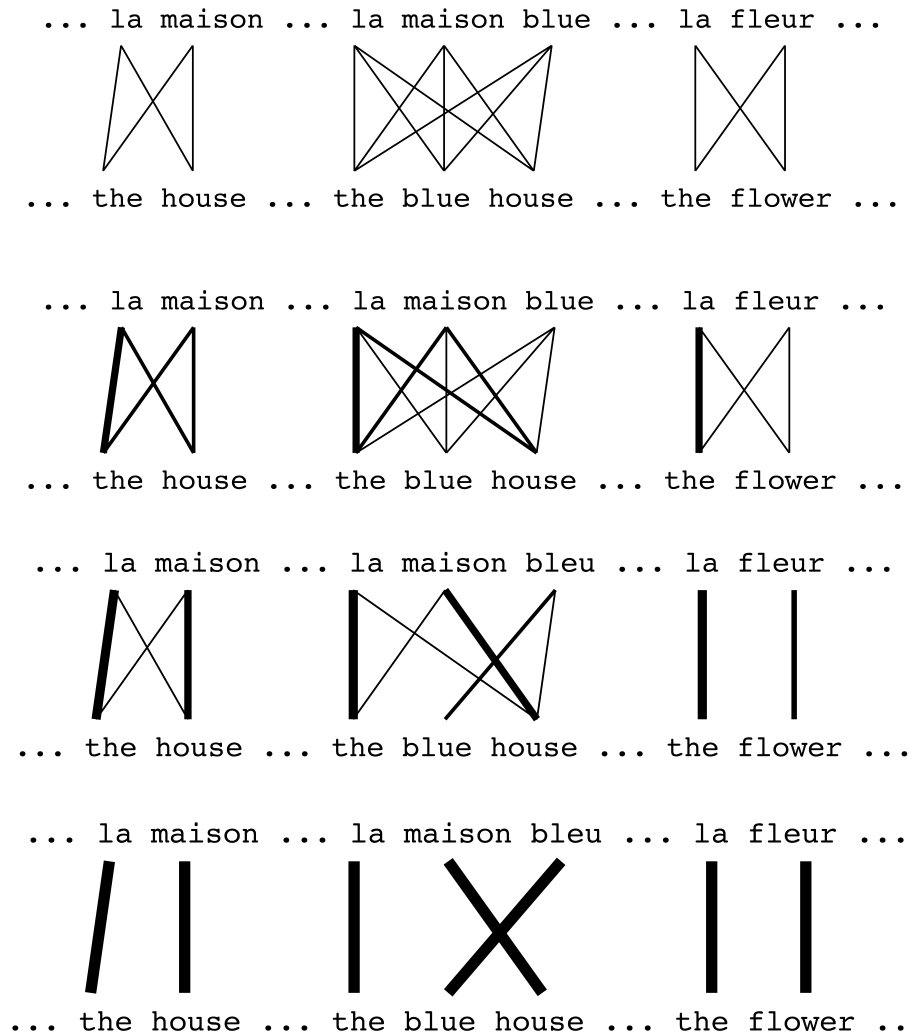
Step 1

Step 2

Step 3

…

Step N



Example from Philipp Koehn

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

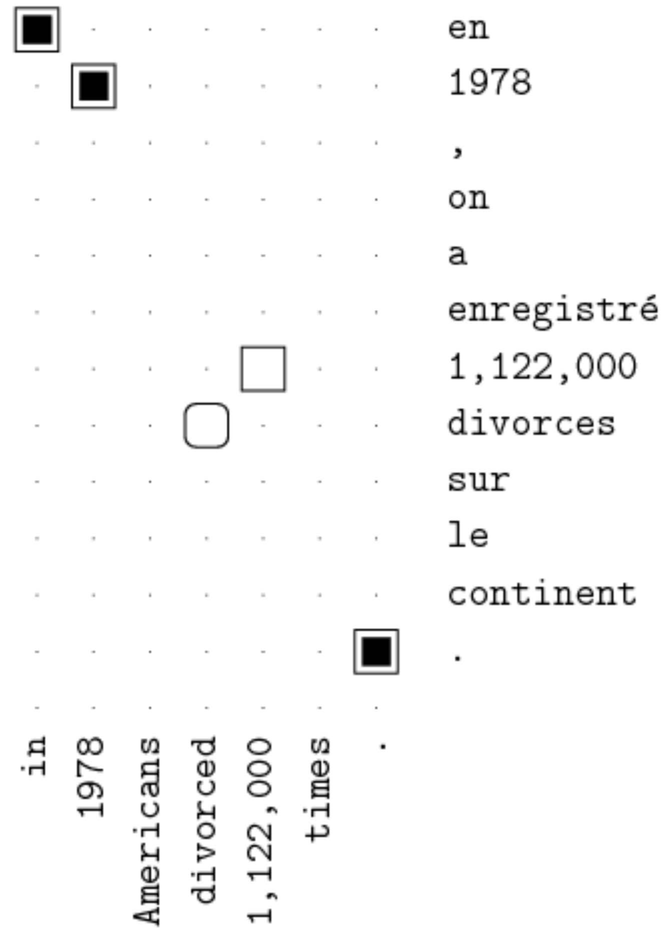
□ = Sure

▢ = Possible

■ = Predicted

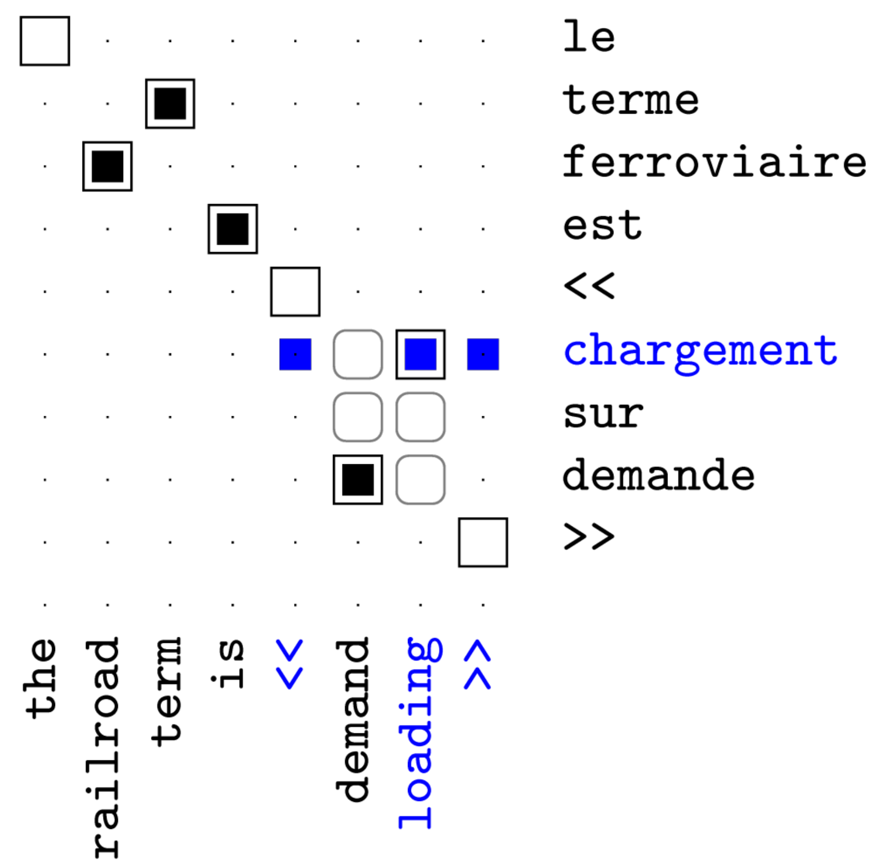
$$AER(A, S, P) = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right)$$

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



Problems with Model 1

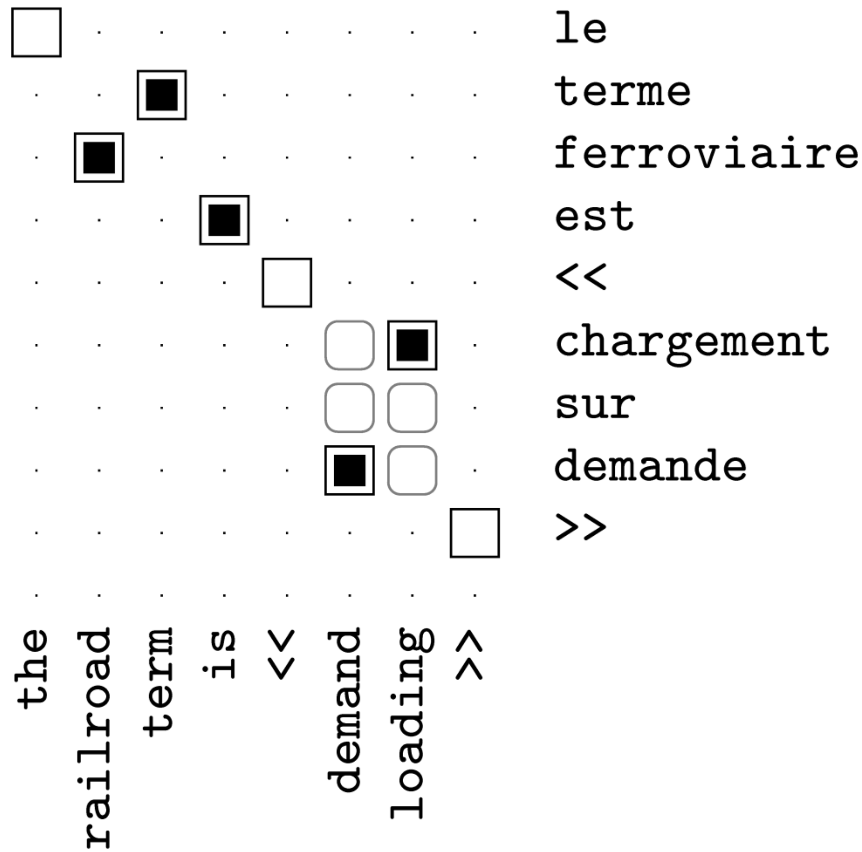
- There's a reason they designed models 2-5!
- Problems: alignments jump around, align everything to rare words
- Experimental setup:
  - Training data: 1.1M sentences of French-English text, Canadian Hansards
  - Evaluation metric: alignment error Rate (AER)
  - Evaluation data: 447 hand-aligned sentences



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?

- 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

Monotonic Translation

Japan shaken by two new quakes

Le Japon secoué par deux nouveaux séismes

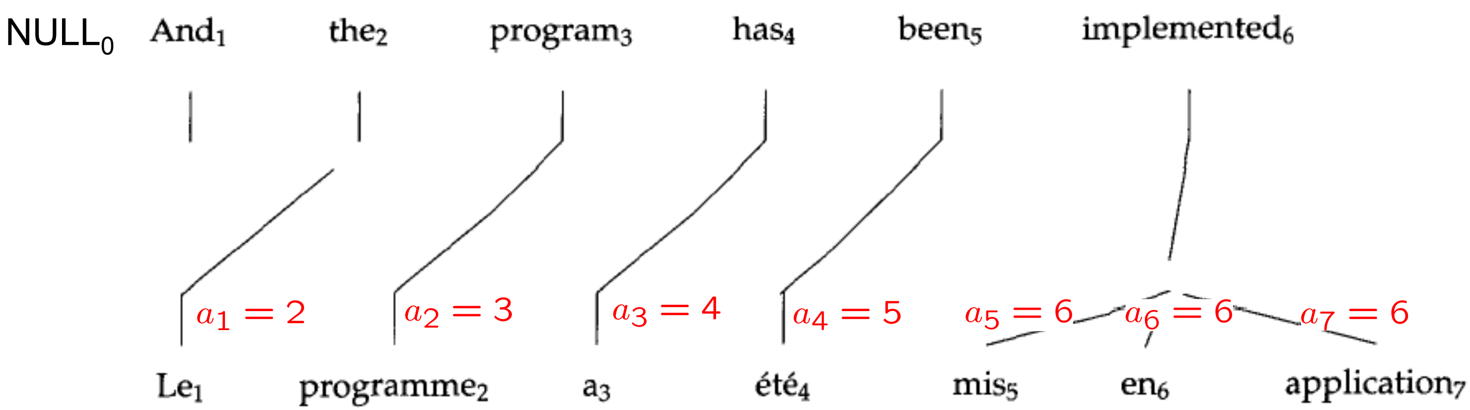
Local Order Change

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.



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

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

IBM Model 2: Learning

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

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

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

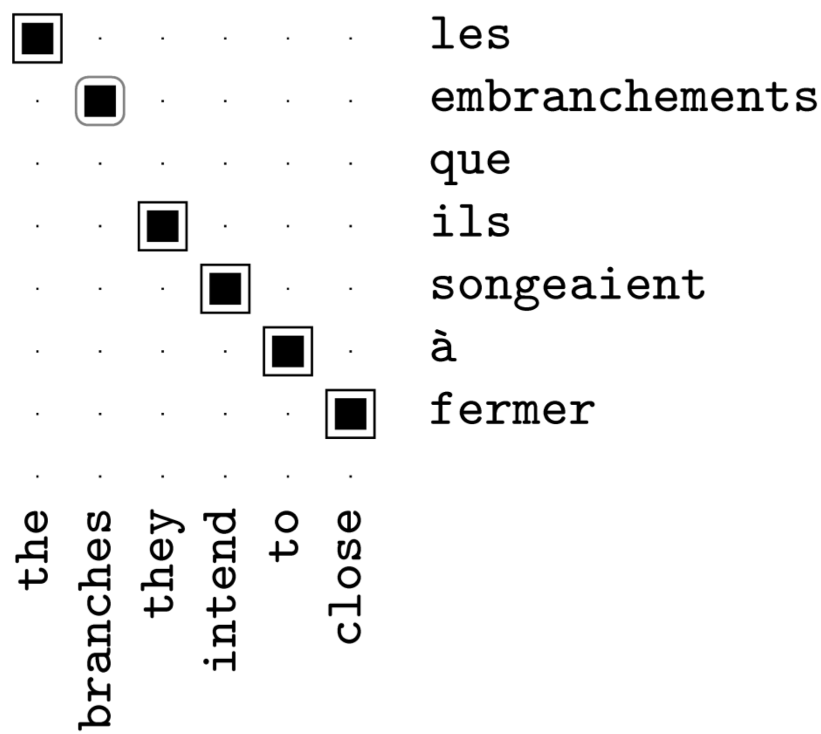
$$c(e,f) = \sum_k \sum_{i \text{ s.t. } e_i = e} \sum_{j \text{ s.t. } f_j = f} \delta(k,i,j)$$

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

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

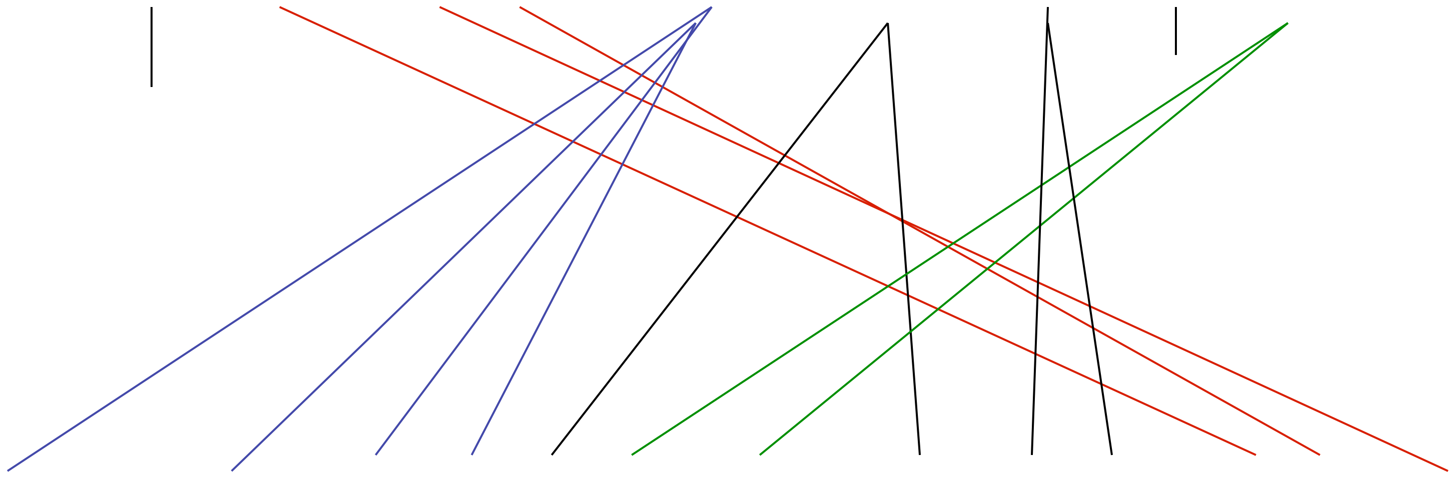
- Basic idea: compute expected source for each word, update co-occurrence statistics, repeat
- Q: What about inference? Is it hard?

Example


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

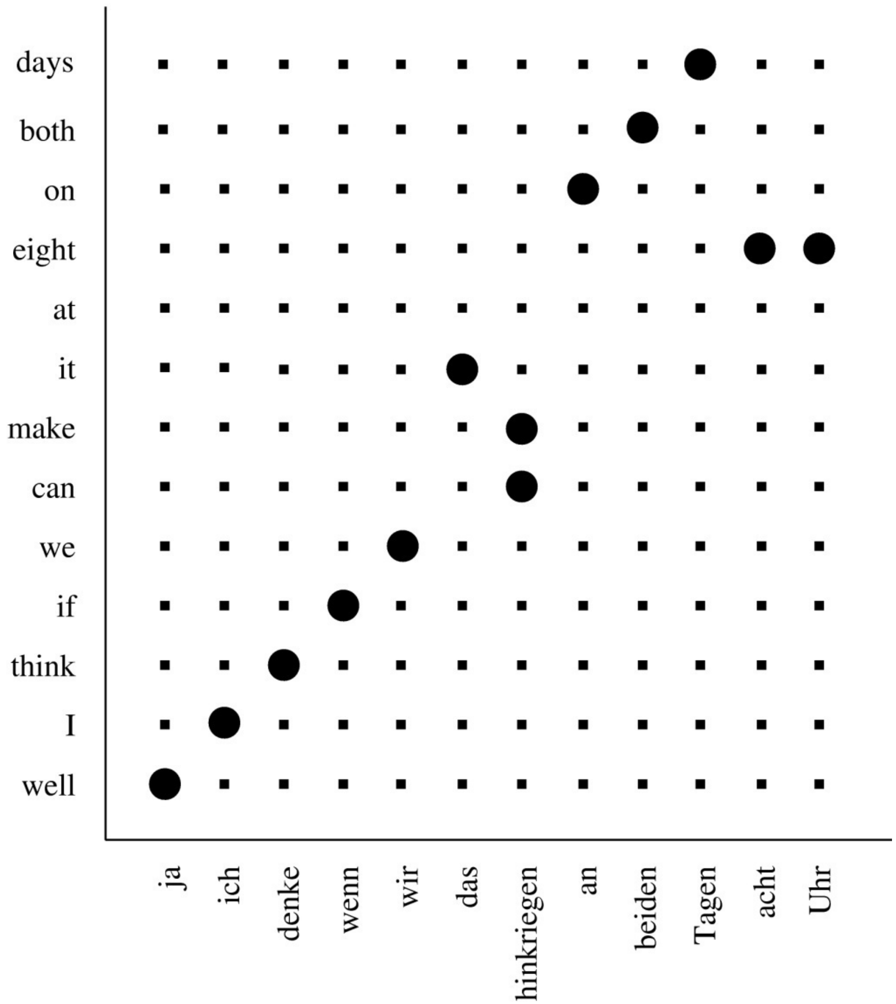
Phrase Movement

On Tuesday Nov. 4, earthquakes rocked Japan once again

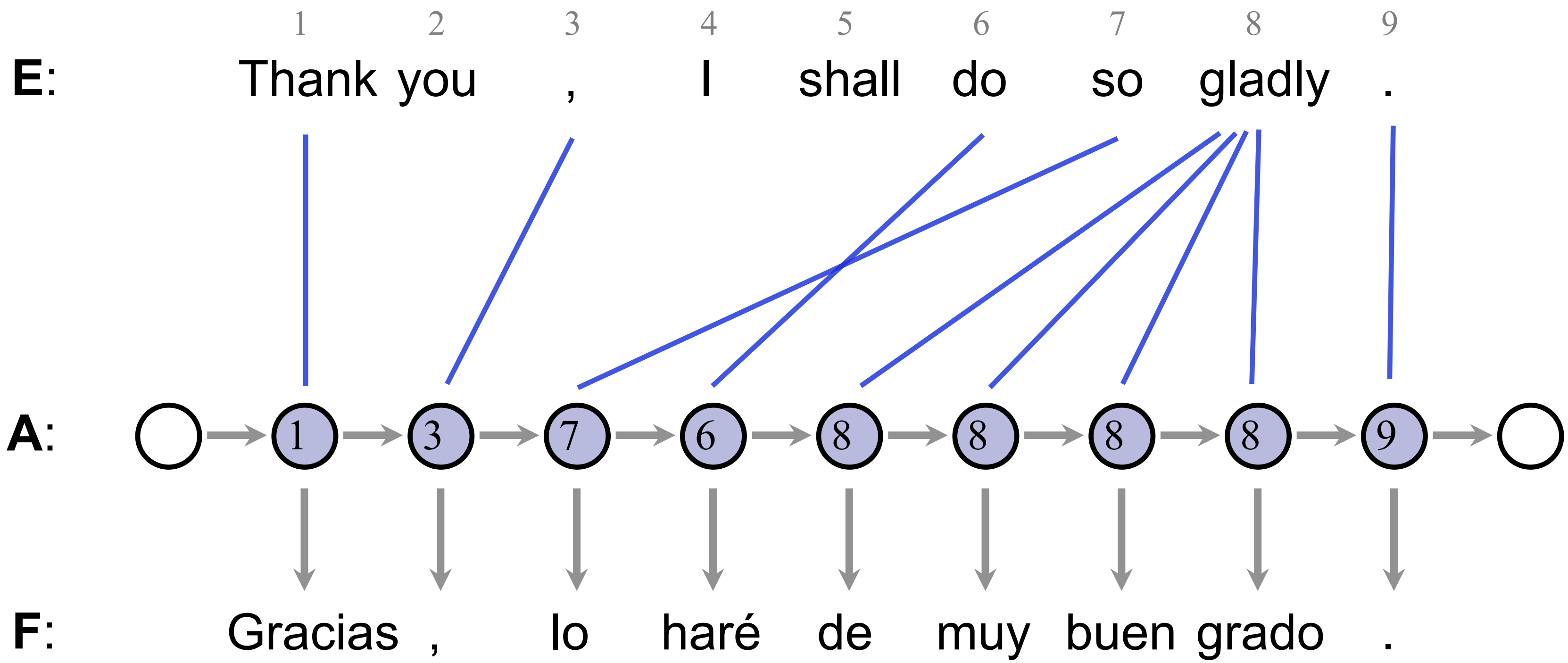


Des tremblements de terre ont à nouveau touché le Japon jeudi 4 novembre.

Phrase Movement



The HMM Model



Model Parameters

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

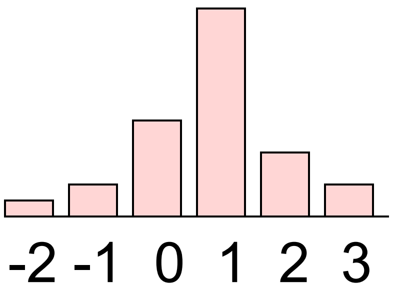
The HMM Model

- Model 2 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

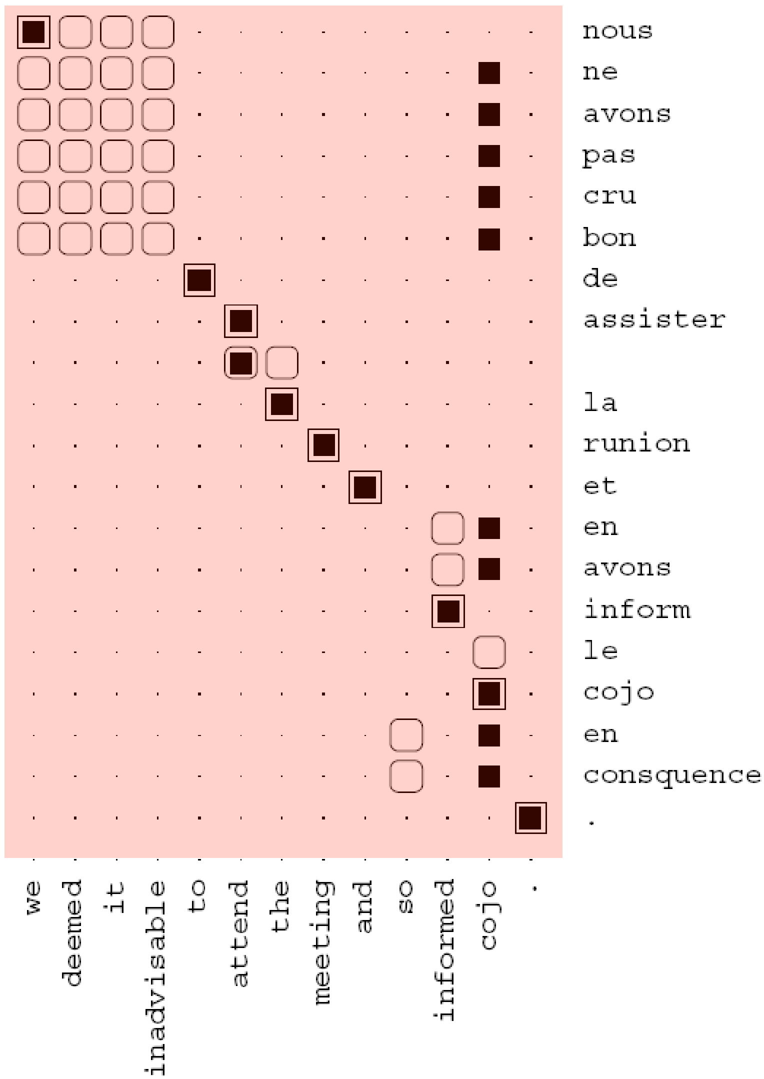
$$P(f, a|e) = \prod_j P(a_j|a_{j-1})P(f_j|e_i)$$

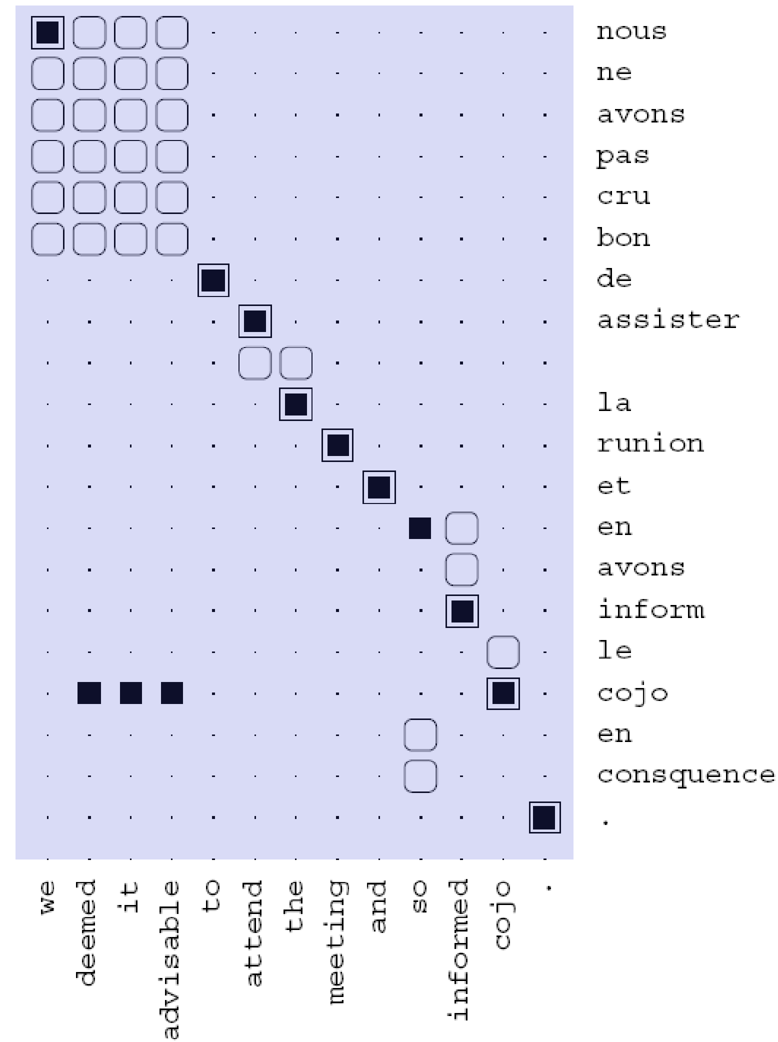
$$P(a_j - a_{j-1})$$



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

HMM Examples

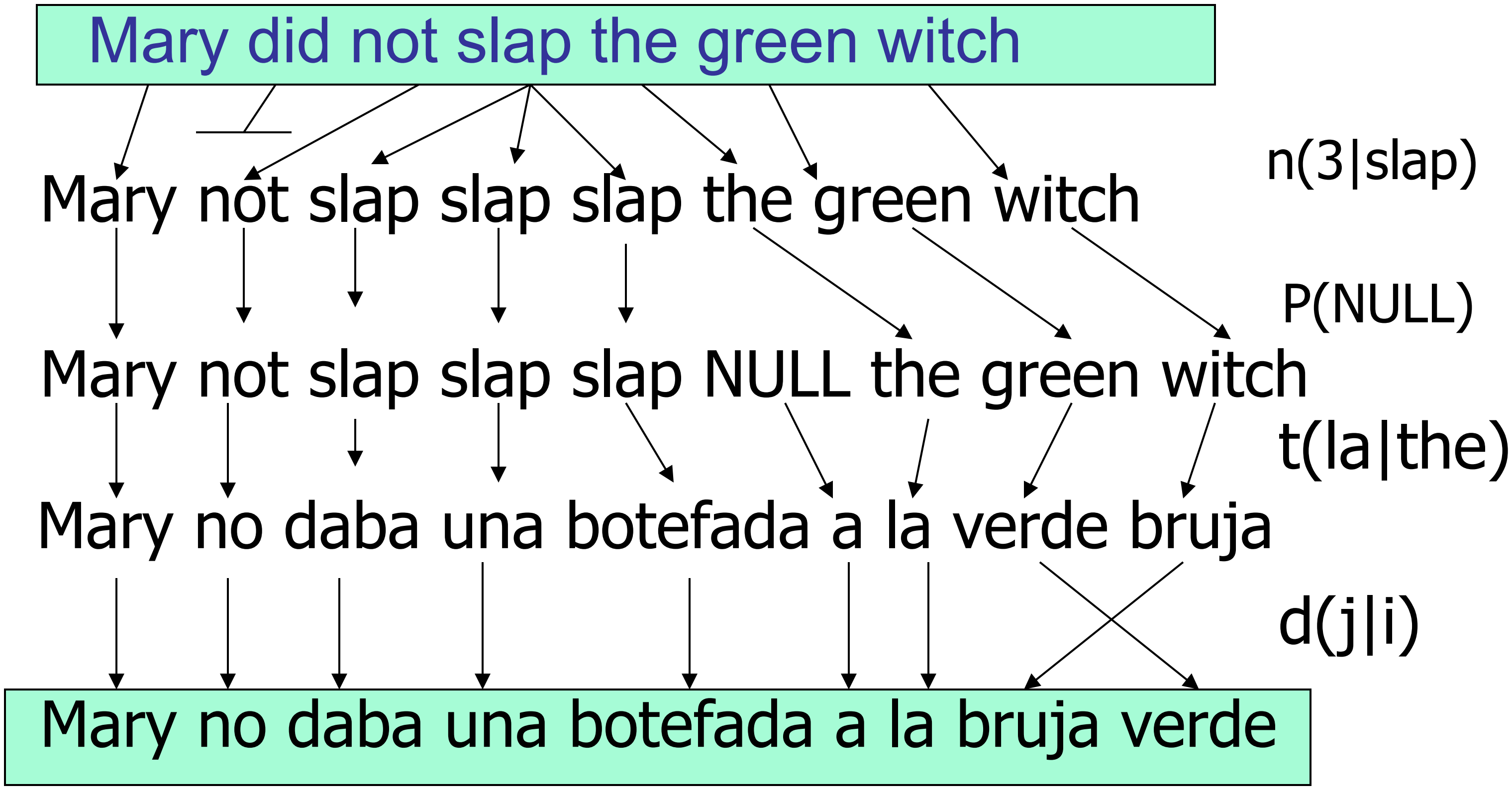




AER for HMMs

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

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		

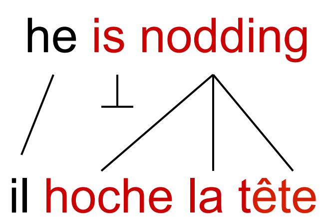
not

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

Decoding

- In these word-to-word models
  - Finding best alignments is easy
  - Finding translations is hard (why?)

```
it is not clear .
|  \    |    \    \
|   \   +     \    \
|    \/ \      \    \
|    /\  \      \    \
CE NE EST PAS CLAIR   .
```

Bag "Generation" (Decoding)

Exact reconstruction

Please give me your response as soon as possible.
$\Rightarrow$ Please give me your response as soon as possible.

Reconstruction preserving meaning

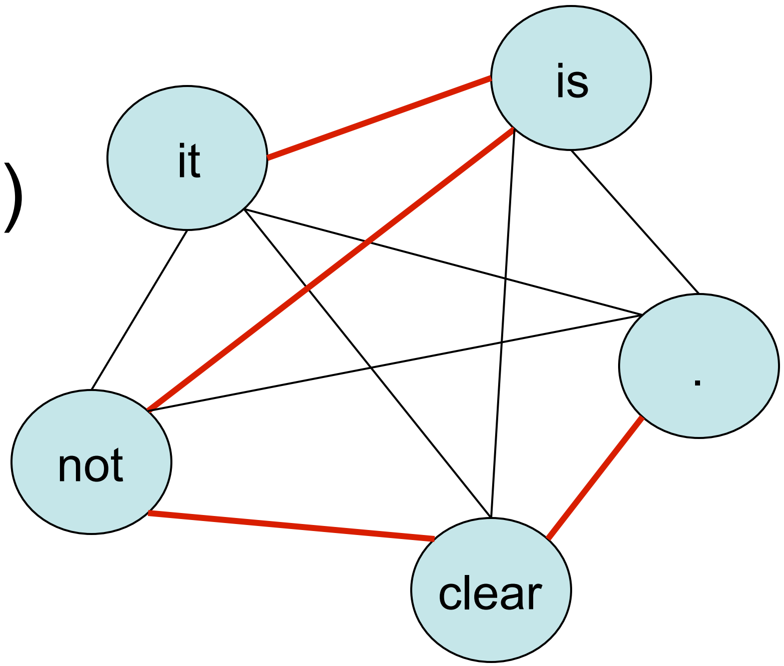
Now let me mention some of the disadvantages.
$\Rightarrow$ Let me mention some of the disadvantages now.

Garbage reconstruction

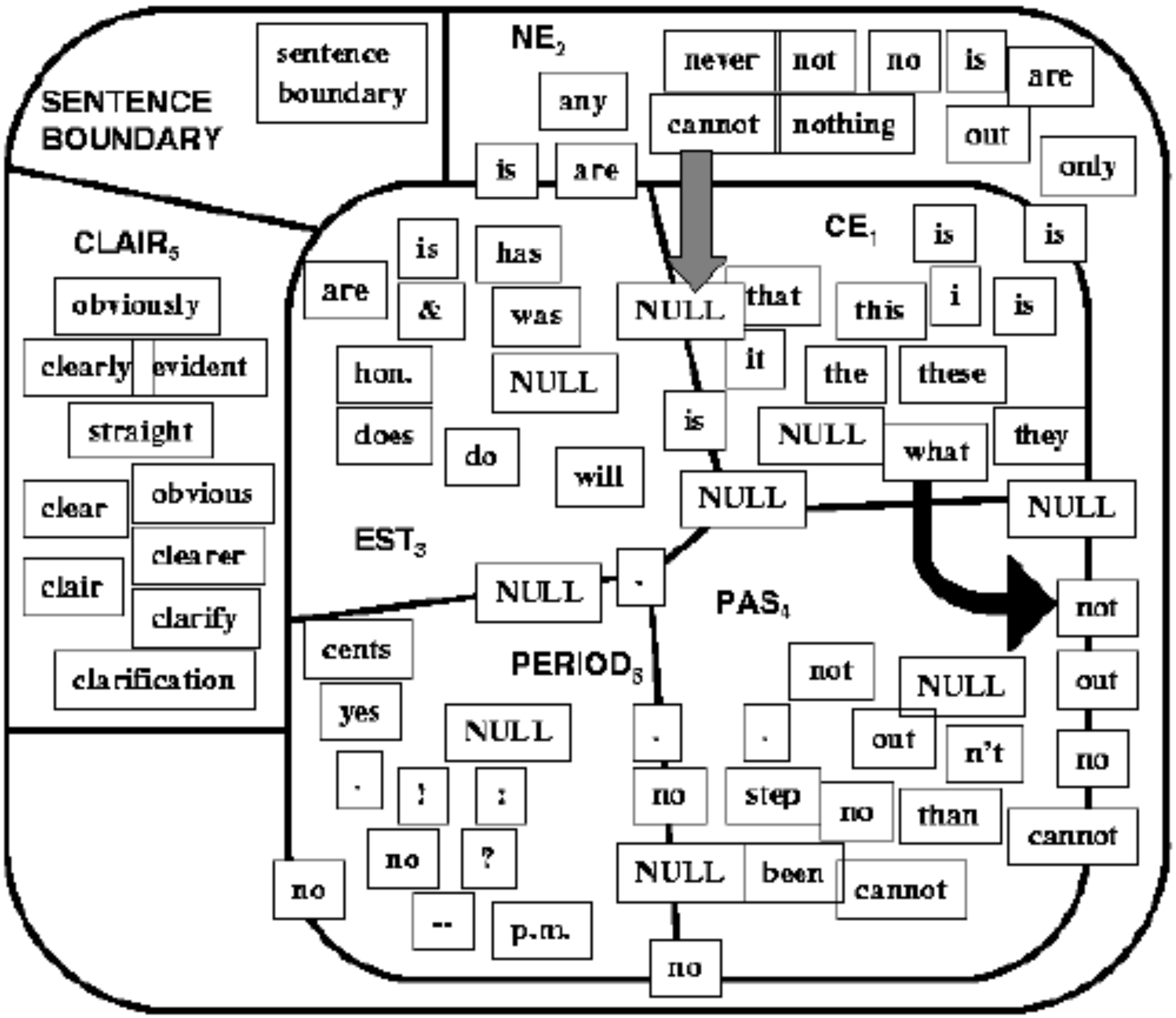
In our organization research has two missions.
$\Rightarrow$ In our missions research organization has two.

Bag Generation as a TSP

- Imagine bag generation with a bigram LM
  - Words are nodes
  - Edge weights are P(w|w')
  - Valid sentences are Hamiltonian paths
- Not the best news for word-based MT!



IBM Decoding as a TSP

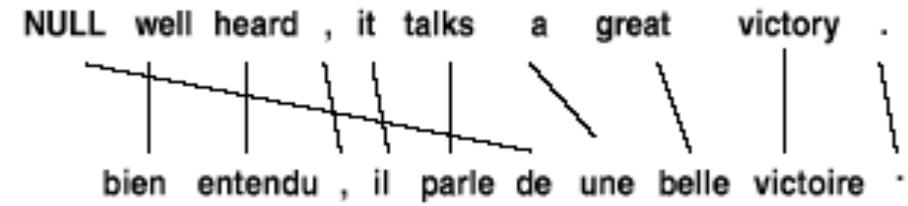


Decoding, Anyway

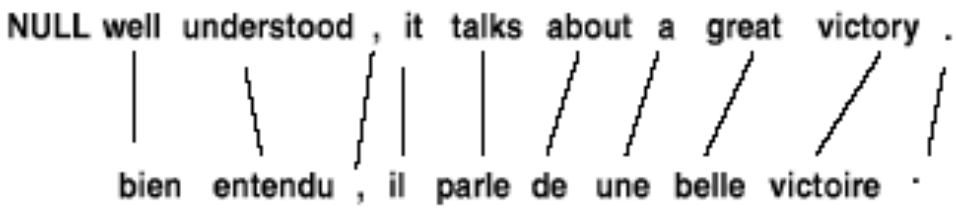
- Simplest possible decoder:
  - Enumerate sentences, score each with TM and LM

- Greedy decoding:
  - Assign each French word it’s most likely English translation
  - Operators:
    - Change a translation
    - Insert a word into the English (zero-fertile French)
    - Remove a word from the English (null-generated French)
    - Swap two adjacent English words
  - Do hill-climbing (or your favorite search technique)

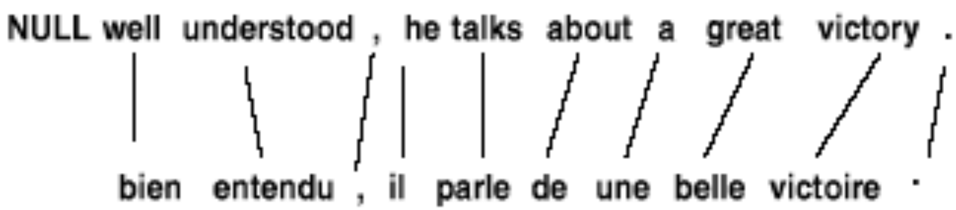
Greedy Decoding



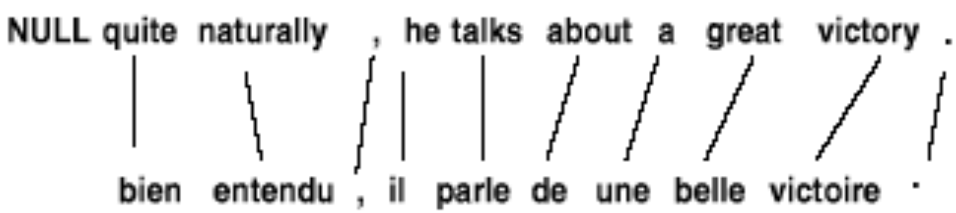
translateTwoWords(2,understood,0,about)



translateOneWord(4,he)



translateTwoWords(1,quite,2,naturally)



Stack Decoding

- Stack decoding:
  - Beam search
  - Usually A* estimates for completion cost
  - One stack per candidate sentence length
- Other methods:
  - Dynamic programming decoders possible if we make assumptions about the set of allowable permutations

sent length	decoder type	time (sec/sent)	search errors	translation errors (semantic and/or syntactic)	NE	PME	DSE	FSE	HSE	CE
6	IP	47.50	0	57	44	57	0	0	0	0
6	stack	0.79	5	58	43	53	1	0	0	4
6	greedy	0.07	18	60	38	45	5	2	1	10
8	IP	499.00	0	76	27	74	0	0	0	0
8	stack	5.67	20	75	24	57	1	2	2	15
8	greedy	2.66	43	75	20	38	4	5	1	33