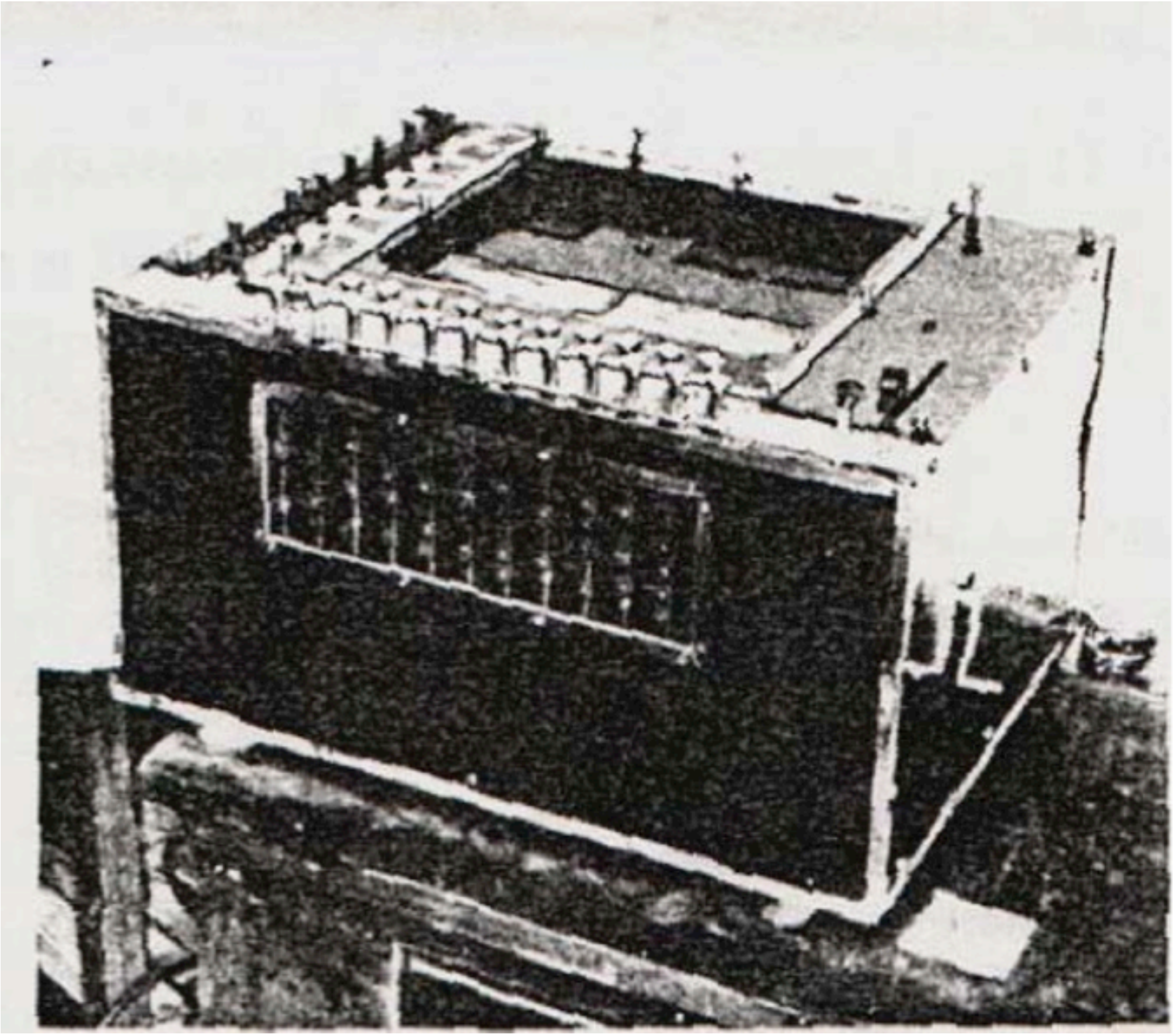
Machine Translation Part 2, and the EM Algorithm

CS 585, Fall 2015
Introduction to Natural Language Processing
http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

[Some slides borrowed from mt-class.org]



Georges Artrouni's “mechanical brain”, a translation device patented in France in 1933. (Image from Corbé by way of John Hutchins)

IBM Model 1: Inference and learning

- Alignment inference:
  Given lexical translation probabilities,
  infer posterior or Viterbi alignment

$$\arg\max_{\mathbf{a}} p(\mathbf{a} \mid \mathbf{e}, \mathbf{f}, \theta)$$

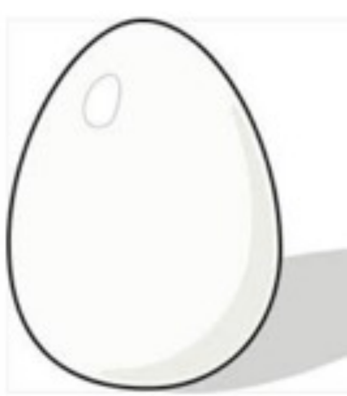
- Translation: incorporate into noisy channel
  (this model isn't good at this)

$$\arg\max_{\mathbf{f}} p(\mathbf{e} \mid \mathbf{f}, \theta) \;\; p(\mathbf{f})$$

- How do we learn translation parameters?
  EM Algorithm

$$\arg\max_{\theta} p(\mathbf{e} \mid \mathbf{f}, \theta)$$

- Chicken and egg problem:
  If we knew alignments, translation
  parameters would be trivial (just counting)



Exercise

1a. Garcia and associates.
1b. Garcia y asociados.

2a. Carlos Garcia has three associates.
2b. Carlos Garcia tiene tres asociados.

3a. his associates are not strong.
3b. sus asociados no son fuertes.

4a. Garcia has a company also.
4b. Garcia tambien tiene una empresa.

5a. its clients are angry.
5b. sus clientes están enfadados.

6a. the associates are also angry.
6b. los asociados tambien están enfadados.

7a. the clients and the associates are enemies.
7b. los clientes y los asociados son enemigos.

8a. the company has three groups.
8b. la empresa tiene tres grupos.

9a. its groups are in Europe.
9b. sus grupos están en Europa.

10a. the modern groups sell strong pharmaceuticals.
10b. los grupos modernos venden medicinas fuertes.

11a. the groups do not sell zanzanine.
11b. los grupos no venden zanzanina.

12a. the small groups are not modern.
12b. los grupos pequeños no son modernos.

MLE

- Maximum Likelihood Estimation:
  general method to learn parameters **theta**
  from observed data **x**

$$\arg\max_\theta P(x \mid \theta)$$

- Turns out ... for simple multinomial models, the MLE is simply normalized counts!

$$\theta_{\text{dog}} \equiv P(w = \text{“dog”} \mid \theta)$$

$$\theta^{MLE} = P(corpus \mid \theta)$$

$$\Rightarrow$$

$$\theta_{\text{dog}}^{MLE} = \frac{\text{count of “dog”}}{\text{num tokens total}}$$

Naive Bayes: x: text, z: classes

Supervised Learning
Given z, learn θ

MLE algorithm:
Count words per class
θ = count(w,k)/count(k)

Unsupervised Learning
Learn z,θ at once
(Clustering)

Naive Bayes: x: text, z: classes

Supervised Learning
Given z, learn θ

MLE algorithm:
Count words per class
θ = count(w,k)/count(k)

Unsupervised Learning
Learn z,θ at once
(Clustering)

Hard EM algorithm:
Randomly initialize θ
Iterate:
1. Predict each document class
   z := argmax_z P(z | x,theta)
2. Count words per class
   θ = count(w,k)/count(k)

Soft EM:
"Expectation"-step:
Calculate z posterior values, and
M-step: *fractional* counts

EM

- Motivation: Want to learn parameters with observed data (text) but the model wants *Latent/missing* variables (alignments)
- Applications
  - Unsupservised learning: e.g. unsup. NB, unsup. HMM
  - Alignment models: e.g. IBM Model 1
    - Is Model 1 “unsupervised”?

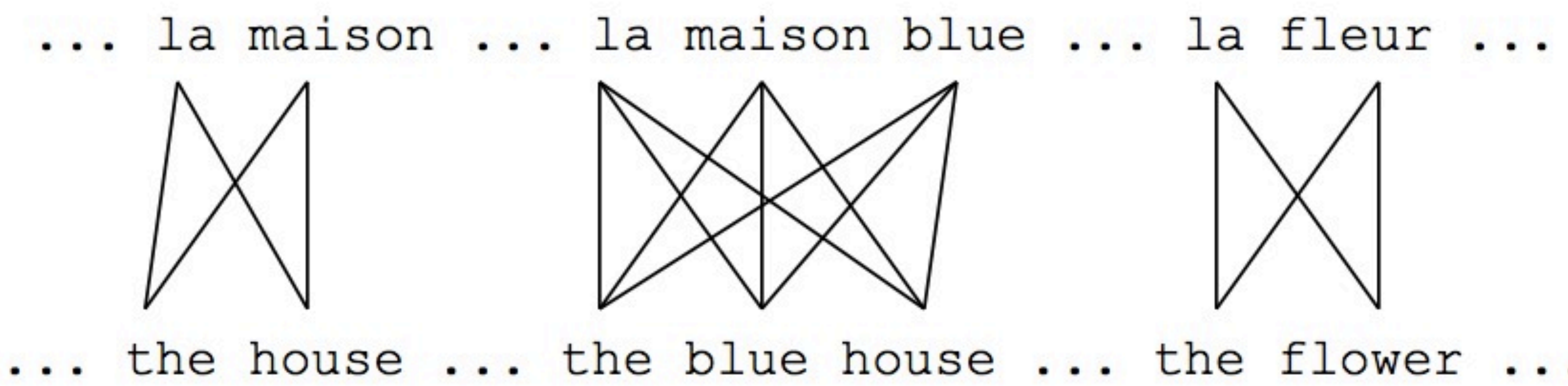
EM Algorithm

- pick some random (or uniform) parameters
- Repeat until you get bored (~ 5 iterations for lexical translation models)
  - using your current parameters, compute "expected" alignments for every target word token in the training data

    $$p(a_i \mid \mathbf{e}, \mathbf{f})$$ (on board)

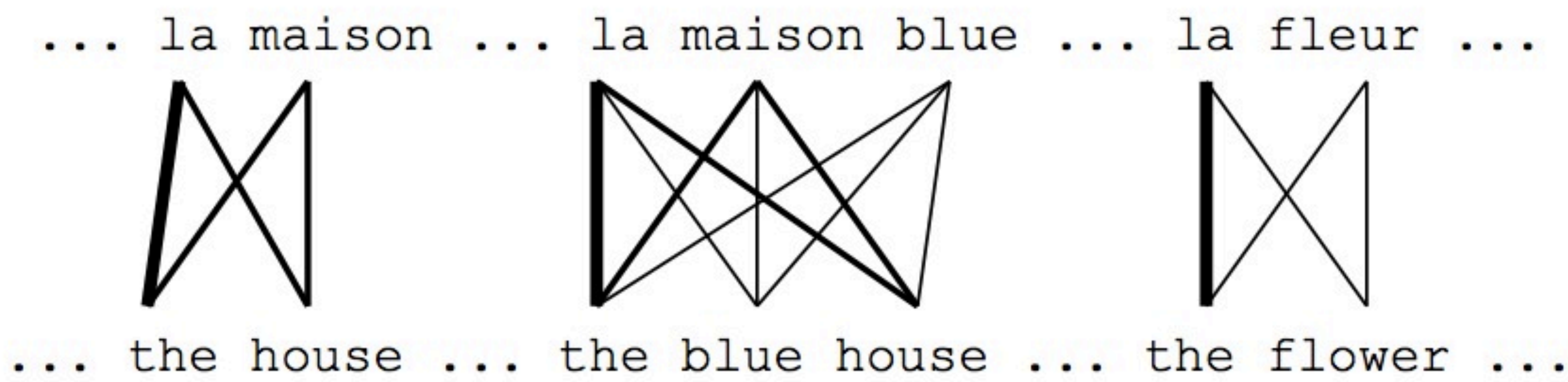
  - keep track of the expected number of times f translates into e throughout the whole corpus
  - keep track of the expected number of times that f is used as the source of any translation
  - use these expected counts as if they were "real" counts in the standard MLE equation

EM for Model 1



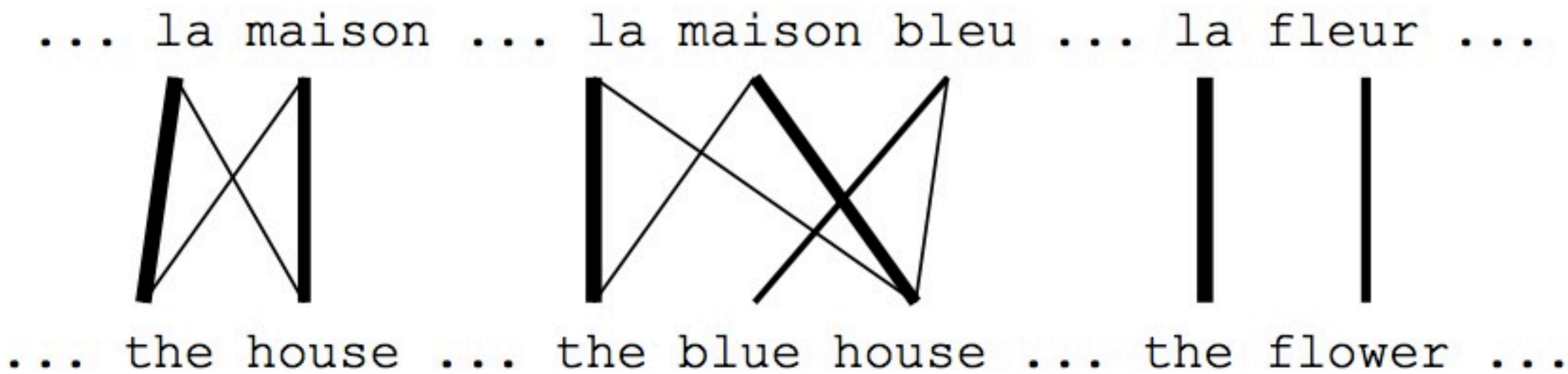
- Initial step: all alignments equally likely
- Model learns that, e.g., la is often aligned with the

EM for Model 1



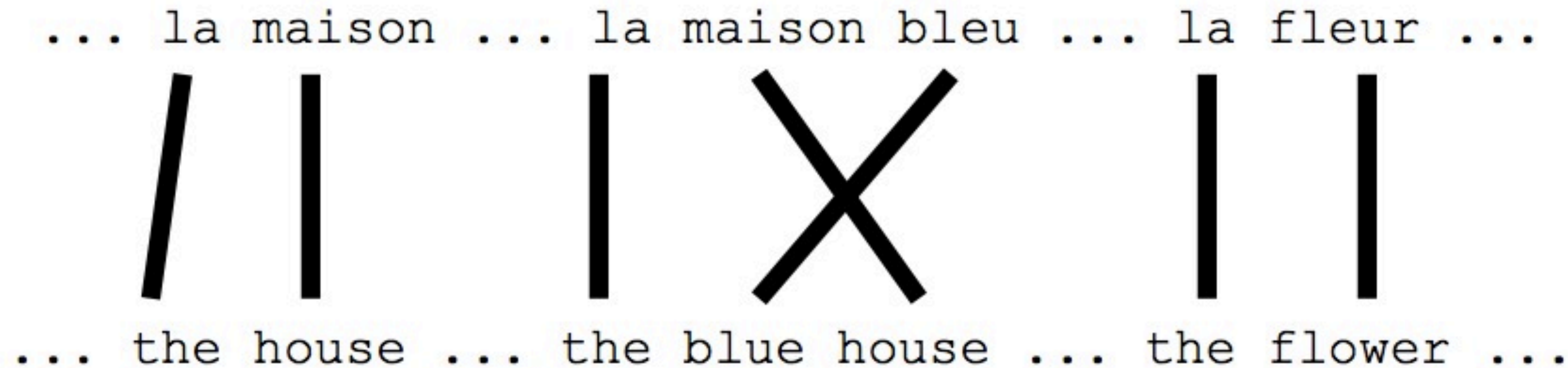
- After one iteration
- Alignments, e.g., between la and the are more likely

EM for Model 1



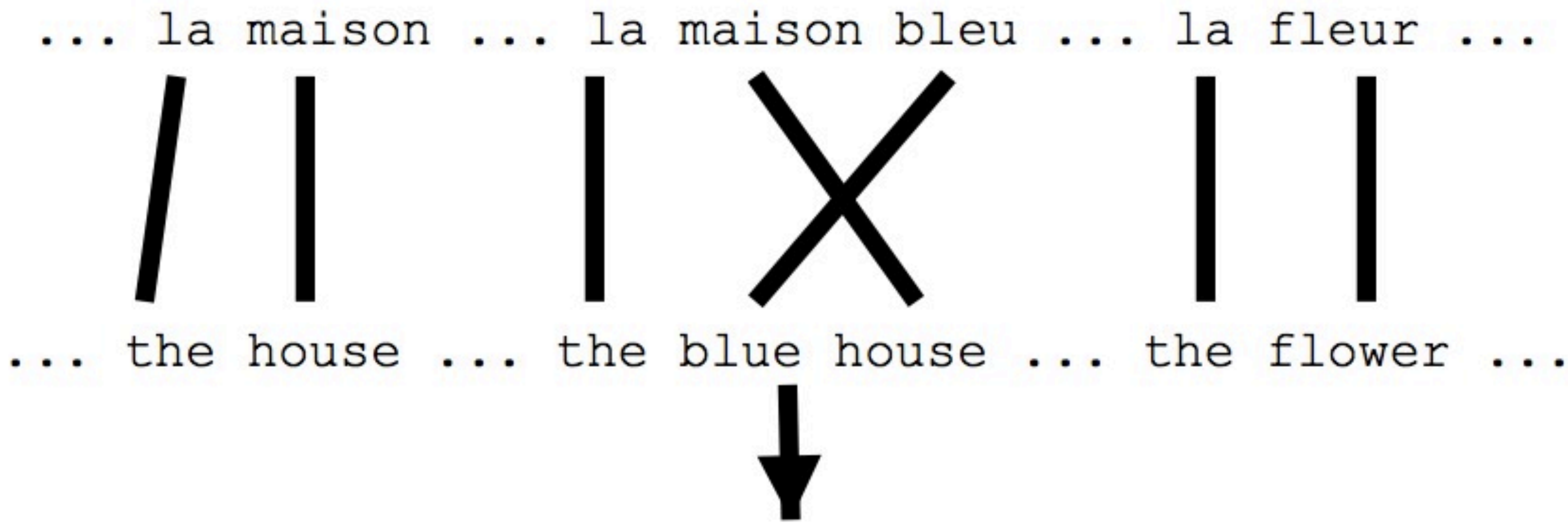
- After another iteration
- It becomes apparent that alignments, e.g., between fleur and flower are more likely (pigeon hole principle)

EM for Model 1



- Convergence
- Inherent hidden structure revealed by EM

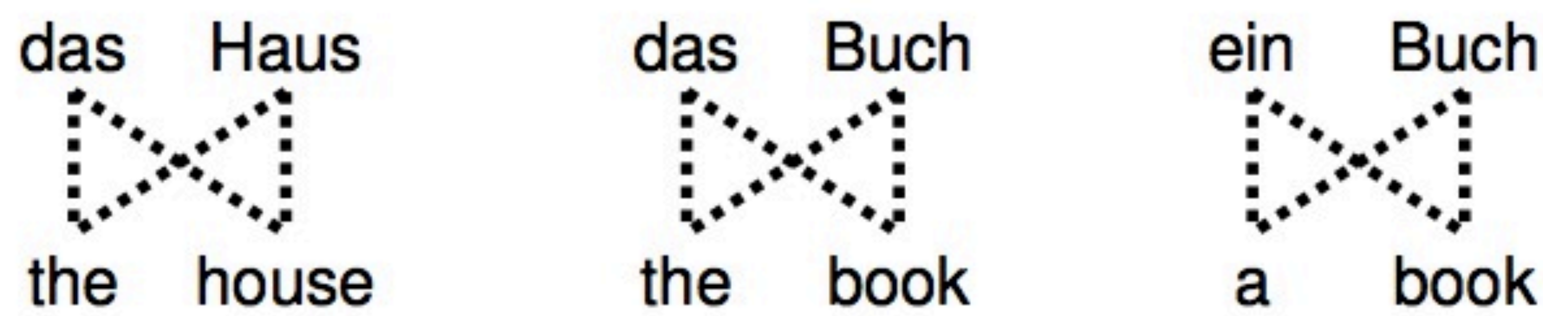
EM for Model 1



p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...

- Parameter estimation from the aligned corpus

Convergence



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

- stopped here

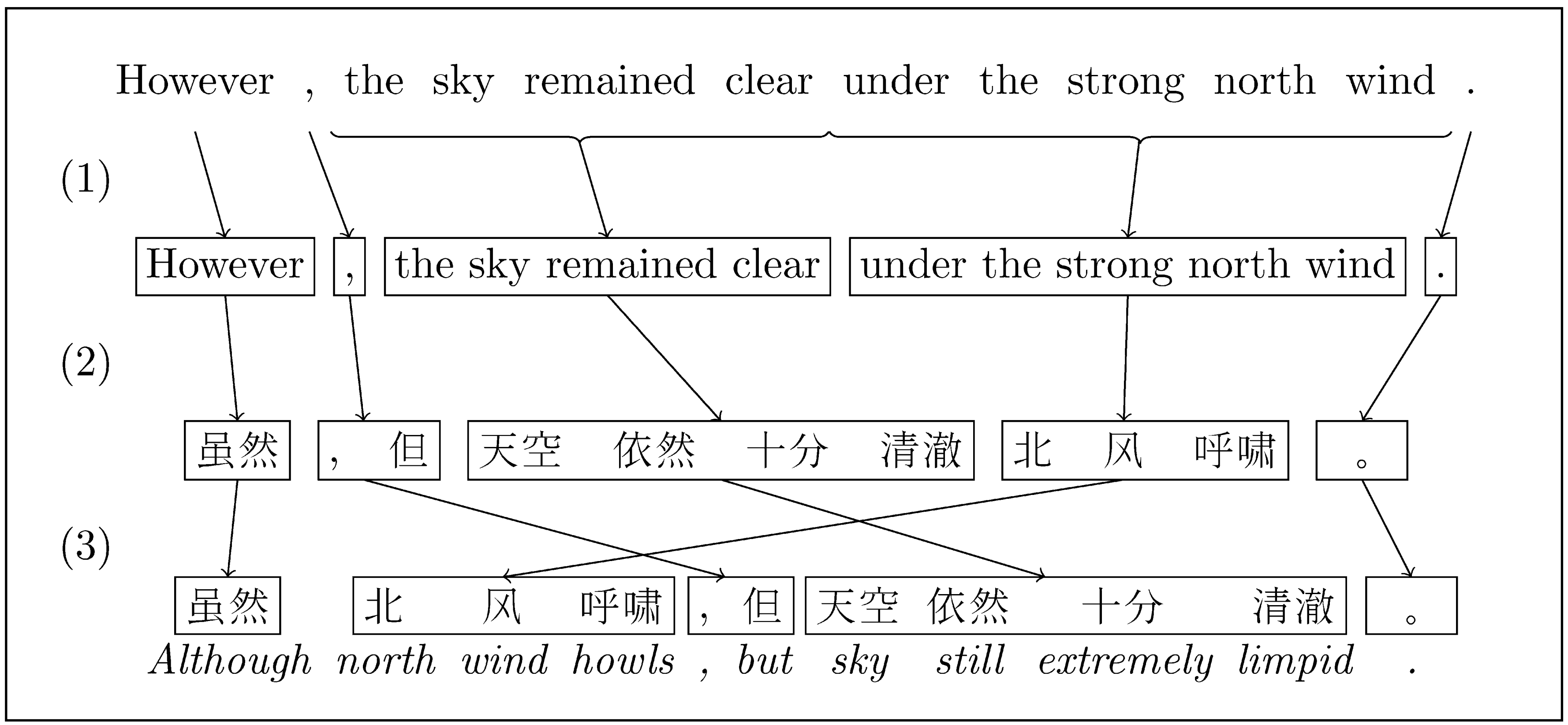
MT

- Phrase-based models
- Evaluation

Phrase-based MT

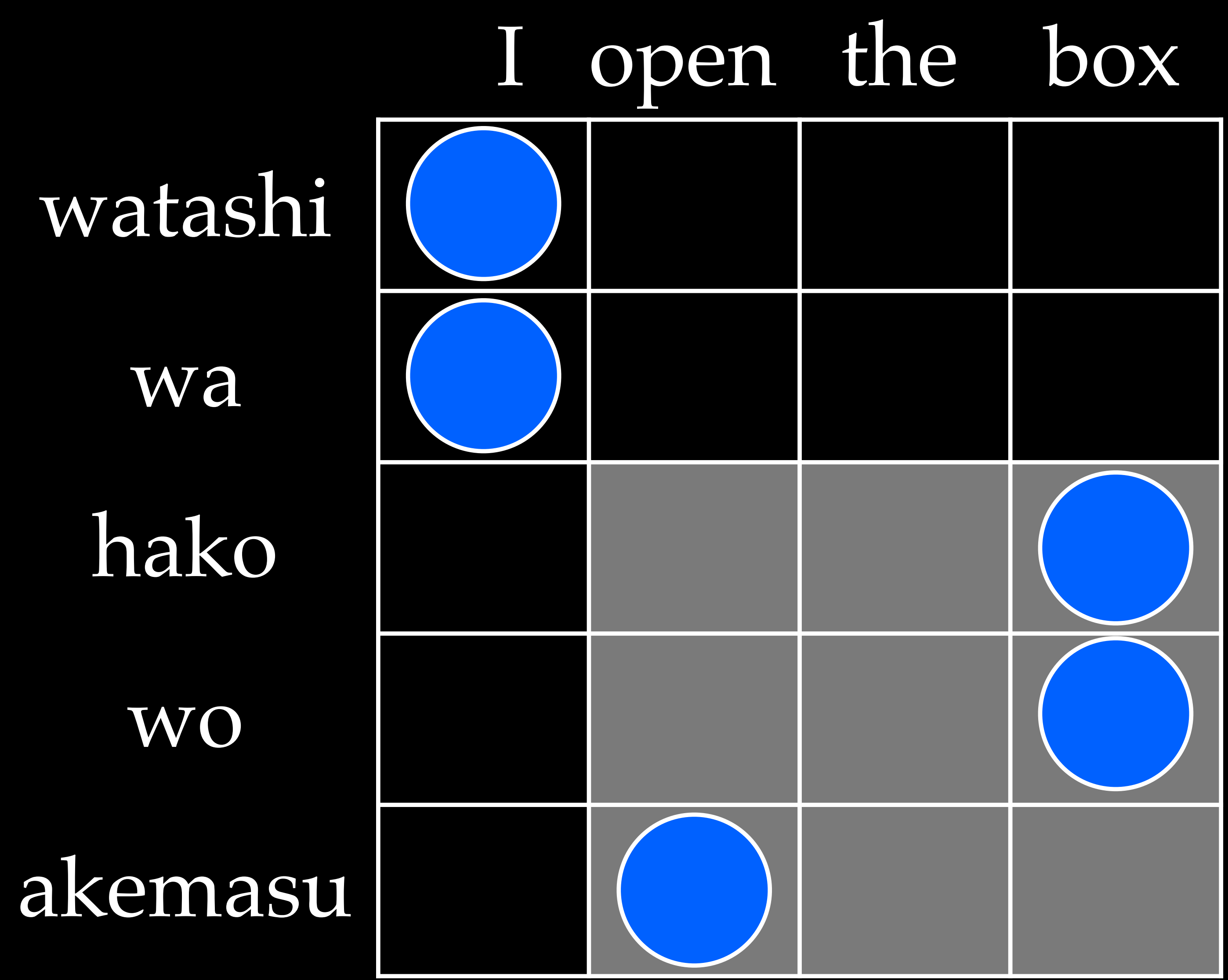
$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \underline{p(\mathbf{f} \mid \mathbf{e}, \mathbf{a})}\; p(\mathbf{a} \mid \mathbf{e})$$

Phrase-to-phrase translations



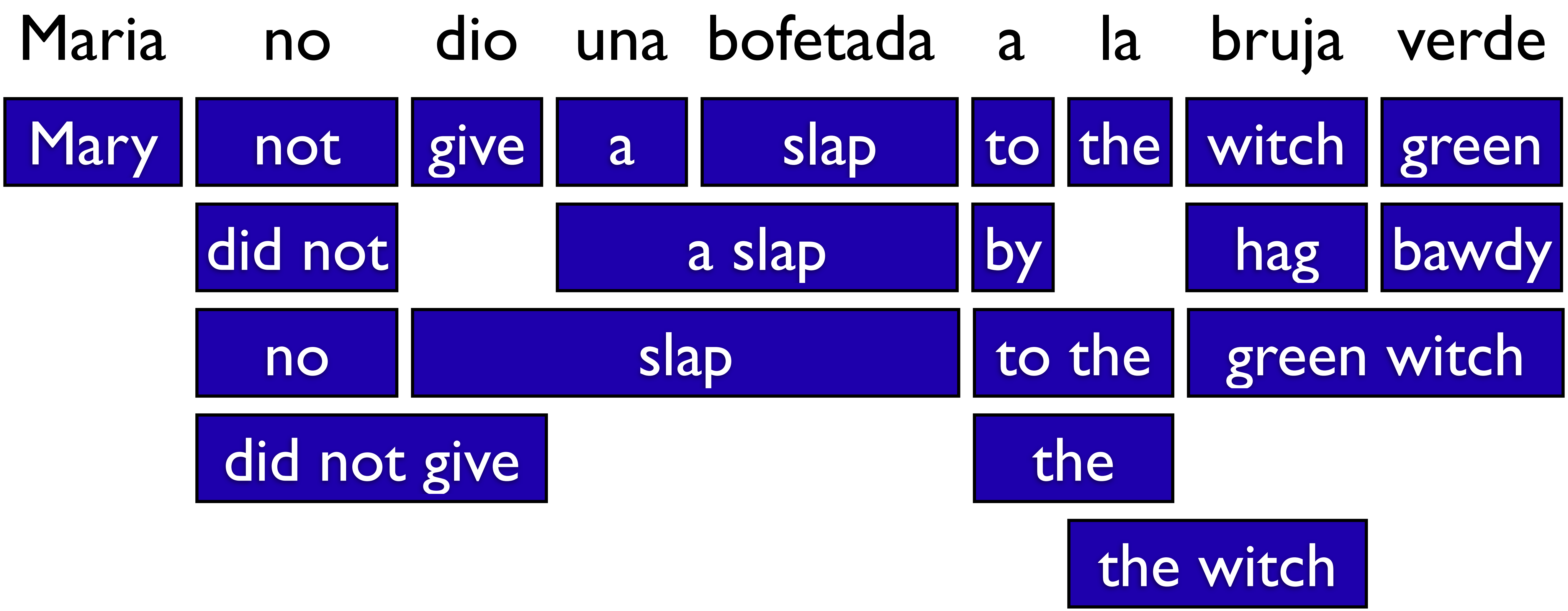
- Phrases can memorize local reorderings
- State-of-the-art (currently or very recently) in industry, e.g. Google Translate

Phrase extraction for training:
Preprocess with IBM Models to predict alignments

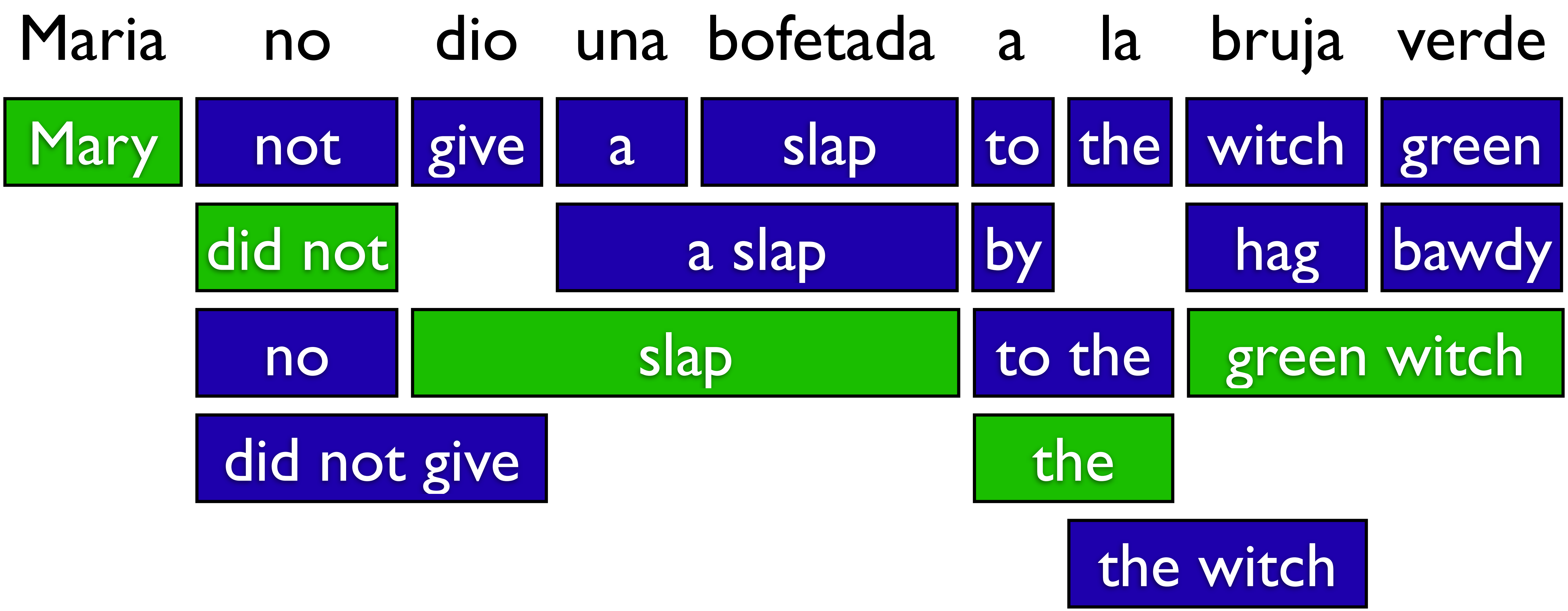


hako wo akemasu / open the box

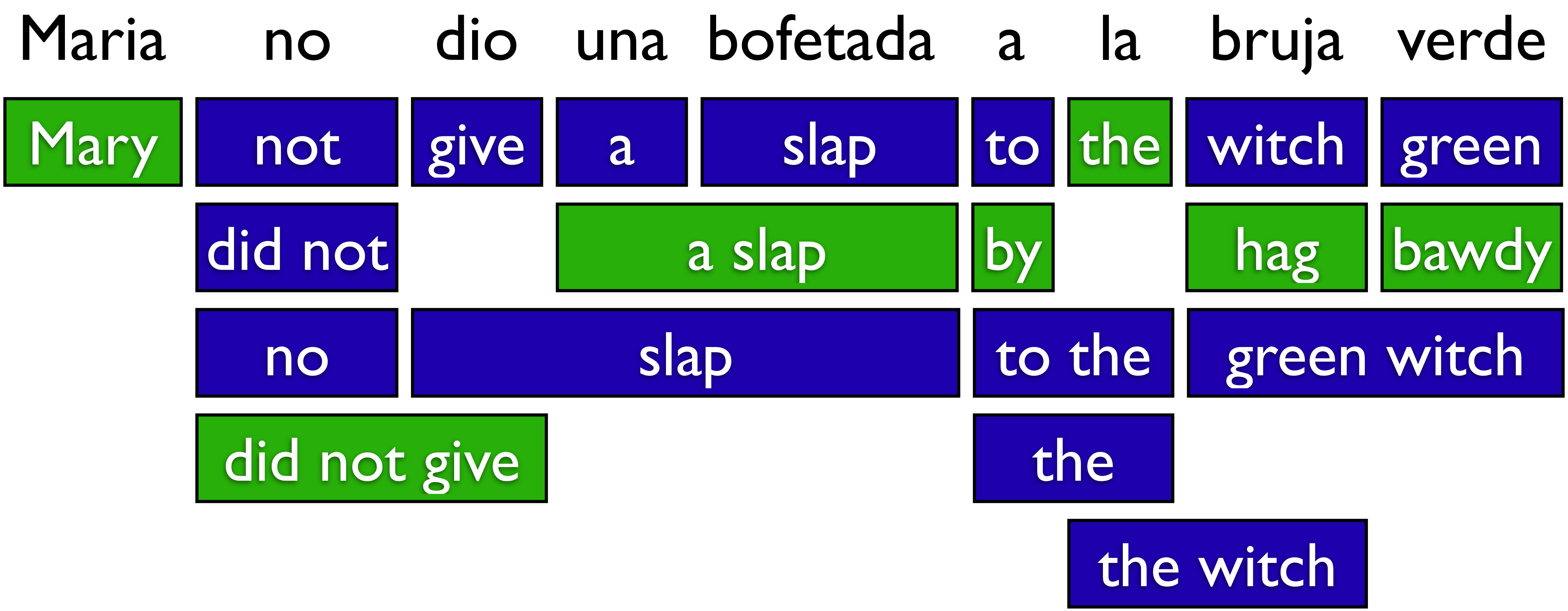
Decoding


Maria
no
dio
una
bofetada
a
la
bruja
verde
Mary
not
give
a
slap
to
the
witch
green
did not
a slap
by
hag
bawdy
no
slap
to the
green witch
did not give
the
the witch

Decoding


Maria no dio una bofetada a la bruja verde
Mary
not
give
a
slap
to
the
witch
green
did not
a slap
by
hag
bawdy
no
slap
to the
green witch
did not give
the
the witch

Decoding


Maria no dio una bofetada a la bruja verde
Mary
not
give
a
slap
to
the
witch
green
did not
a slap
by
hag
bawdy
no
slap
to the
green witch
did not give
the
the witch

MT Evaluation

Illustrative translation results

- *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)

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign
direct investment 40.007 billion US dollars today provide data include
that year to November china actually using foreign 46.959 billion US dollars and

MT Evaluation

- Manual (the best!?):
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**

- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)

- Automatic metric:
  - WER (word error rate) – why problematic?
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

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.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words *airport* are only correct if two reference words *airport;* can't cheat by typing out "the the the the the")
  - Do count unigrams also in a bigram for unigram precision, etc.
- Brevity Penalty
  - Can't just type out single word "the" (precision 1.0!)
- It was thought quite hard to "game" the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

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.

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

BLEU in Action

枪手被警方击毙。	(Foreign Original)
the gunman was shot to death by the police .	(Reference Translation)
the gunman was police kill .	#1
wounded police jaya of	#2
the gunman was shot dead by the police .	#3
the gunman arrested by police kill .	#4
the gunmen were killed .	#5
the gunman was shot to death by the police .	#6
gunmen were killed by police ?SUB>0 ?SUB>0	#7
al by the police .	#8
the ringer is killed by the police .	#9
police killed the gunman .	#10

green	= 4-gram match	(good!)
red	= word not matched	(bad!)

Multiple Reference Translations

Reference translation 1:
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.

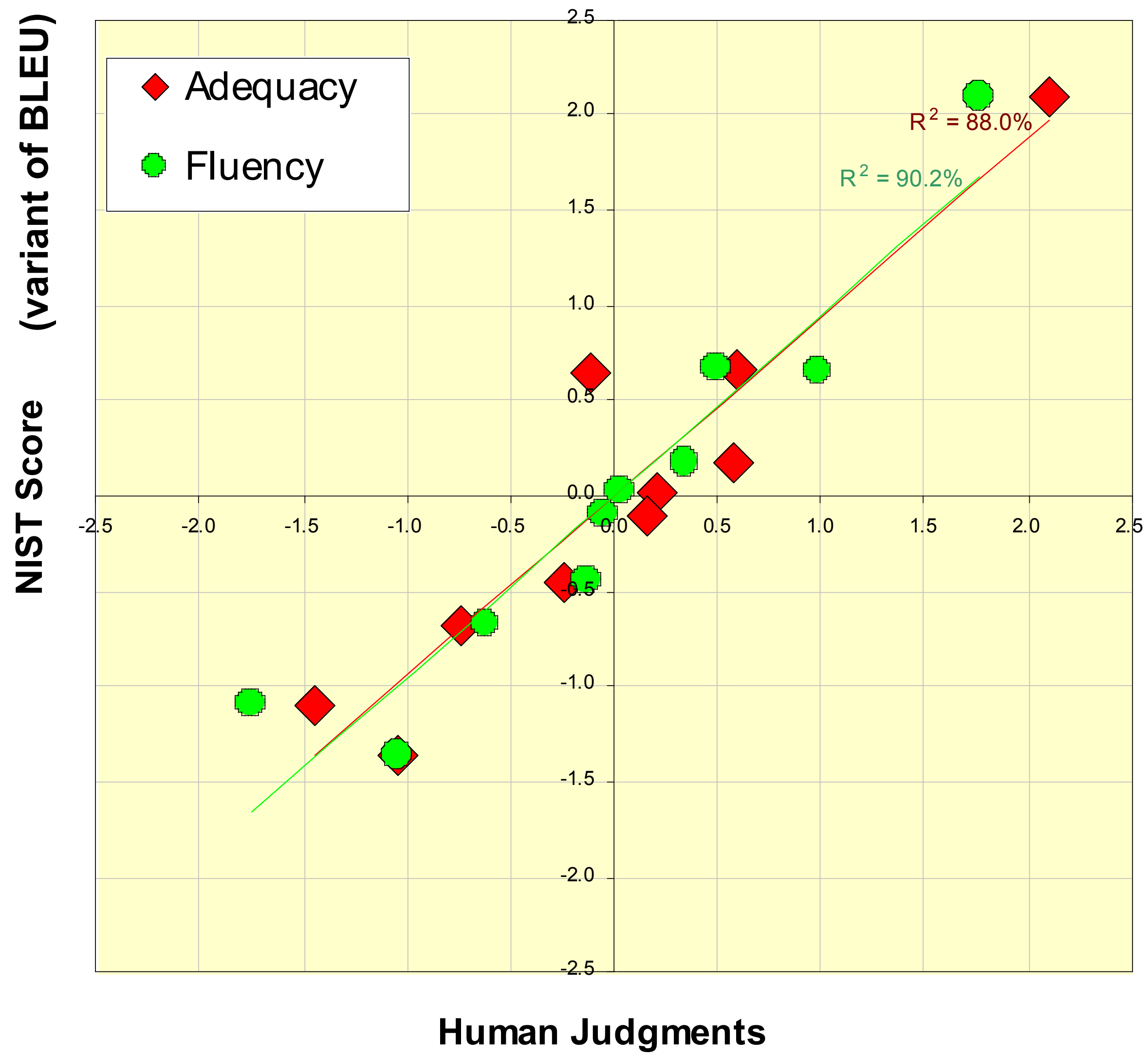
Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

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.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

Initial results showed that BLEU predicts human judgments well



slide from G. Doddington (NIST)