Machine Translation: Word alignment models

Bill MacCartney

CS224N / Ling 284

[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky, and Chris Manning]

Let's start with some live translations!

Le Monde

読売新聞

What does a YouTube funny animal video have in common with machine translation?

1. Google now dominates in both areas

2. Sir, [fully automatic machine translation] is like a dog's walking on his hind legs. It is not done well; but you are surprised to find it done at all.

   [with apologies to Samuel Johnson]

OK, just one more fluffy diversion …

Word Lens

"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, March 1947

"… as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague … to make any quasi-mechanical translation scheme very hopeful."
– Norbert Wiener, April 1947

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat . process of elimination
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

cognate?

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

It’s Really Spanish/English

Clients do not sell pharmaceuticals in Europ **=>** **Clientes no venden medicinas en Europa**

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Statistical MT

Suppose we had a probabilistic model of translation
P(e | f)

Suppose f is *de rien*

P(*you're welcome* | *de rien*)	= 0.45
P(*nothing* | *de rien*)	= 0.13
P(*piddling* | *de rien*)	= 0.01
P(*underpants* | *de rien*)	= 0.000000001

A Bayesian approach

$$\hat{e} = \text{argmax}_e \, P(e \mid f)$$

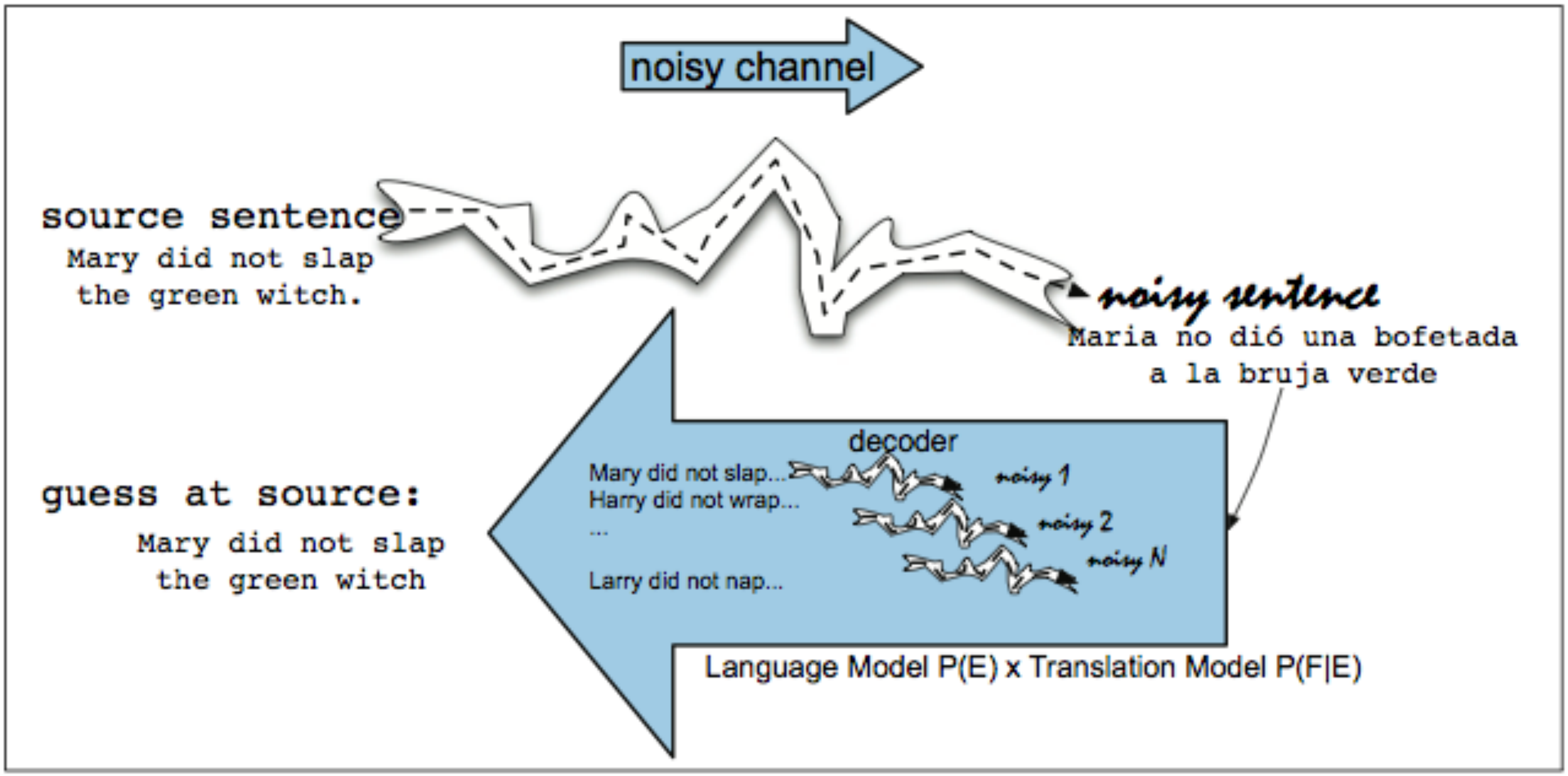
$$= \text{argmax}_e \, \frac{P(f \mid e) \, P(e)}{P(f)}$$

$$= \text{argmax}_e \, P(f \mid e) \, P(e)$$

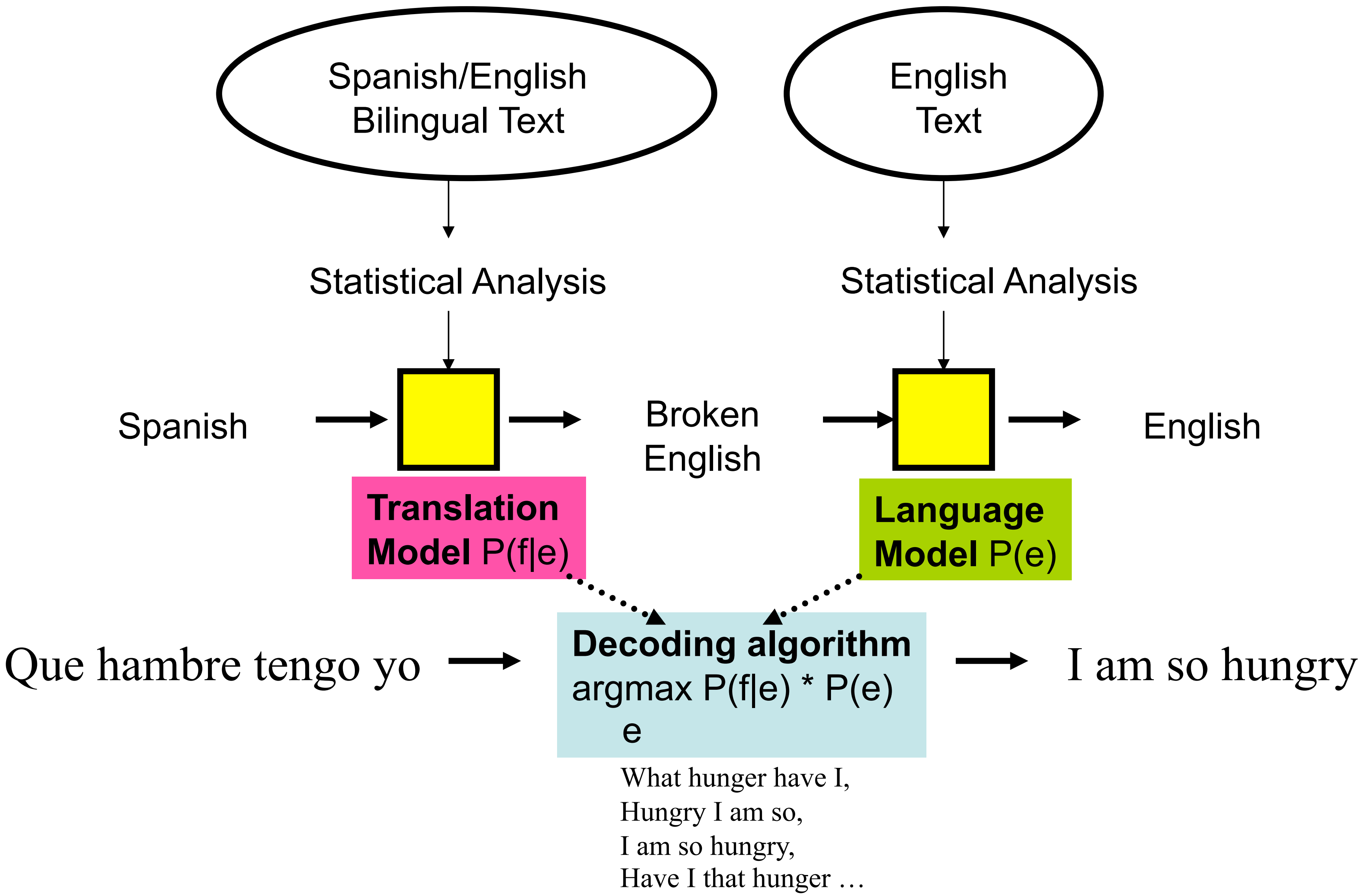
translation model
(fidelity)

language model
(fluency)

The “noisy channel” model



Statistical MT Systems



A division of labor

- Use of Bayes Rule ("the noisy channel model") allows a division of labor:
  - Job of the translation model P(f|e) is just to model how various English words typically get translated into French (perhaps in a certain context)
    - P(f|e) doesn't have to worry about language-particular facts about English word order: that's the job of P(e)
  - The job of the language model is to choose felicitous bags of words and to correctly order them for English
    - P(e) can do bag generation: putting a bag of words in order:
      - E.g., hungry I am so → I am so hungry
- Both can be incomplete/sloppy

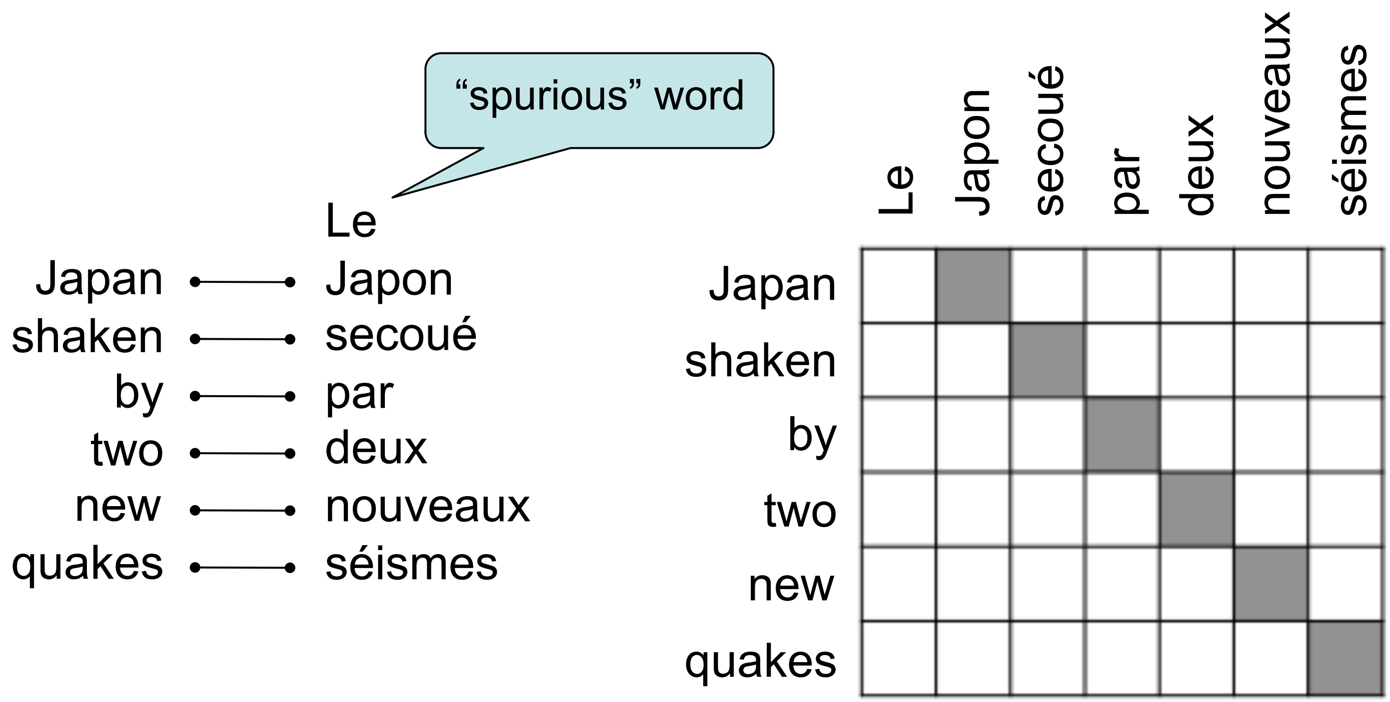
Plan of action

Statistical MT in five easy lectures!

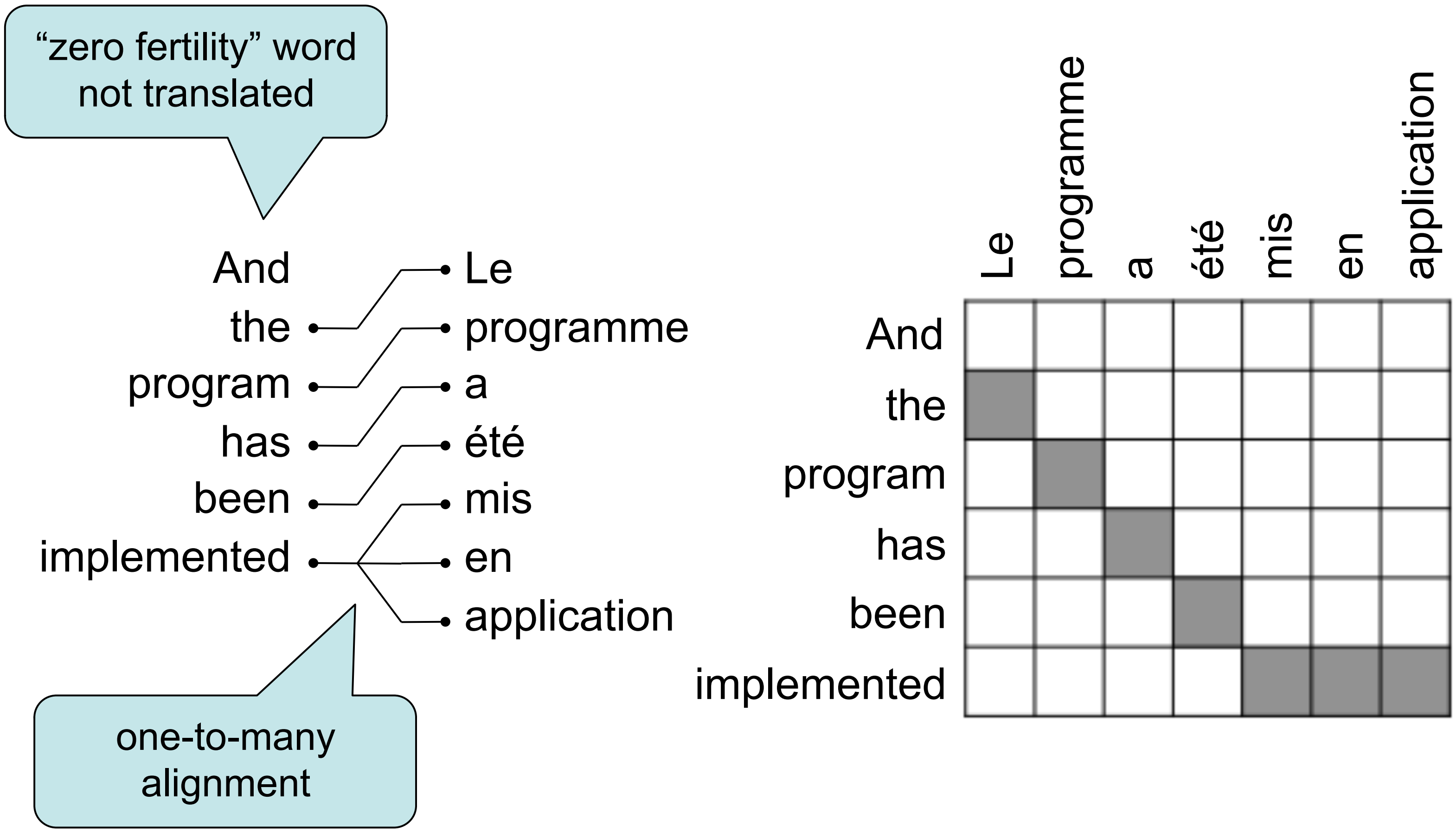
- Last time: language models
- Today: translation models
  - word alignments
  - the IBM sequence of translation models
- Next time: EM for word alignment models
- Then: MT systems, decoding, evaluation
- Then: phrase-based MT, syntactic MT

Alignments

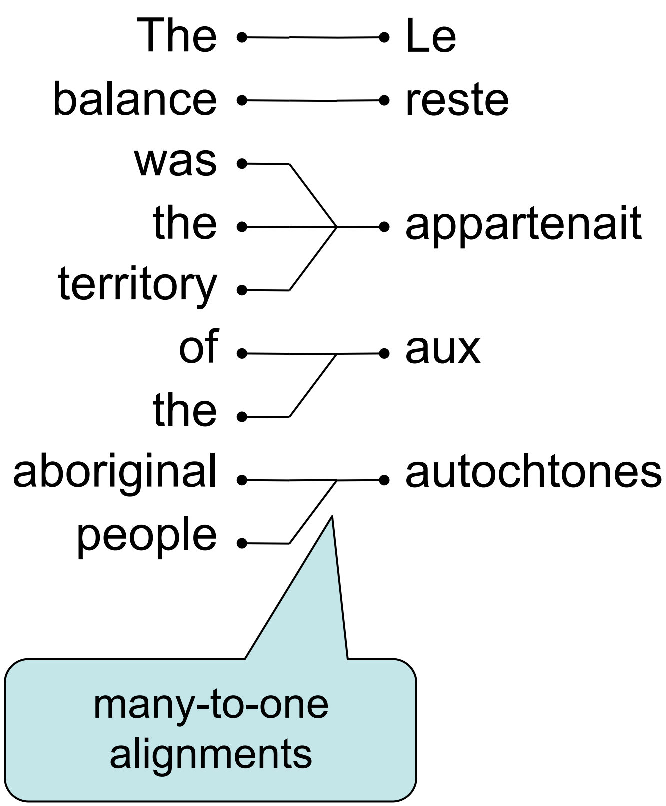
We can factor the translation model $P(f \mid e)$ by identifying *alignments* (correspondences) between words in *e* and words in *f*

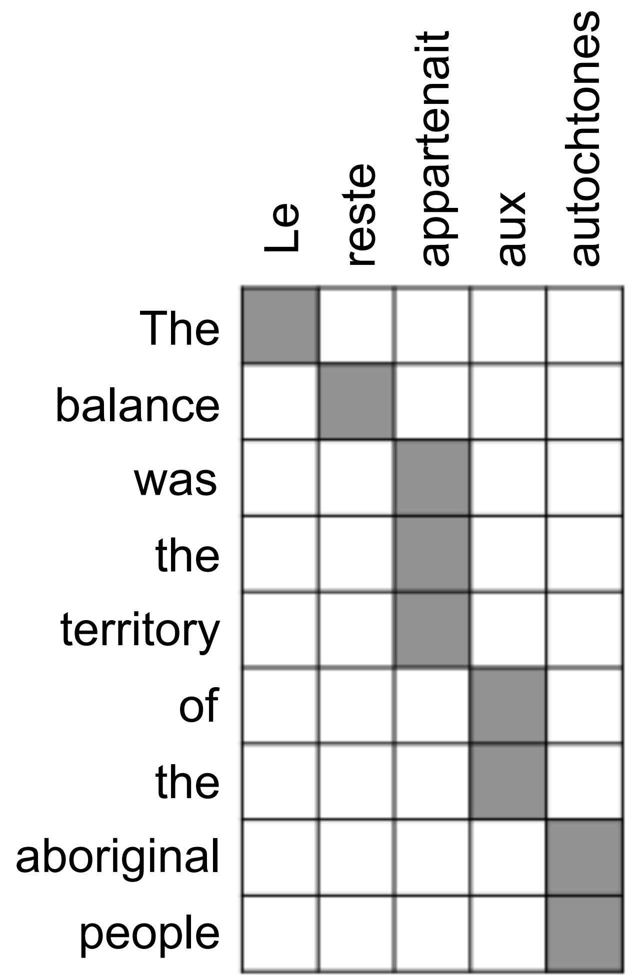


Alignments: harder

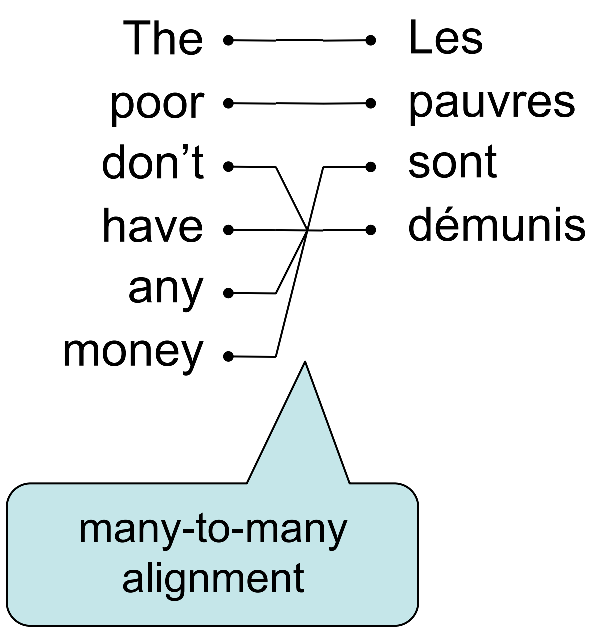


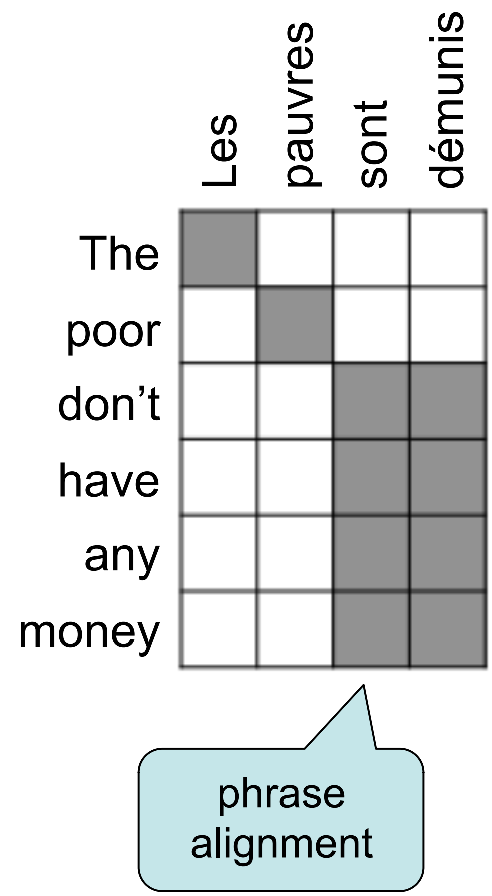
Alignments: harder



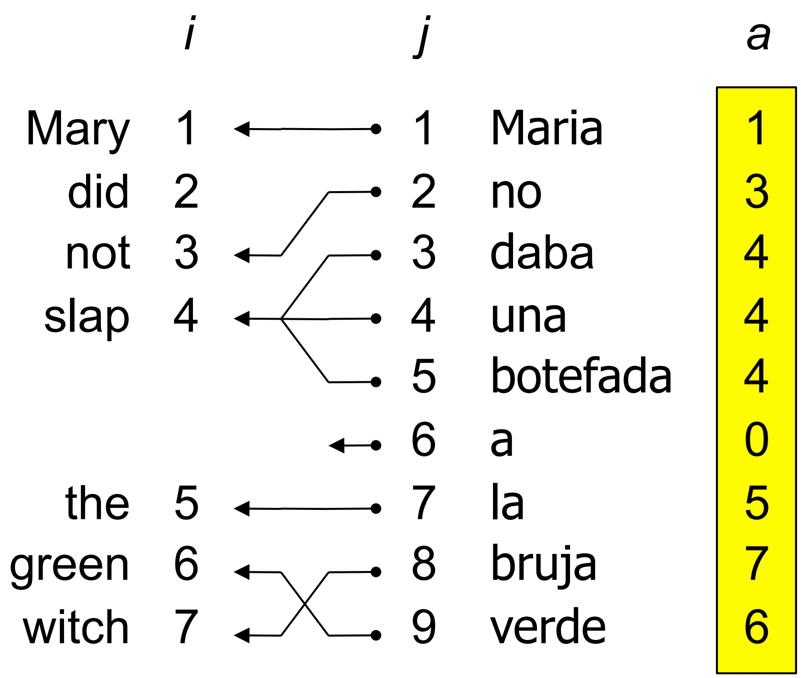


Alignments: hardest





Alignment as a vector



- used in all IBM models
- *a* is vector of length *J*
- maps indexes *j* to indexes *i*
- each $a_j \in \{0, 1 \ldots I\}$
- $a_j = 0 \Leftrightarrow f_j$ is "spurious"
- no many-to-one alignments
- no many-to-many alignments
- but provides foundation for phrase-based alignment

Today's (easy!) quiz question

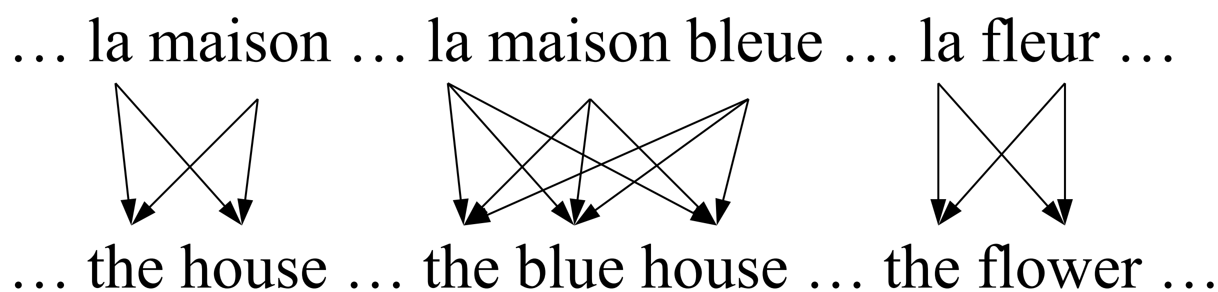
How many possible (IBM-style) alignments?

	i	*j*		*a*
Mary	1	1	Maria	-
did	2	2	no	-
not	3	3	daba	-
slap	4	4	una	-
		5	botefada	-
		6	a	-
the	5	7	la	-
green	6	8	bruja	-
witch	7	9	verde	-

A. 362880
B. 4782969
C. 40353607
D. 43046721
E. 134217728

Unsupervised Word Alignment

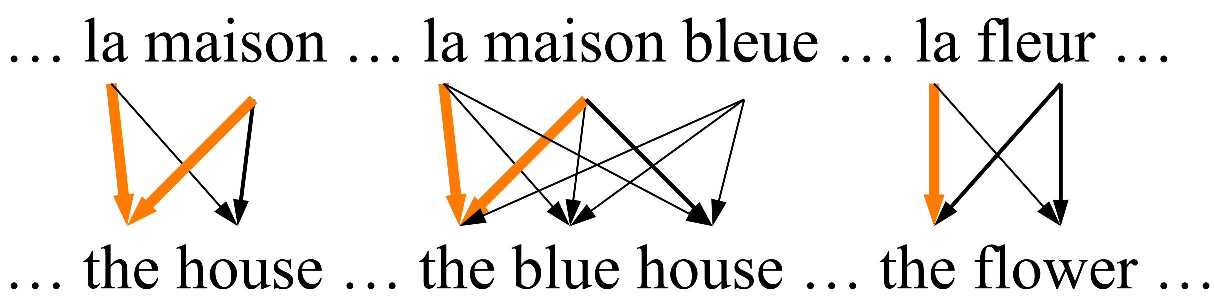
Input: a *bitext*: pairs of translated *sentences*
Output: *alignments*: pairs of translated *words*



All word alignments equally likely

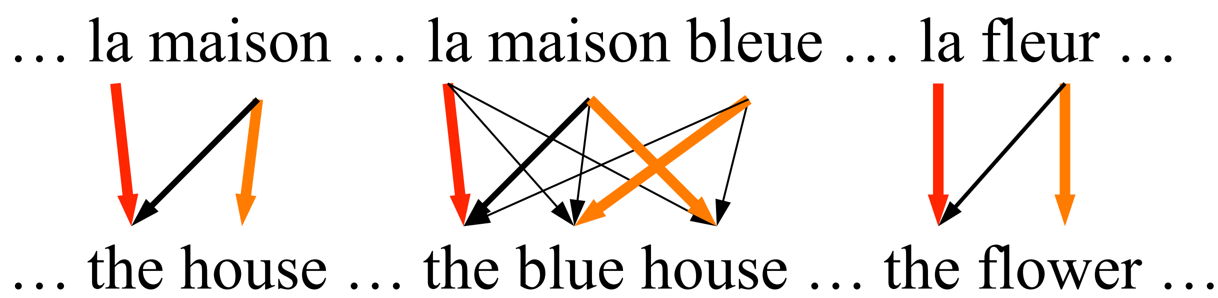
All P(french-word | english-word) equally likely

Unsupervised Word Alignment



“la” and “the” observed to co-occur frequently,
so P(la | the) is increased.

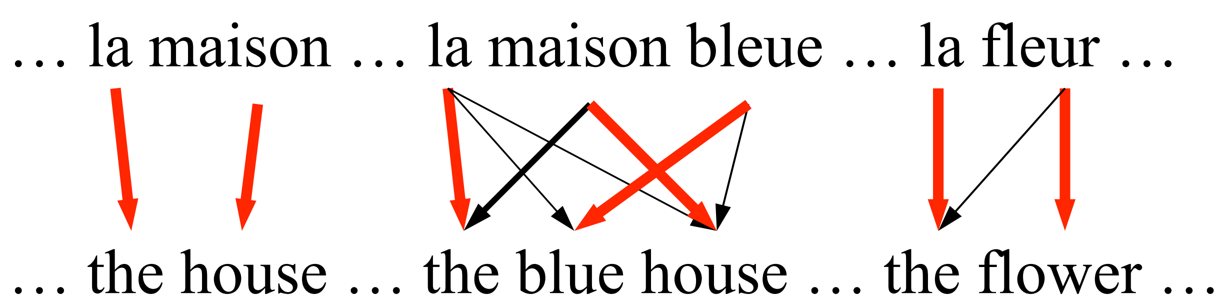
Unsupervised Word Alignment



“maison” co-occurs with both “the” and “house”, but P(maison | house) can be raised without limit, to 1.0, while P(maison | the) is limited because of “la”

(pigeonhole principle)

Unsupervised Word Alignment

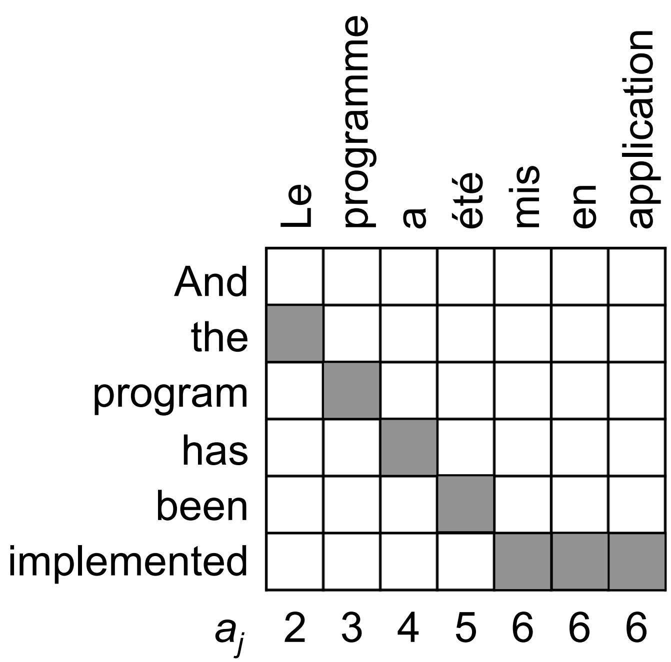


settling down after another iteration

That was the idea of IBM Model 1. For details, see the next slides and:

- "A Statistical MT Tutorial Workbook" (Knight, 1999).
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

IBM Model 1 generative story



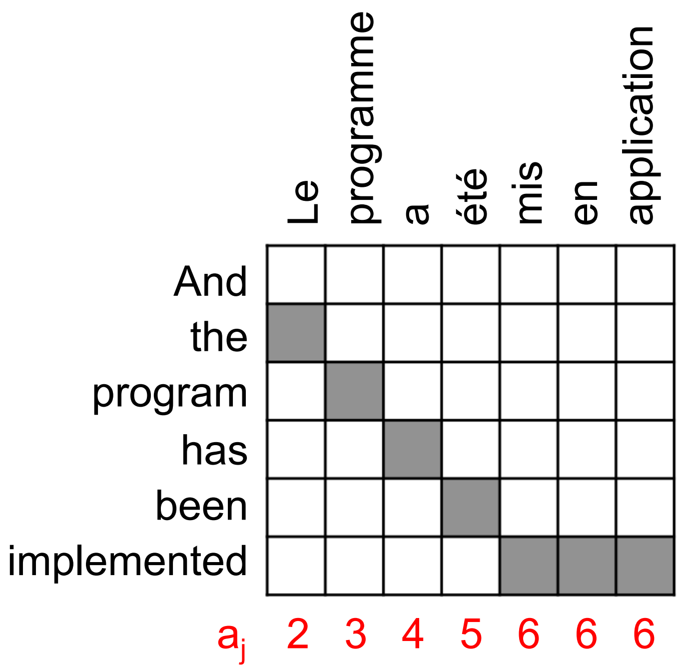
Given English sentence $e_1, e_2, \ldots e_I$

Choose length J for French sentence

For each j in 1 to J:

- Choose a_j uniformly from 0, 1, … I
- Choose f_j by translating e_{aj}

IBM Model 1 parameters



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

$$= \epsilon \prod_j P(a_j) P(f_j|e_{a_j})$$

$$= \epsilon \prod_j \frac{1}{I+1} P(f_j|e_{a_j})$$

$$= \frac{\epsilon}{(I+1)^J} \prod_j P(f_j|e_{a_j})$$

Applying Model 1*

P(*f*, *a* | *e*) can be used as a *translation model* or an *alignment model*

As translation model

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

As alignment model

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

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

* Actually, any *P*(*f*, *a* | *e*), e.g., any IBM model

Applying Model 1 *efficiently*

(see Knight 99, section 31)

	Le	programme	a	été	mis	en	application
And	0.01	0.02	0.01	0.02	0.01	0.03	0.01
the	0.49	0.01	0.02	0.03	0.01	0.02	0.01
program	0.01	0.63	0.01	0.02	0.01	0.03	0.01
has	0.02	0.01	0.43	0.11	0.06	0.04	0.01
been	0.02	0.01	0.17	0.39	0.01	0.02	0.01
implemented	0.01	0.02	0.01	0.01	0.11	0.12	0.37

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

$$\propto \sum_a \prod_j P(f_j|e_{a_j})$$

exponential?

$$= \sum_{a_1=0}^{I} \dots \sum_{a_J=0}^{I} \prod_j P(f_j|e_{a_j})$$

$$\propto \prod_j \sum_i P(f_j|e_i)$$

quadratic!

Model 1: Word alignment learning with Expectation-Maximization (EM)

- Start with $P(f^p|e^q)$ uniform, including $P(f^j|\text{NULL})$
- For each sentence
  - For each French position *j*
    - Calculate posterior over English positions $P(a_j \mid e, f)$

$$P(a_j = i \mid f, e) = \frac{P(f_j \mid e_i)}{\sum_{i'} P(f_j \mid e_{i'})}$$

    - Increment count of word f_j with word e_{a_j}
      - $C(f_j|e_i) \mathrel{+}= P(a_j = i \mid f,e)$
- Renormalize counts to give probs $P(f^p \mid e^q) = \dfrac{C(f^p \mid e^q)}{\sum_{f^x} C(f^x \mid e^q)}$
- Iterate until convergence

IBM StatMT Translation Models

- IBM1 – lexical probabilities only
- IBM2 – lexicon plus absolute position
- HMM – lexicon plus relative position
- IBM3 – plus fertilities
- IBM4 – inverted relative position alignment
- IBM5 – non-deficient version of model 4

- All the models we discuss today handle 0:1, 1:0, 1:1, 1:n alignments *only*

[Brown et al. 93, Vogel et al. 96]

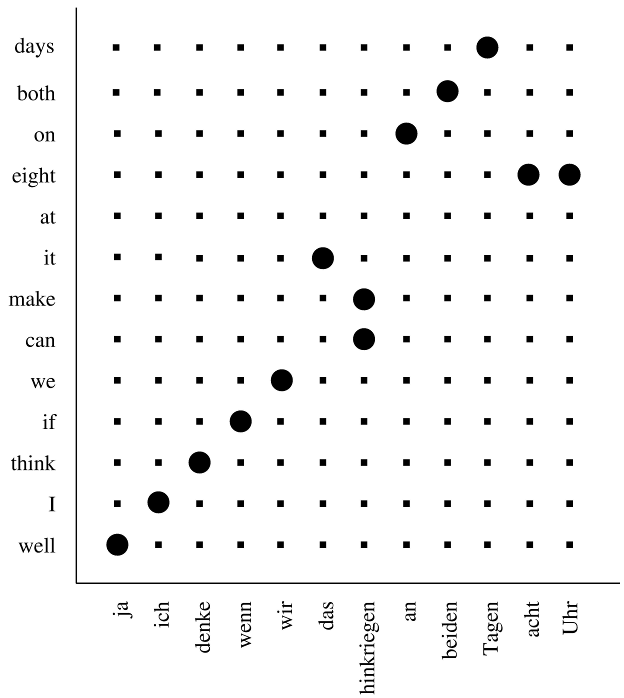
Comparative results

[Och & Ney 03]

		Size of training corpus			
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

IBM models 1,2,3,4,5

- In Model 2, the placement of a word in the French depends on where it was in the English



•Unlike Model 1, Model 2 captures the intuition that translations should usually “lie along the diagonal”.

•The main focus of PA #2.

IBM Models 1,2,<u>3</u>,4,5

- In Model 3 we model how many French words an English word can produce, using a concept called *fertility*

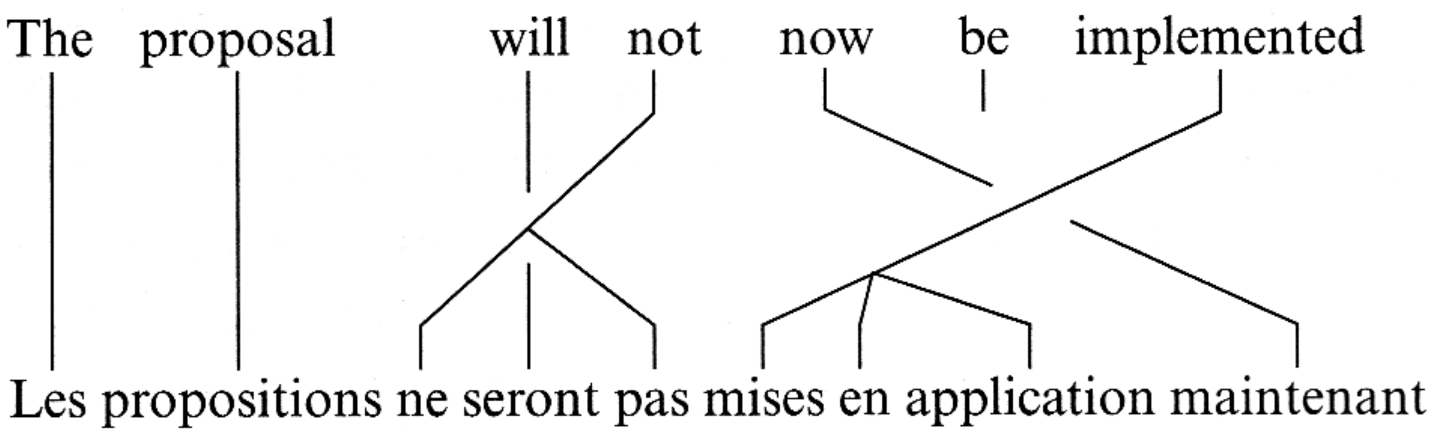
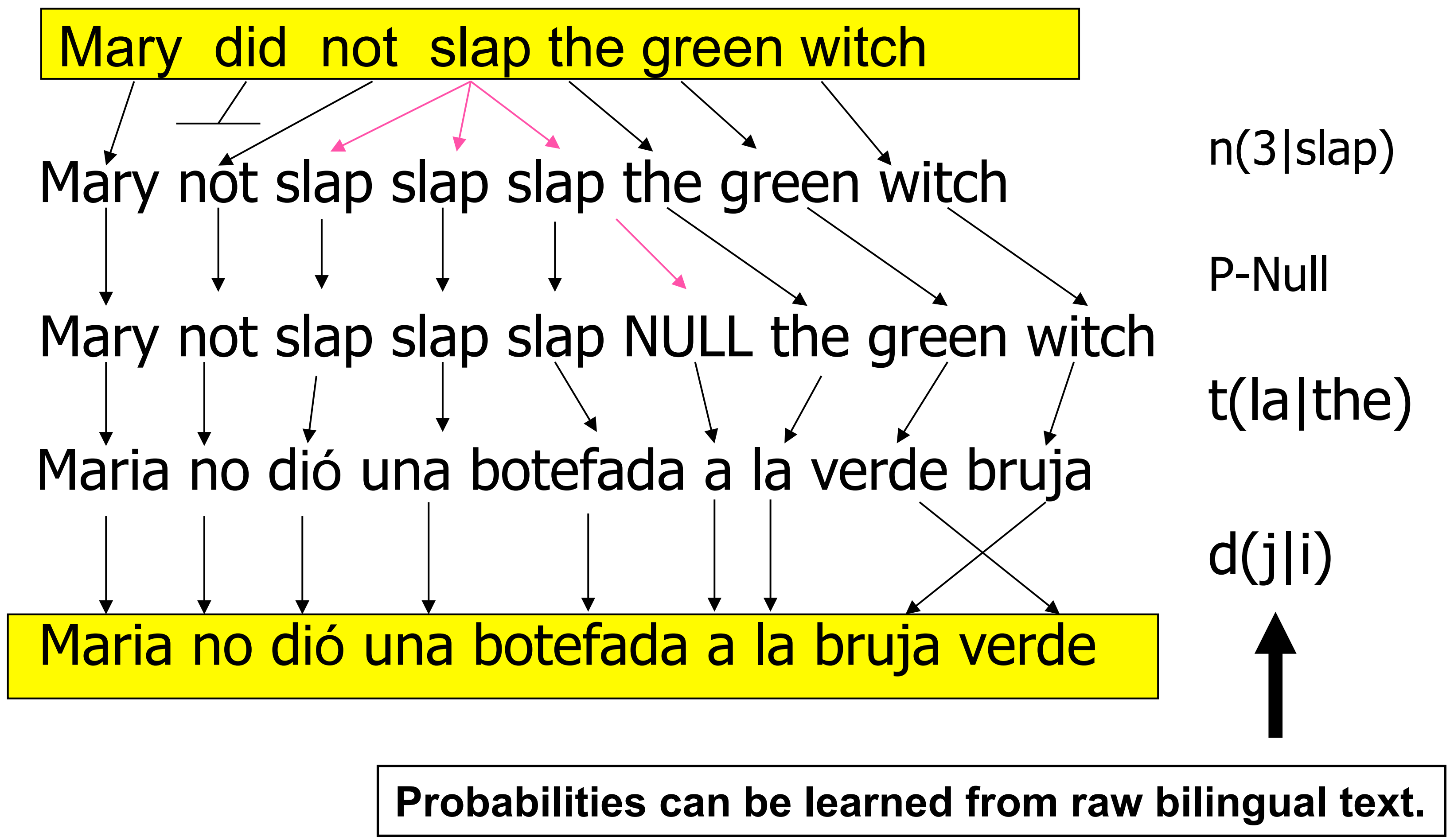


Figure 32.3
Alignment example.

Model 3 generative story



IBM Model 3 (from Knight 99)

- For each word e_i in English sentence, choose a fertility Φ_i. The choice of Φ_i depends only on e_i, not other words or Φ's.
- For each word e_i, generate Φ_i Spanish words. Choice of French word depends only on English word e_i, not English context or any Spanish words.
- Permute all the Spanish words. Each Spanish word gets assigned absolute target position slot (1,2,3, etc). Choice of Spanish word position dependent only on absolute position of English word generating it.

Model 3: P(f|e) parameters

- What are the parameters for this model?
- Words: P(casa|house)
- Spurious words: P(a|null)
- Fertilities: n(1|house): prob that "house" will produce 1 Spanish word whenever it appears.
- Distortions: d(5|2) prob that word in position 2 of English sentence generates word in position 5 of French translation
  - Actually, distortions are d(5|2,4,6) where 4 is length of English sentence, 6 is Spanish length

Spurious words

- We could have n(3|NULL) (probability of being exactly 3 spurious words in a Spanish translation)
- But instead, of n(0|NULL), n(1|NULL) … n(25|NULL), have a single parameter p1
- After assign fertilities to non-NULL English words we want to generate (say) z Spanish words.
- As we generate each of z words, we optionally toss in spurious Spanish word with probability p_1
- Probability of not adding spurious word: $p_0 = 1 – p_1$

Distortion probabilities for spurious words

- Can't just have d(5|0,4,6), I.e. chance that NULL word will end up in position 5.
- Why? These are spurious words! Could occur anywhere!! Too hard to predict
- Instead,
  - Use normal-word distortion parameters to choose positions for normally-generated Spanish words
  - Put NULL-generated words into empty slots left over
  - If three NULL-generated words, and three empty slots, then there are 3!, or six, ways for slotting them all in
  - We'll assign a probability of 1/6 for each way

Real Model 3 story

- For each word e_i in English sentence, choose fertility Φ_i with prob $n(\Phi_i | e_i)$
- Choose number Φ_0 of spurious Spanish words to be generated from e_0=NULL using p_1 and sum of fertilities from step 1
- Let m be sum of fertilities for all words including NULL
- For each i = 0,1,2,…l , k=1,2,… Φ_i :
  - choose Spanish word τ_{ik} with probability $t(\tau_{ik}|e_i)$
- For each i=1,2,…l , k=1,2,… Φ_i :
  - choose target Spanish position π_{ik} with prob $d(\pi_{ik}|I,L,m)$
- For each k=1,2,…, Φ_0 choose position π_{0k} from Φ_0 – k+1 remaining vacant positions in 1,2,…m for total prob of 1/ Φ_0!
- Output Spanish sentence with words τ_{ik} in positions π_{ik} (0<=I<=1,1<=k<= Φ_I)

Model 3 parameters

- n, t, p, d
- Again, if we had complete data of English strings and step-by-step rewritings into Spanish, we could:
  - Compute n(0|did) by locating every instance of “did”, and seeing how many words it translates to
  - t(maison|house) how many of all French words generated by “house” were “maison”
  - d(5|2,4,6) out of all times some word2 was translated, how many times did it become word5?

Since we don’t have word-aligned data…

- We bootstrap alignments from incomplete data
- From a sentence-aligned bilingual corpus
  1) Assume some startup values for n, d, Φ, etc.
  2) Use values for n, d, Φ, etc to use model 3 to work out chances of different possible alignments. Use these alignments to retrain n, d, Φ, etc
  3) Go to 2
- This is a more complicated case of the EM algorithm

Examples: translation & 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		

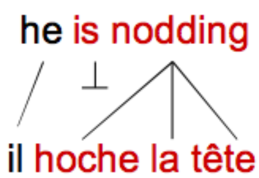
not

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

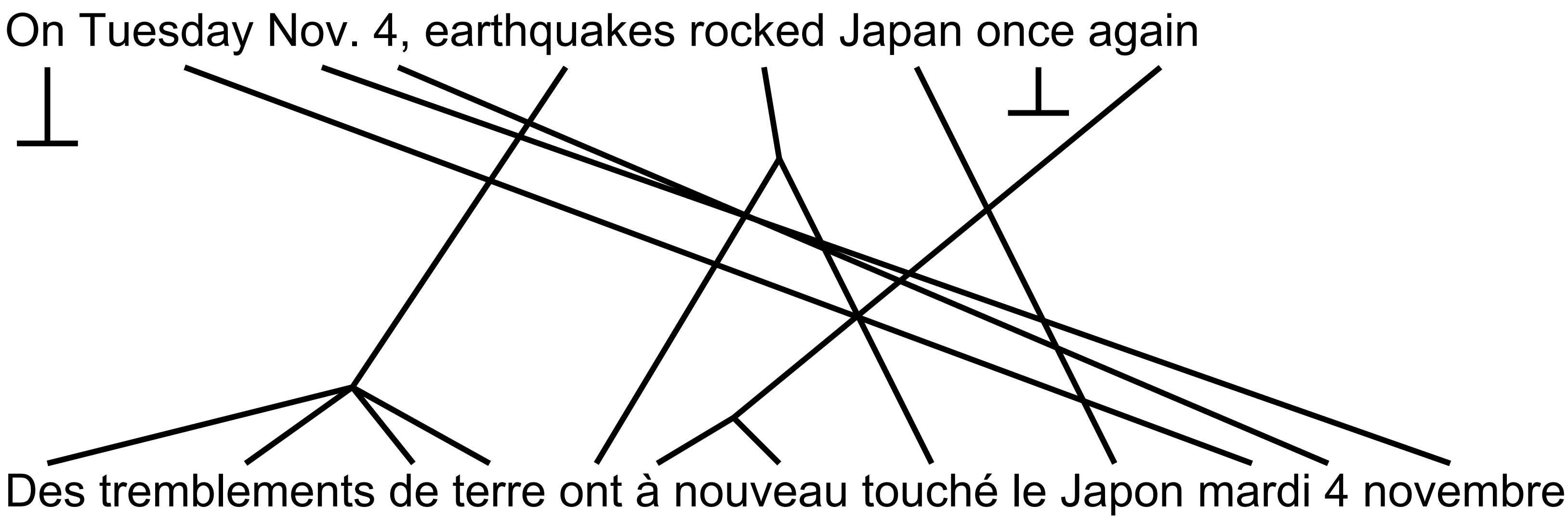
should

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		

IBM models 1,2,3,4,5

- In model 4 the placement of later French words produced by an English word depends on what happened to earlier French words generated by that same English word

Alignments: linguistics



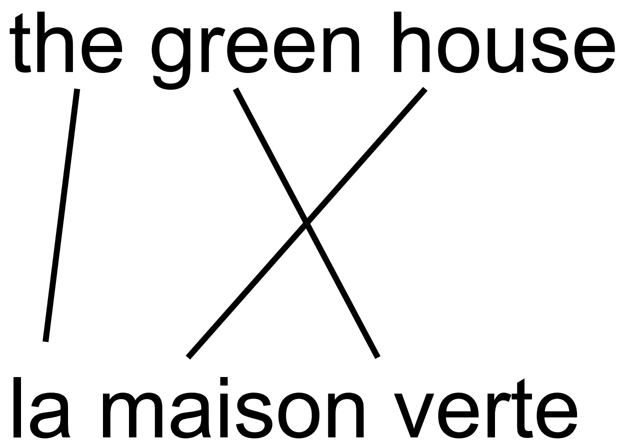
IBM models 1,2,3,4,5

- In model 5 they do non-deficient alignment. That is, you can't put probability mass on impossible things.

Why all the models?

- We don’t start with aligned text, so we have to get initial alignments from somewhere.
- Model 1 is words only, and is relatively easy and fast to train.
- We are working in a space with many local maxima, so output of model 1 can be a good place to start model 2. Etc.
- The sequence of models allows a better model to be found faster [the intuition is like deterministic annealing].

Alignments: linguistics

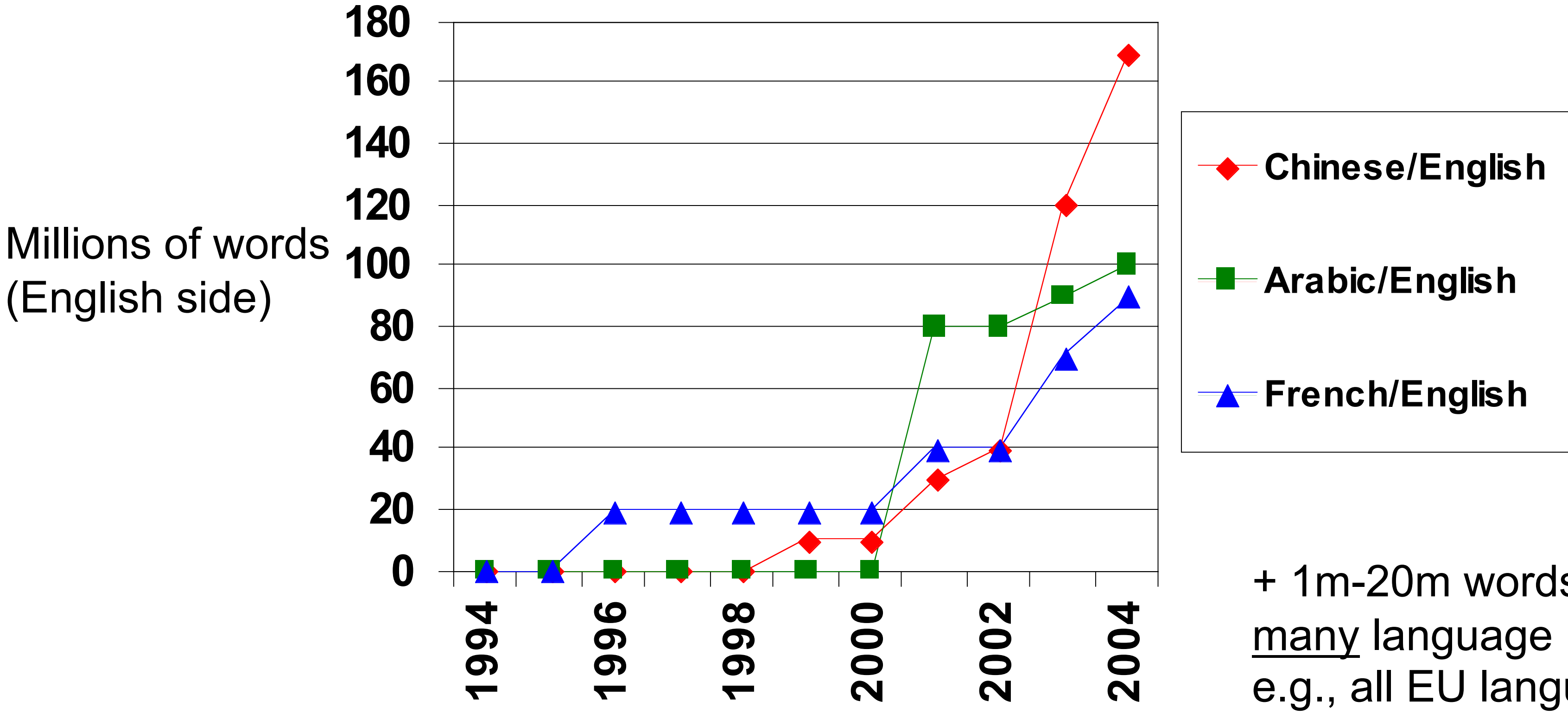


- There isn’t enough linguistics to explain this in the translation model … have to depend on the language model … that may be unrealistic … and may be harming our translation model

From No Data to Sentence Pairs

- Really hard way: pay $$$
  - Suppose one billion words of parallel data were sufficient
  - At 20 cents/word, that's $200 million
- Pretty hard way: Find it, and then earn it!
  - De-formatting
  - Remove strange characters
  - Character code conversion
  - Document alignment
  - **Sentence alignment**
  - **Tokenization (also called Segmentation)**
- Easy way: Linguistic Data Consortium (LDC)

Ready-to-Use Online Bilingual Data



+ 1m-20m words for <u>many</u> language pairs, e.g., all EU languages)

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Tokenization (or Segmentation)

- English
  - Input (some character stream):

    `"There," said Bob.`

  - Output (7 “tokens” or “words”):

    `" There , " said Bob .`

- Chinese
  - Input (char stream):

    美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。

  - Output:

    美国 关岛国 际机 场 及其 办公 室 均接获 一名 自称 沙地 阿拉 伯富 商拉登 等发 出 的 电子邮件。

Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Done by Dynamic Programming: see FSNLP ch. 13 for details