IBM Model 1
and
Machine Translation

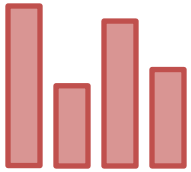
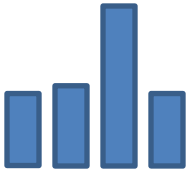
Recap

Expectation Maximization (EM)

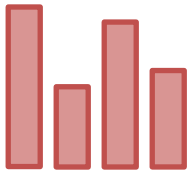
0. Assume *some* value for your parameters

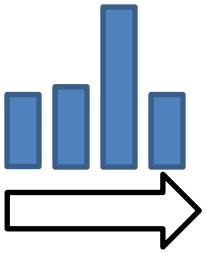
Two step, iterative algorithm

1. E-step: count under uncertainty, assuming these parameters

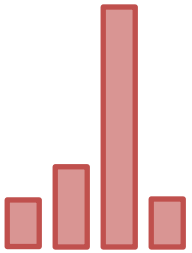
$p(z_i)$  $\text{count}(z_i, w_i)$ 

2. M-step: maximize log-likelihood, assuming these uncertain counts

$p^{(t)}(z)$ 



estimated counts

$p^{(t+1)}(z)$ 

Three Coins/Unigram With Class Example

Imagine three coins

Flip 1st coin (penny)

unobserved:
vowel or consonant?
part of speech?

If heads: flip 2nd coin (dollar coin)

If tails: flip 3rd coin (dime)

observed:
a, *b*, *e*, etc.
We run the code, vs.
The *run* failed

Three Coins/Unigram With Class Example

Imagine three coins

Flip 1st coin (penny)

$$p(\text{heads}) = \lambda \qquad p(\text{tails}) = 1 - \lambda$$

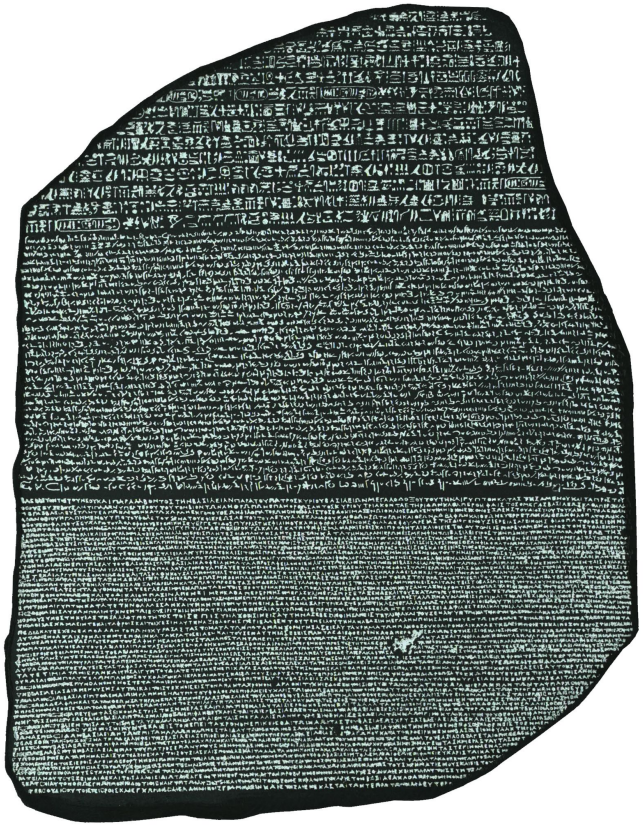
If heads: flip 2nd coin (dollar coin)

$$p(\text{heads}) = \gamma \qquad p(\text{tails}) = 1 - \gamma$$

If tails: flip 3rd coin (dime)

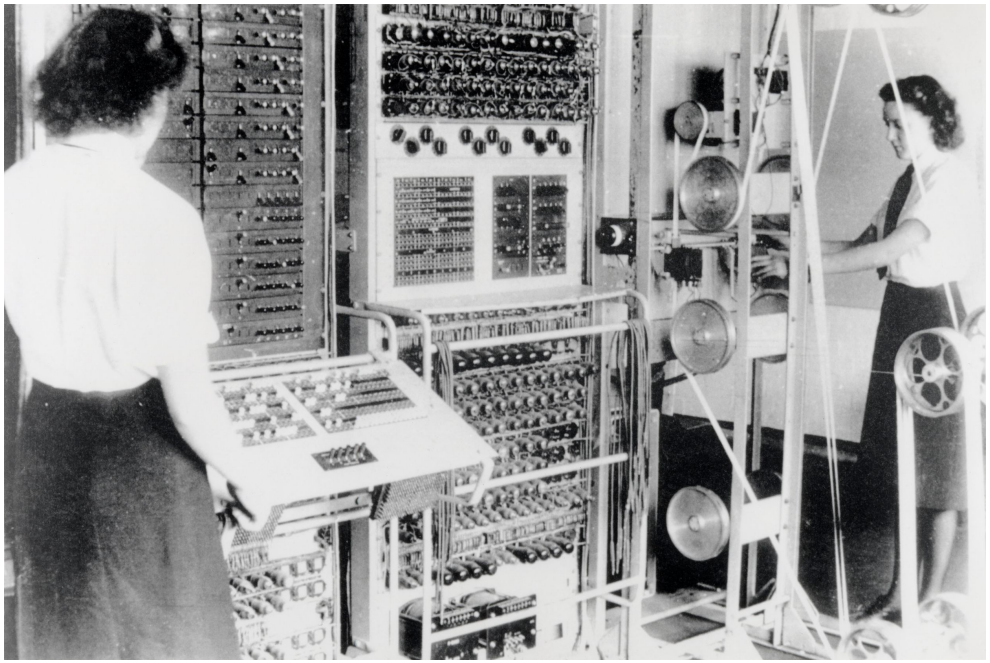
$$p(\text{heads}) = \psi \qquad p(\text{tails}) = 1 - \psi$$

Machine Translation



https://upload.wikimedia.org/wikipedia/commons/c/ca/Rosetta_Stone_BW.jpeg

Historical Context: World War II



From the National Archives (United Kingdom), via Wikimedia Commons, https://commons.wikimedia.org/wiki/File%3AColossus.jpg

By Antoine Taveneaux (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons https://commons.wikimedia.org/wiki/File%3ATuring-statue-Bletchley_14.jpg

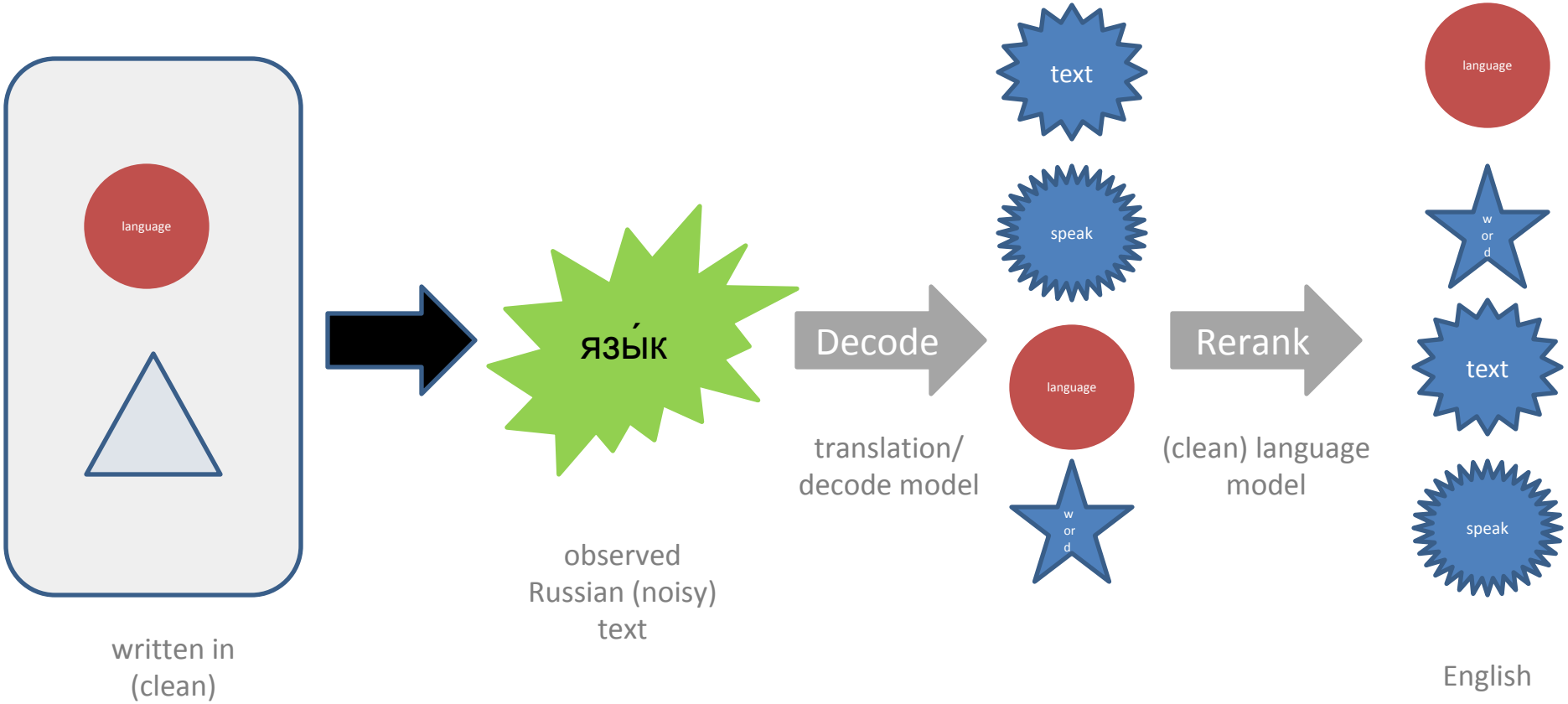
Warren Weaver's Note

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."

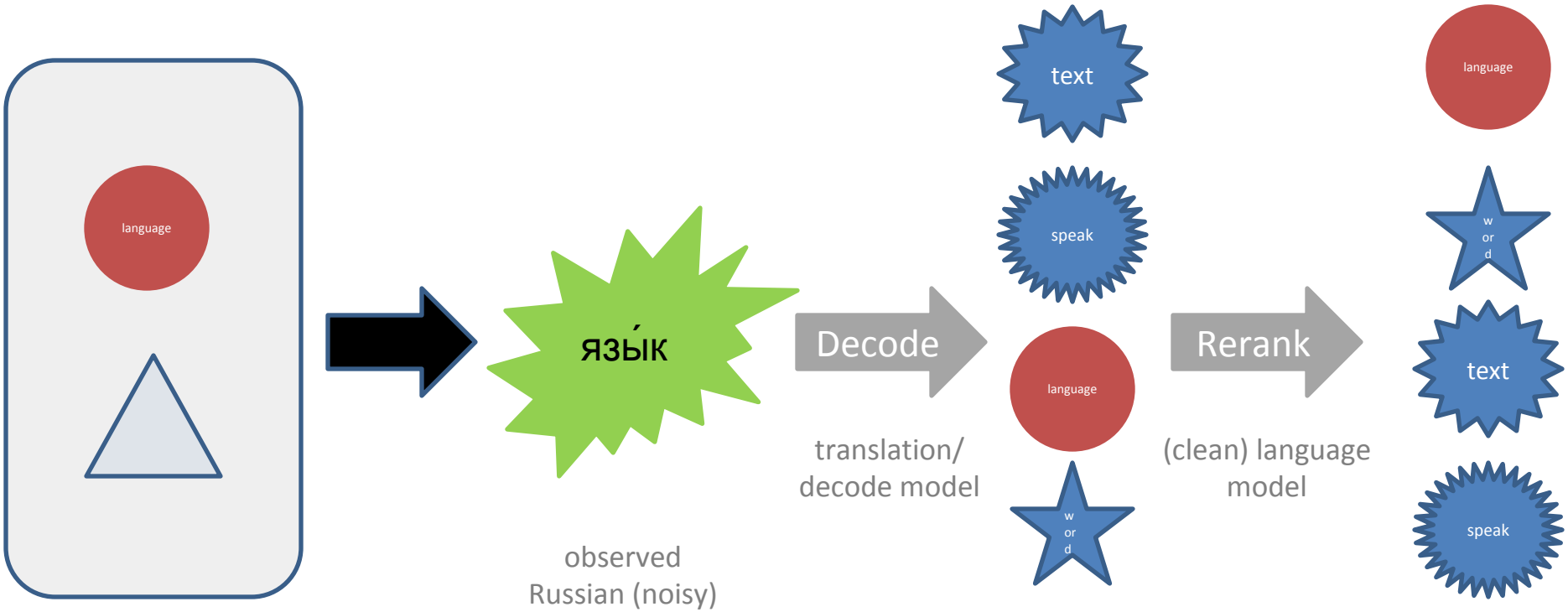
(Warren Weaver, 1947)

http://www.mt-archive.info/Weaver-1949.pdf

Noisy Channel Model

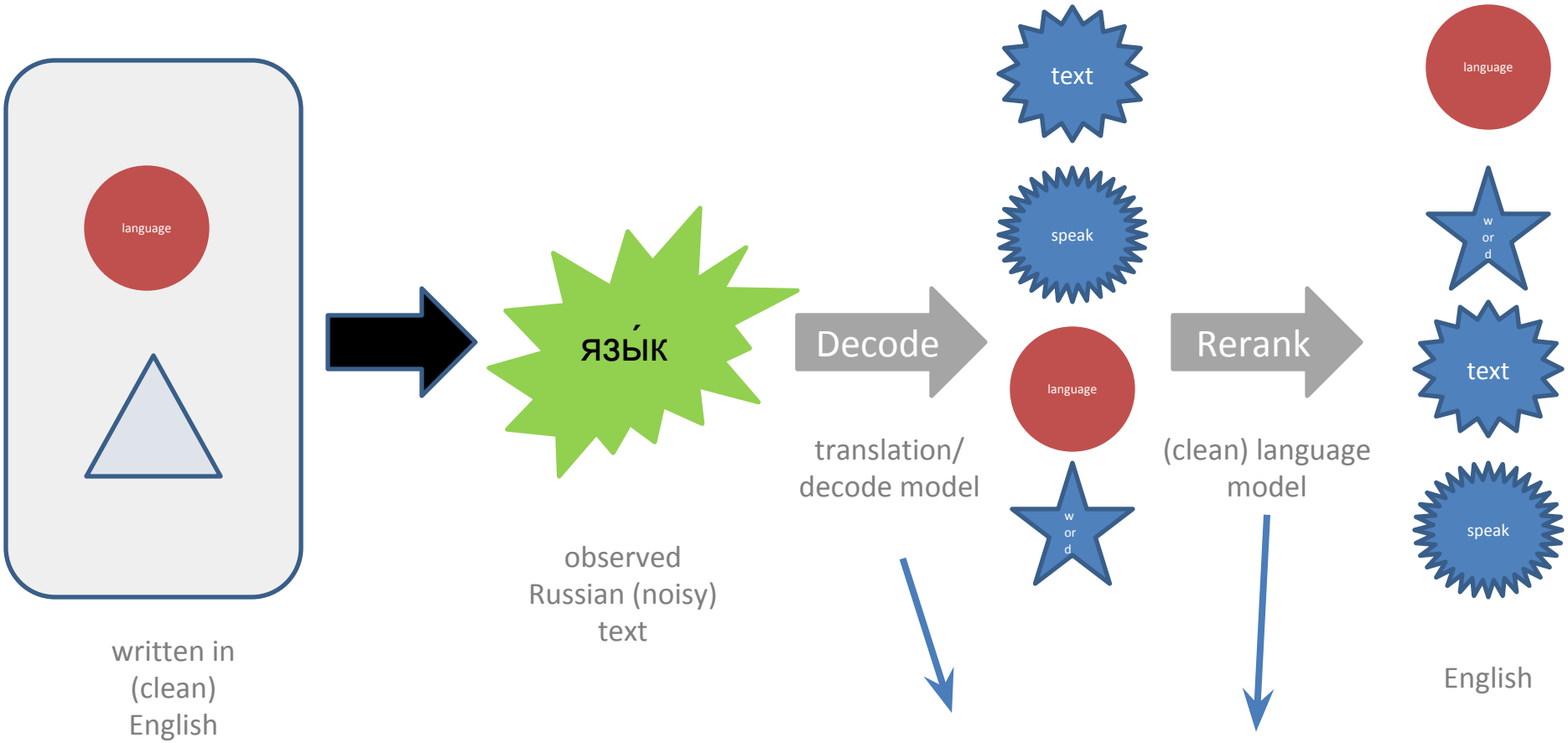


Noisy Channel Model



$$p(X \mid Y) \propto p(Y \mid X) * p(X)$$

Noisy Channel Model

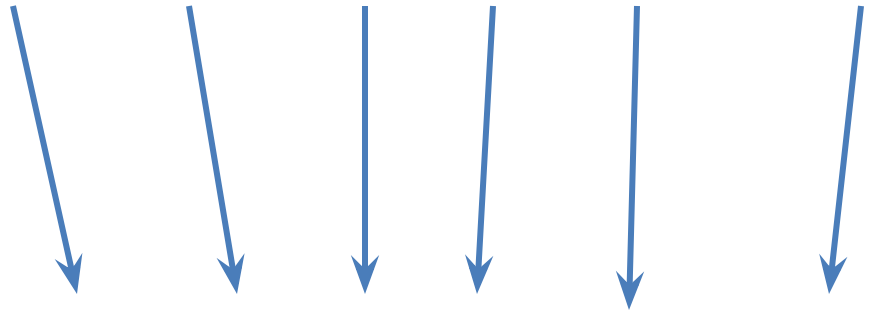


$$p(X \mid Y) \propto p(Y \mid X) * p(X)$$

Translation

Translate French (observed) into English:

Le chat est sur la chaise.



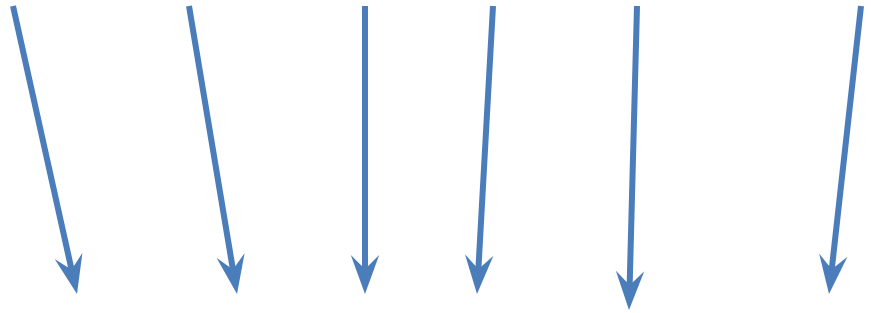
The cat is on the chair.

$$p(English|French)$$

Translation

Translate French (observed) into English:

Le chat est sur la chaise.



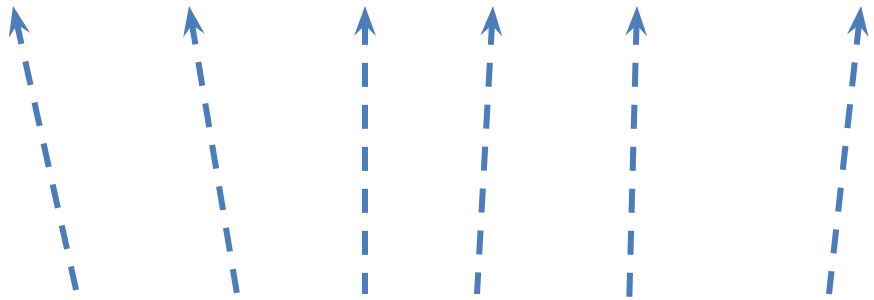
The cat is on the chair.

$$p(English|French) \propto p(French|English) * p(English)$$

Translation

Translate French (observed) into English:

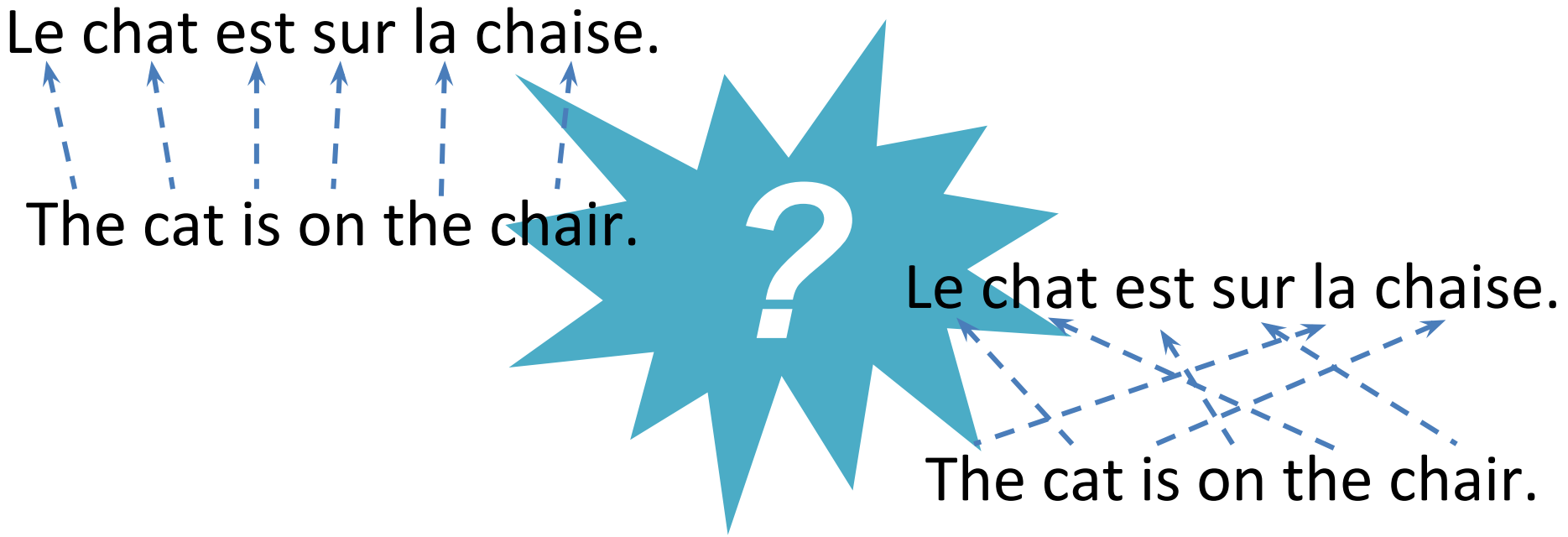
Le chat est sur la chaise.



The cat is on the chair.

$$p(English|French) \propto p(French|English) * p(English)$$

Alignment



$$p(English|French) \propto p(French|English) * p(English)$$

Parallel Texts

Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world,

Whereas disregard and contempt for human rights have resulted in barbarous acts which have outraged the conscience of mankind, and the advent of a world in which human beings shall enjoy freedom of speech and belief and freedom from fear and want has been proclaimed as the highest aspiration of the common people,

Whereas it is essential, if man is not to be compelled to have recourse, as a last resort, to rebellion against tyranny and oppression, that human rights should be protected by the rule of law,

Whereas it is essential to promote the development of friendly relations between nations,
…

http://www.un.org/en/universal-declaration-human-rights/

Yolki, pampa ni tlatepanitalotl, ni tlasenkauajkayotl iuan ni kuali nemilistli ipan ni tlalpan, yaya ni moneki moixmatis uan monemilis, ijkinoj nochi kuali tiitstosej ika touampoyouaj.

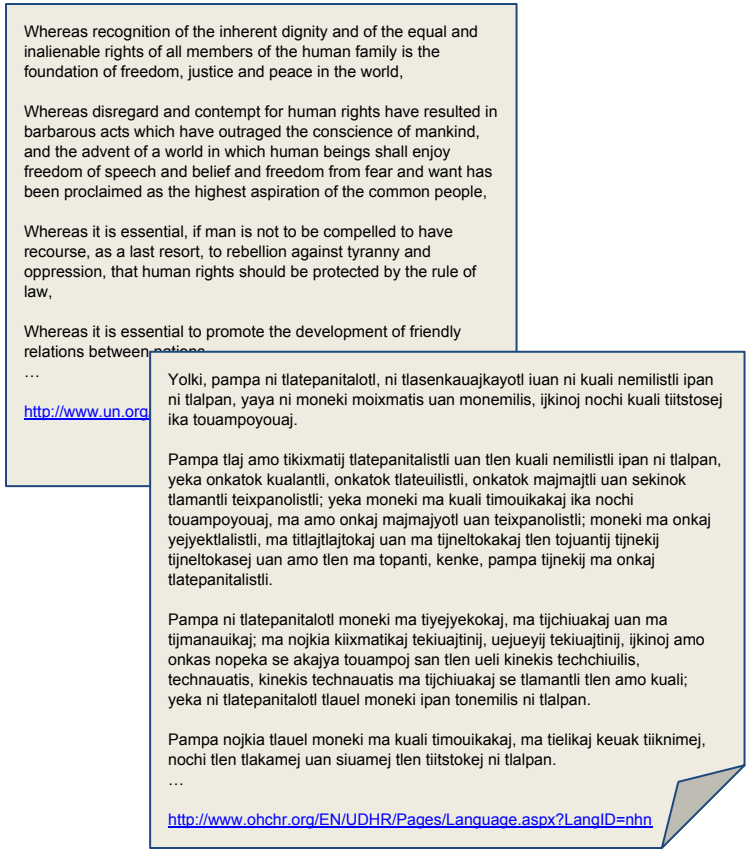
Pampa tlaj amo tikixmatij tlatepanitalistli uan tlen kuali nemilistli ipan ni tlalpan, yeka onkatok kualantli, onkatok tlateuilistli, onkatok majmajtli uan sekinok tlamantli teixpanolistli; yeka moneki ma kuali timouikakaj ika nochi touampoyouaj, ma amo onkaj majmajyotl uan teixpanolistli; moneki ma onkaj yejyektlalistli, ma titlajtlajtokaj uan ma tijneltokakaj tlen tojuantij tijnekij tijneltokasej uan amo tlen ma topanti, kenke, pampa tijnekij ma onkaj tlatepanitalistli.

Pampa ni tlatepanitalotl moneki ma tiyejyekokaj, ma tijchiuakaj uan ma tijmanauikaj; ma nojkia kiixmatikaj tekiuajtinij, uejueyij tekiuajtinij, ijkinoj amo onkas nopeka se akajya touampoj san tlen ueli kinekis techchiuilis, technauatis, kinekis technauatis ma tijchiuakaj se tlamantli tlen amo kuali; yeka ni tlatepanitalotl tlauel moneki ipan tonemilis ni tlalpan.

Pampa nojkia tlauel moneki ma kuali timouikakaj, ma tielikaj keuak tiiknimej, nochi tlen tlakamej uan siuamej tlen tiitstokej ni tlalpan.
…

http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=nhn

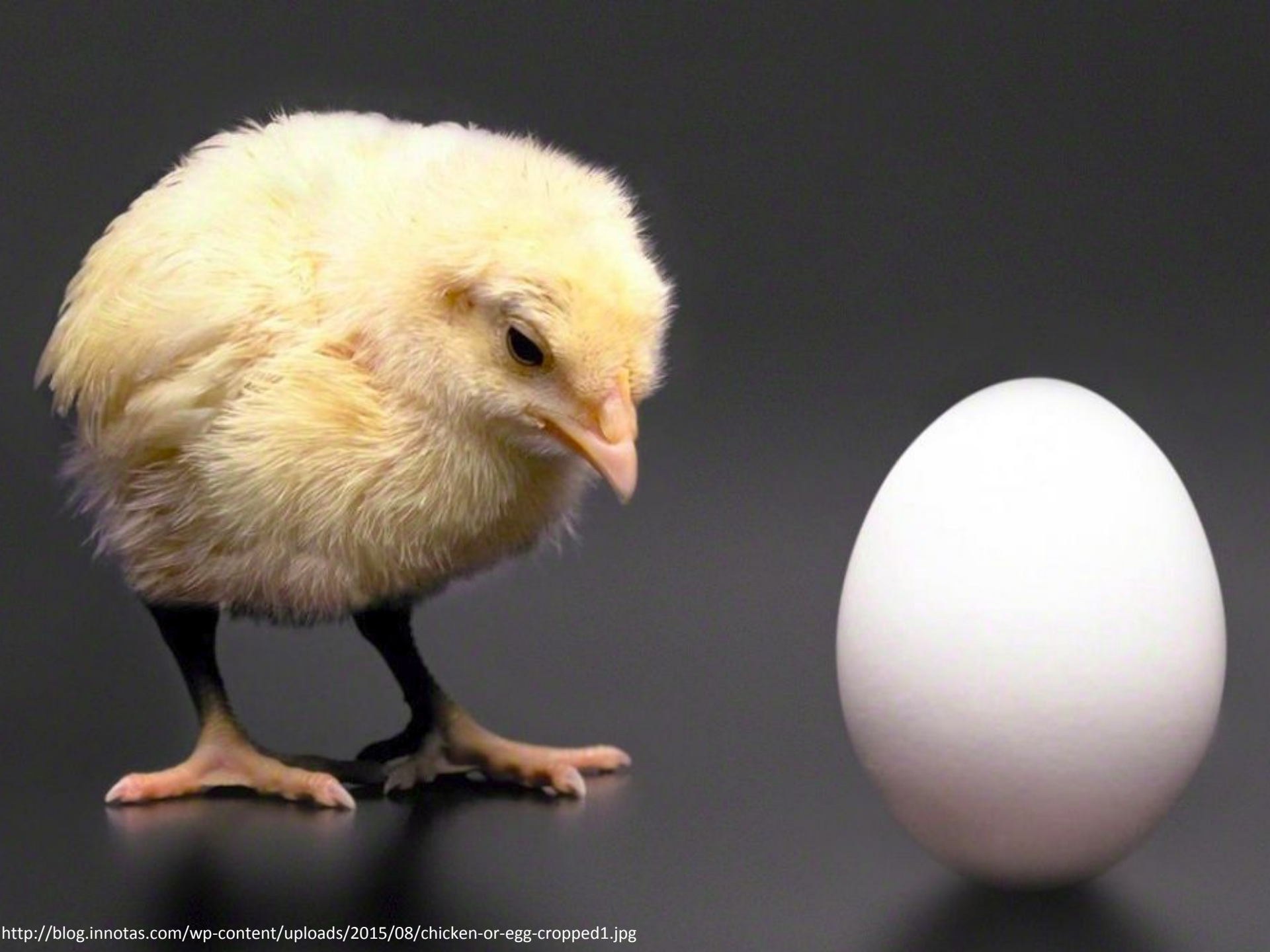
Preprocessing



- Sentence align
- Clean corpus
- Tokenize
- Handle case
- Word segmentation (morphological, BPE, etc.)
- Language-specific preprocessing (example: pre-reordering)
- …

Alignments

- If we had word-aligned text, we could easily estimate P($f|e$).
  - But we don't usually have word alignments, and they are expensive to produce by hand…
- If we had P($f|e$) we could produce alignments automatically.



http://blog.innotas.com/wp-content/uploads/2015/08/chicken-or-egg-cropped1.jpg

IBM Model 1 (1993)

- Lexical Translation Model
- Word Alignment Model
- The simplest of the original IBM models
- For all IBM models, see the original paper (Brown et al, 1993): http://www.aclweb.org/anthology/J93-2003

Simplified IBM 1

- We'll work through an example with a simplified version of IBM Model 1
- Figures and examples are drawn from A Statistical MT Tutorial Workbook, Section 27, (Knight, 1999)
- **Simplifying assumption:** each source word must translate to exactly one target word and vice versa

IBM Model 1 (1993)

- *f*: vector of French words

  (visualization of alignment)

- *e*: vector of English words

- *a*: vector of alignment indices

Le chat est sur la chaise verte

The cat is on the green chair

0 1 2 3 4 6 5

IBM Model 1 (1993)

- *f*: vector of French words
  (visualization of alignment)
- *e*: vector of English words

Le chat est sur la chaise verte

The cat is on the green chair

- *a*: vector of alignment indices

0 1 2 3 4 6 5

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j}) = t(f_1|e_{a_1}) \cdots t(f_m|e_{a_m})$$

- $t(f_j|e_i)$: translation probability of the word f_j given the word e_i

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

Model and Parameters

Want: P($f|e$)

But don't know how to train this directly…

Solution: Use P(a, $f|e$), where a is an alignment

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

Model and Parameters: Intuition

Translation prob.: $t(f_j|e_i)$

Example:

$$t(chaise|chair) > t(chaise|the)$$

Interpretation:

How probable is it that we see f_j given e_i

Model and Parameters: Intuition

Alignment/translation prob.: $P(a, f|e)$

Example (visual representation of *a*):

Interpretation:

How probable are the alignment *a* and the translation *f* (given *e*)

Model and Parameters: Intuition

Alignment prob.: $P(a|e,f)$

Example:

P( | “le chat”, “the cat”) **<** P( | “le chat”, “the cat”)

Interpretation:

How probable is alignment *a* (given *e* and *f*)

Model and Parameters

How to compute:

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j}) = t(f_1|e_{a_1}) \cdots t(f_m|e_{a_m})$$

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

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

Parameters

In the coin example, we had 3 parameters from which we could compute all others:

$$p(\text{heads}) = \lambda$$

$$p(\text{heads}) = \gamma$$

$$p(\text{heads}) = \psi$$

Parameters

For IBM model 1, we can compute all parameters given translation parameters:

$$t(f_j|e_i)$$

How many of these are there?

Parameters

For IBM model 1, we can compute all parameters given translation parameters:

$$t(f_j|e_i)$$

How many of these are there?

|*French vocabulary*| x |*English vocabulary*|

Data

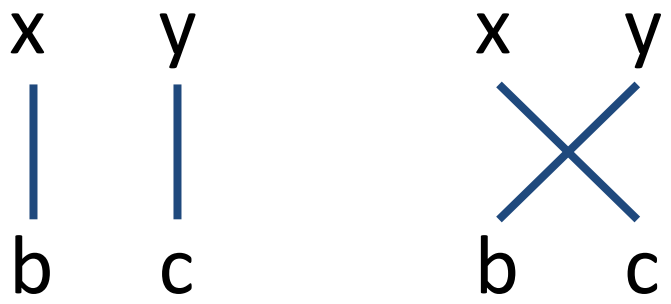
Two sentence pairs:

English	French
b c	x y
b	y

All Possible Alignments

(French: x, y)

(English: b, c)



Remember: simplifying assumption that each word must be aligned exactly once

Expectation Maximization (EM)

0. Assume *some* value for $t(f_j|e_i)$ and compute other parameter values

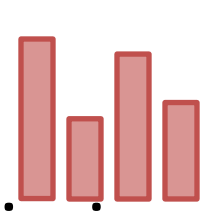
Two step, iterative algorithm

1. E-step: count alignments and translations under uncertainty, assuming these parameters

$t(f_j|e_i)$

$P(a|e,f)$

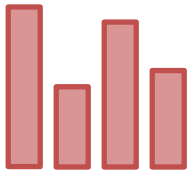
$P(a,f|e)$



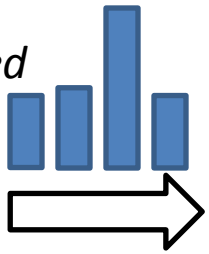
P(le chat | "the cat")

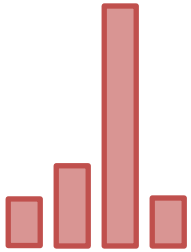
P(le chat | "the cat")

2. M-step: maximize log-likelihood (update parameters), using uncertain counts



estimated counts





EM Step 0: Initialize

Set parameter values uniformly.

All translations have an equal chance of happening.

$$t(x|b) = 1/2$$
$$t(y|b) = 1/2$$
$$t(x|c) = 1/2$$
$$t(y|c) = 1/2$$

E-step: Compute P(*a*,*f*|*e*)

For all alignments, compute P(*a*,*f*|*e*)

Remember:

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

$$t(x|b) = 1/2$$
$$t(y|b) = 1/2$$
$$t(x|c) = 1/2$$
$$t(y|c) = 1/2$$

P(x y | *b c*) = ?

P(x y | *b c*) =

P(y | *b*) = ½

E-step: Compute P(*a,f*|*e*)

For all alignments, compute P(*a,f*|*e*)

Remember:

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

$$t(x|b) = 1/2$$
$$t(y|b) = 1/2$$
$$t(x|c) = 1/2$$
$$t(y|c) = 1/2$$

P(x y | *b c*) = ½*½ = ¼

P(x y | *b c*) = ½*½ = ¼

P(y | *b*) = ½

E-step: Compute P(*a*|*e*,*f*)

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

P( | *b c*) = ¼

P( | *b c*) = ¼

P( | *b*) = ½

P( | *b c, x y*) = (¼)/(2/4) = ½

P(| *b c, x y*) = (¼)/(2/4) = ½

P(| *b*) = (½)/(½) = 1

Collect Counts: Example

P(| || *b c, x y*) = ½

P(X | *b c, x y*) = ½

P(| | *b, y*) = 1

Count instances where *b* and *y* are aligned:

ct(*y*|*b*) = ½ + 1

Collect Counts

P(| || *b c, x y*) = ½

P(X | *b c, x y*) = ½

P(| | *b, y*) = 1

$$tc(x|b) = 1/2$$
$$tc(y|b) = 1/2 + 1 = 3/2$$
$$tc(x|c) = 1/2$$
$$tc(y|c) = 1/2$$

M-step: Normalize

$$tc(x|b) = 1/2$$

$$tc(y|b) = 1/2 + 1 = 3/2$$

$$tc(x|c) = 1/2$$

$$tc(y|c) = 1/2$$

$$t(x|b) = \frac{1/2}{4/2} = 1/4$$

$$t(y|b) = \frac{3/2}{4/2} = 3/4$$

$$t(x|c) = \frac{1/2}{1} = 1/2$$

$$t(y|c) = \frac{1/2}{1} = 1/2$$

E-step (again!): P(*a,f*|*e*)

Compute P(*a,f*|*e*) using new parameters:

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

$$t(x|b) = \frac{1/2}{4/2} = 1/4$$

$$t(y|b) = \frac{3/2}{4/2} = 3/4$$

$$t(x|c) = \frac{1/2}{1} = 1/2$$

$$t(y|c) = \frac{1/2}{1} = 1/2$$

P(x y | *b c*) = ?

P(x y | *b c*) =

P(y | *b*) = 3/4

E-step (again!): P(*a,f*|*e*)

Compute P(*a,f*|*e*) using new parameters:

$$P(a, f|e) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

$$t(x|b) = \frac{1/2}{4/2} = 1/4$$

$$t(y|b) = \frac{3/2}{4/2} = 3/4$$

$$t(x|c) = \frac{1/2}{1} = 1/2$$

$$t(y|c) = \frac{1/2}{1} = 1/2$$

P(x y | *b c*) = ¼*½ = ⅛

P(x y | *b c*) = ½*¾ = 3/8

P(y | *b*) = 3/4

E-step (again): Compute P(*a* | *e*,*f*)

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

P(x y | | | *b c*) = ⅛

P(x y | *b c*) = 3/8

P(y | | *b*) = 3/4

P(| | | *b c, x y*) = (⅛)/(4/8) = ¼

P(X | *b c, x y*) = (⅜)/(4/8) = ¾

P(| | *b*) = (¾)/(¾) = 1

Collect Counts (again)

P(| || *b c, x y*) = ¼

P(X | *b c, x y*) = ¾

P(| | *b ,y*) = 1

$$tc(x|b) = 1/4$$
$$tc(y|b) = 3/4 + 1 = 7/4$$
$$tc(x|c) = 3/4$$
$$tc(y|c) = 1/4$$

M-step (again): Normalize Counts

Collected counts:

$$tc(x|b) = 1/4$$
$$tc(y|b) = 3/4 + 1 = 7/4$$
$$tc(x|c) = 3/4$$
$$tc(y|c) = 1/4$$

Normalized counts:

$$t(x|b) = 1/8$$
$$t(y|b) = 7/8$$
$$t(x|c) = 3/4$$
$$t(y|c) = 1/4$$

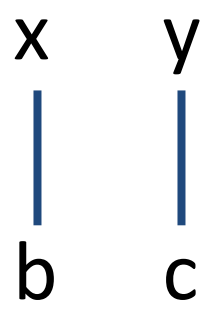
What is happening to $t(f_j|e_i)$?

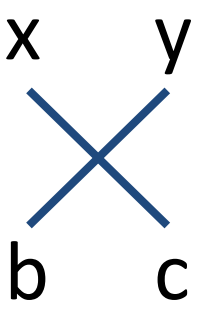
$$t(x|b) = 1/2$$
$$t(y|b) = 1/2$$
$$t(x|c) = 1/2$$
$$t(y|c) = 1/2$$

$$t(x|b) = \frac{1/2}{4/2} = 1/4$$
$$t(y|b) = \frac{3/2}{4/2} = 3/4$$
$$t(x|c) = \frac{1/2}{1} = 1/2$$
$$t(y|c) = \frac{1/2}{1} = 1/2$$

$$t(x|b) = 1/8$$
$$t(y|b) = 7/8$$
$$t(x|c) = 3/4$$
$$t(y|c) = 1/4$$

What does that mean?

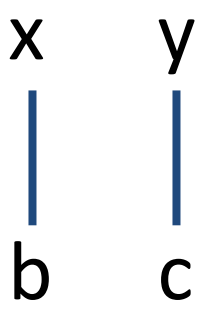


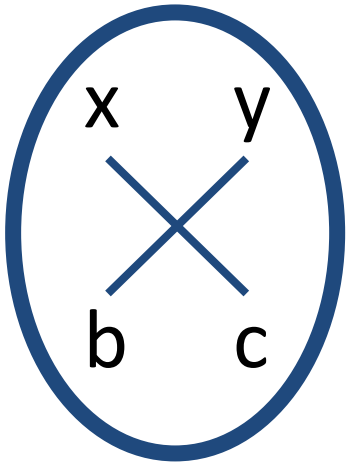


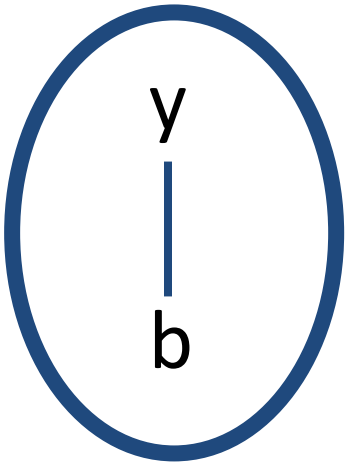
Which alignments are more likely to be correct?

$$t(x|b) = 1/8$$
$$t(y|b) = 7/8$$
$$t(x|c) = 3/4$$
$$t(y|c) = 1/4$$

What does that mean?







Which alignments are more likely to be correct?

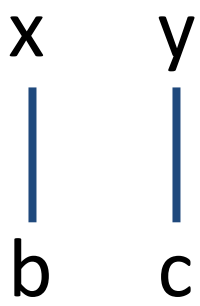
$$t(x|b) = 1/8$$
$$t(y|b) = 7/8$$
$$t(x|c) = 3/4$$
$$t(y|c) = 1/4$$

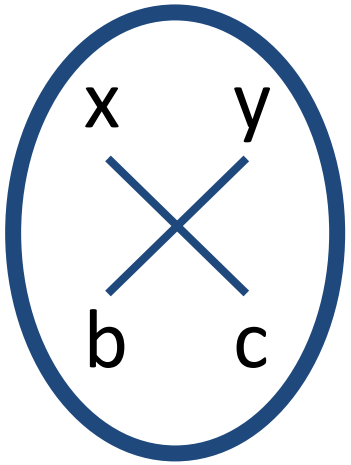
What would happen to $t(f_j|e_i)$…

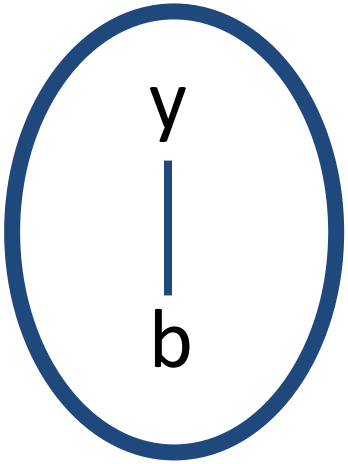
if we repeated these steps many times?

Many Iterations of EM:

$$t(x|b) = 0.0001$$
$$t(y|b) = 0.9999$$
$$t(x|c) = 0.9999$$
$$t(y|c) = 0.0001$$







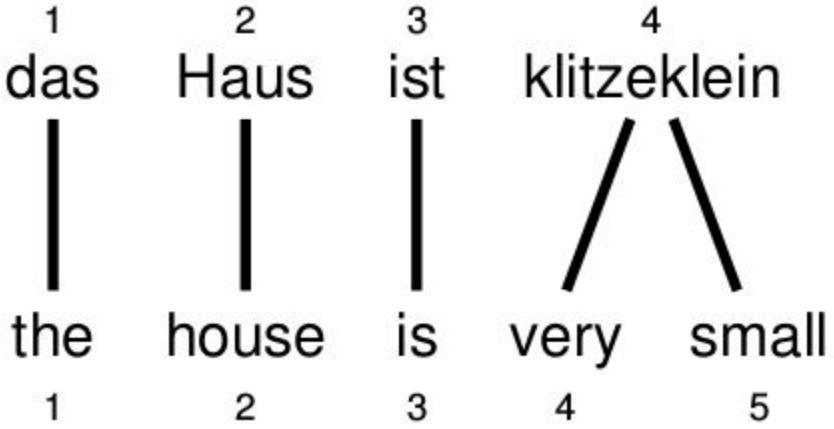
Review of IBM Model 1 & EM

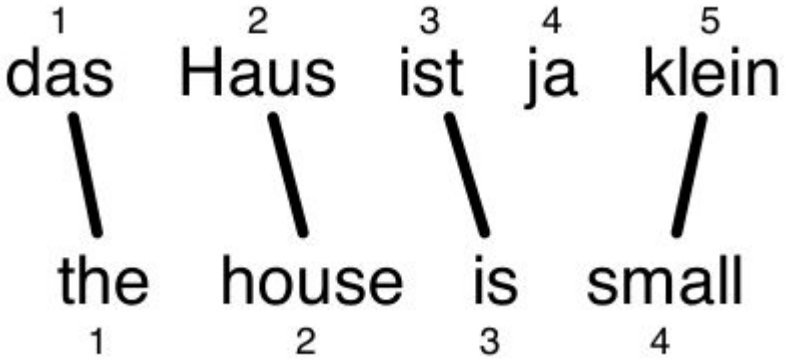
- Iteratively learned an alignment/translation model from sentence-aligned text (without "gold standard" alignments)
- Model can now be used for alignment and/or word-level translation
- We explored a simplified version of this; IBM Model 1 allows more types of alignments

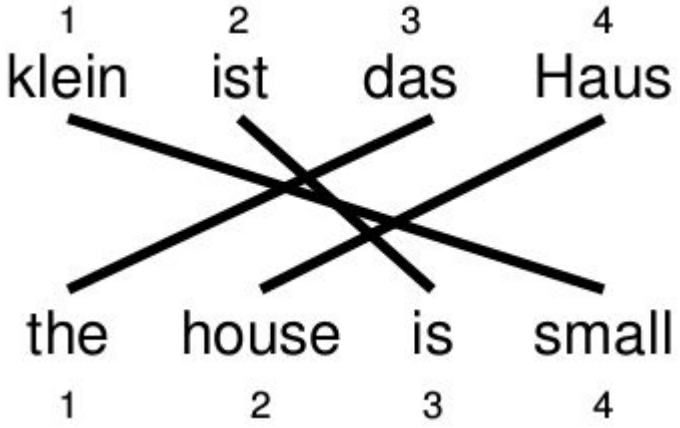
Uses for Alignments

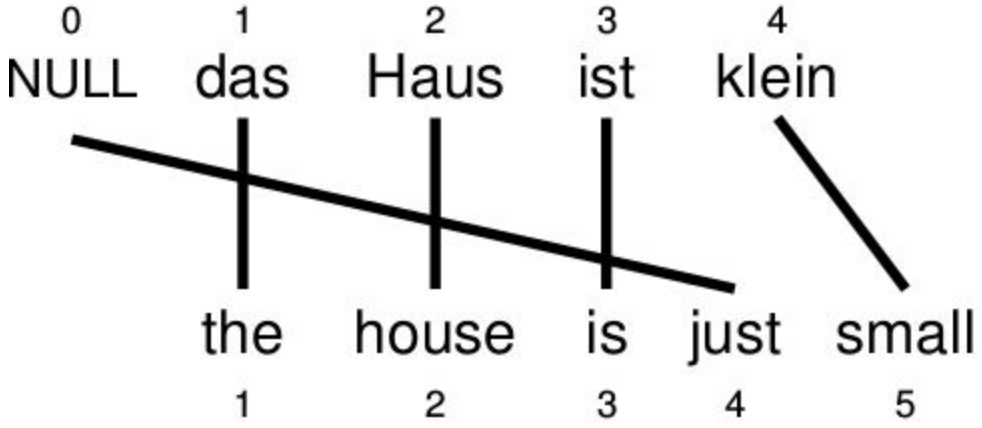
- Component of machine translation systems
- Produce a translation lexicon automatically
- Cross-lingual projection/extraction of information
- Supervision for training other models (for example, neural MT systems)

Alignment Examples (English-> German)









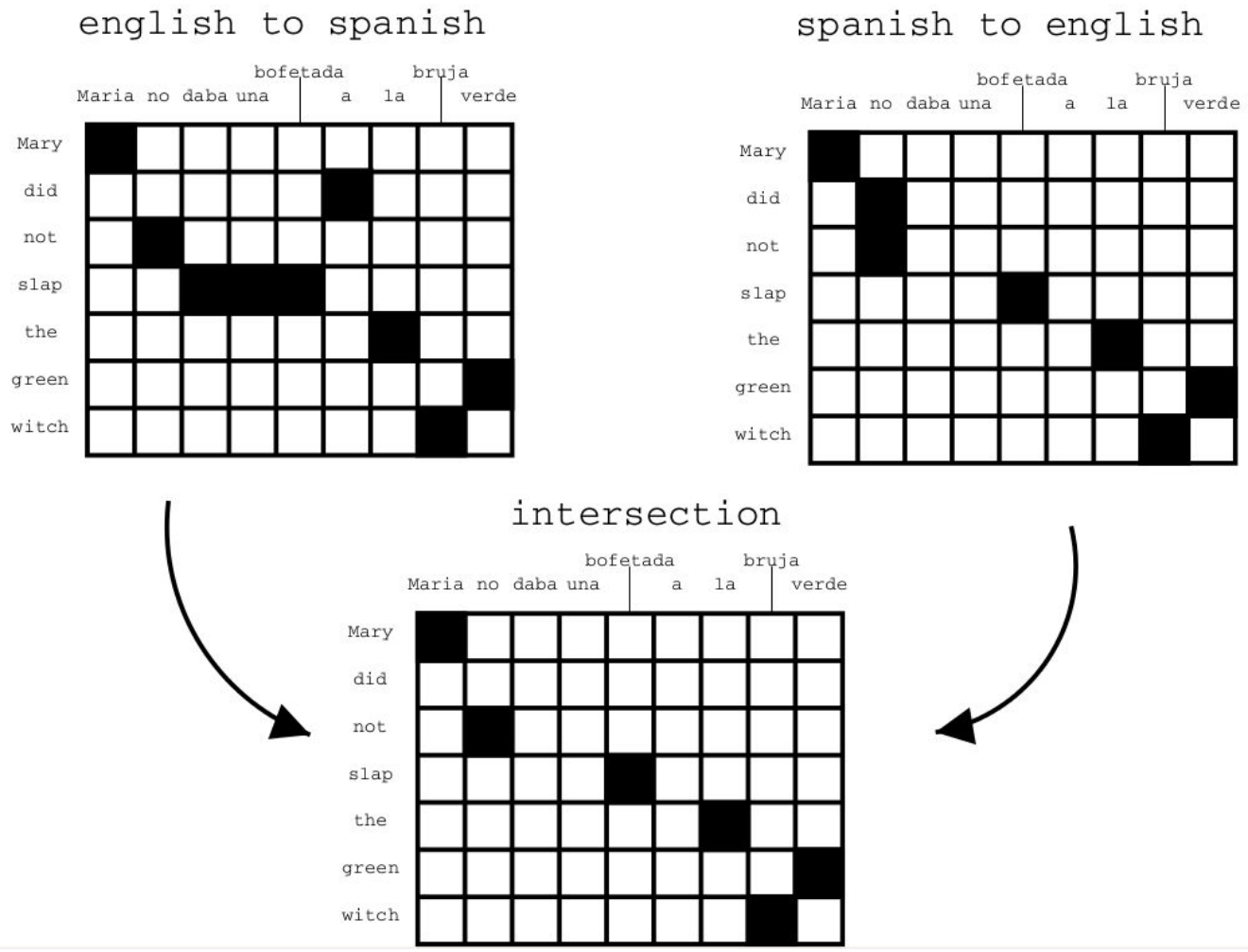
Why is Model 1 insufficient?

- Why won't this produce great translations?

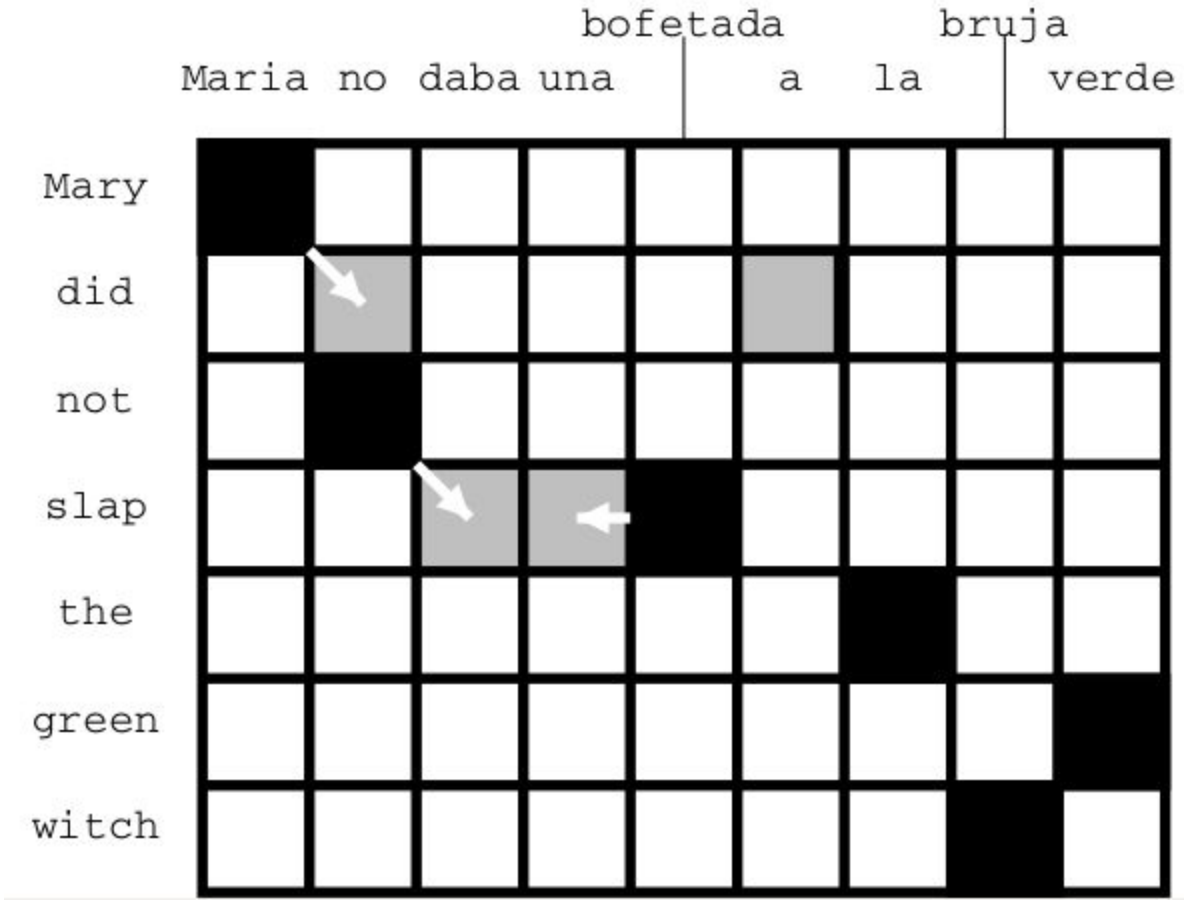
Why is Model 1 insufficient?

- Why won't this produce great translations?
  - Indifferent to order (language model may help?)
  - Translates one word at a time
  - Translates each word in isolation
  - …

Phrases


english to spanish
Maria no daba una bofetada a la bruja verde
Mary
did
not
slap
the
green
witch
spanish to english
Maria no daba una bofetada a la bruja verde
Mary
did
not
slap
the
green
witch
intersection
Maria no daba una bofetada a la bruja verde
Mary
did
not
slap
the
green
witch

Phrases


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

Decoding

What have we done so far?

- We can score alignments.
- We can score translations.

How do we *generate* translations?

- Decoding!

Decoding

Why can’t we just score all possible translations?

What do we do instead?

Decoding

- Many translation options for a word/phrase.
- Decoding is NP-complete (can verify solutions in polynomial time; can't locate solutions efficiently).
- We use heuristics to limit the search space.
- See: statmt.org/book/slides/06-decoding.pdf

The role of the decoder is to:

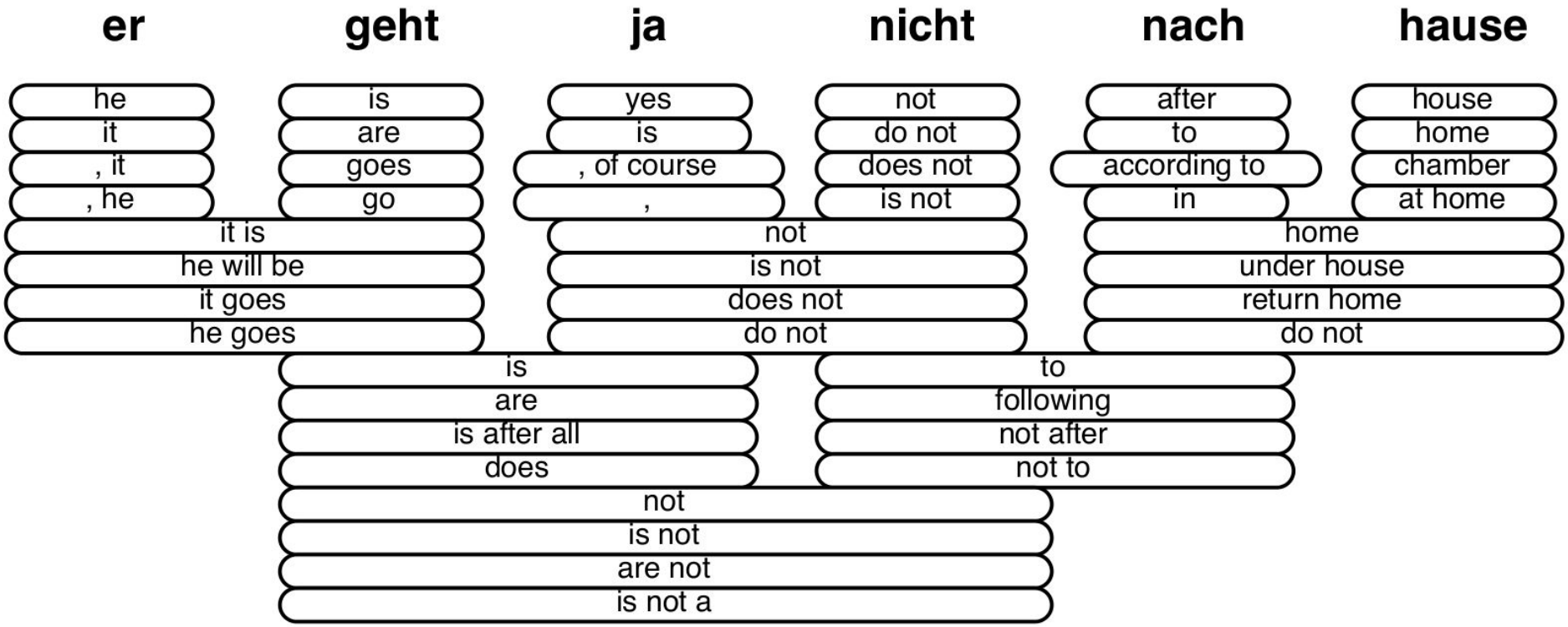
- Choose "good" translation options
- Arrange them in a "good" order

How can this go wrong?

- Search doesn’t find the best translation
  - Need to fix the search

- The best translation found is not good
  - Need to fix the model

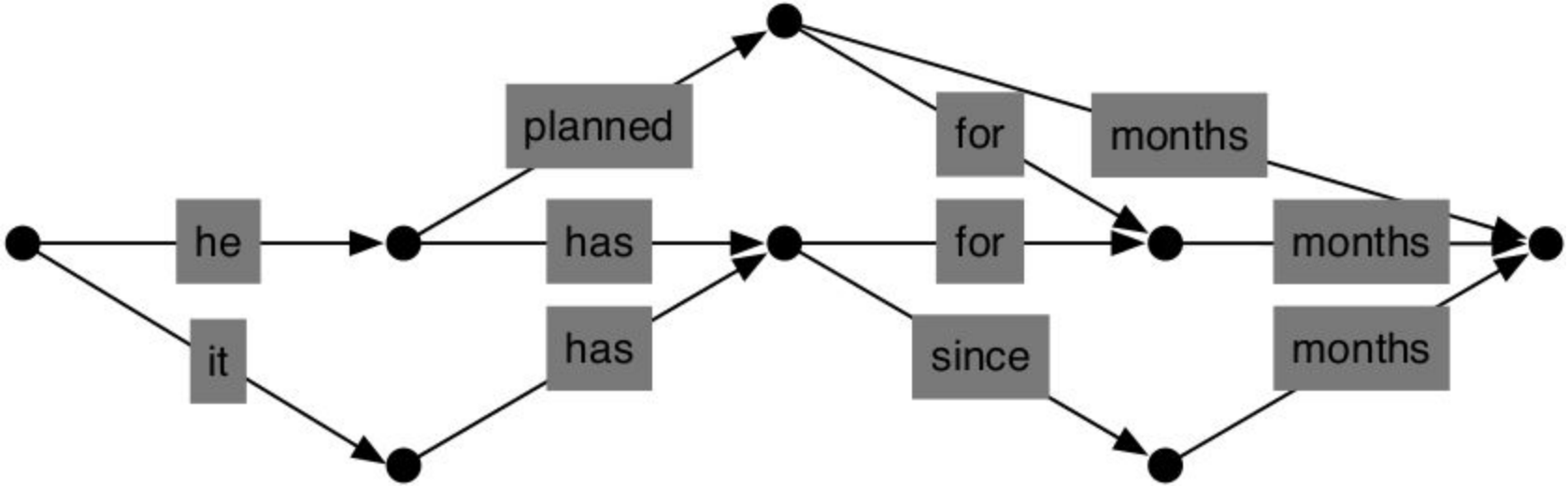
Decoding Options



In this example from Koehn (2017) slides, there might be >2700 phrase pairs for this (short!) sentence. (http://mt-class.org/jhu/slides/lecture-decoding.pdf)

Search Graph

Er hat seit Monaten geplant.

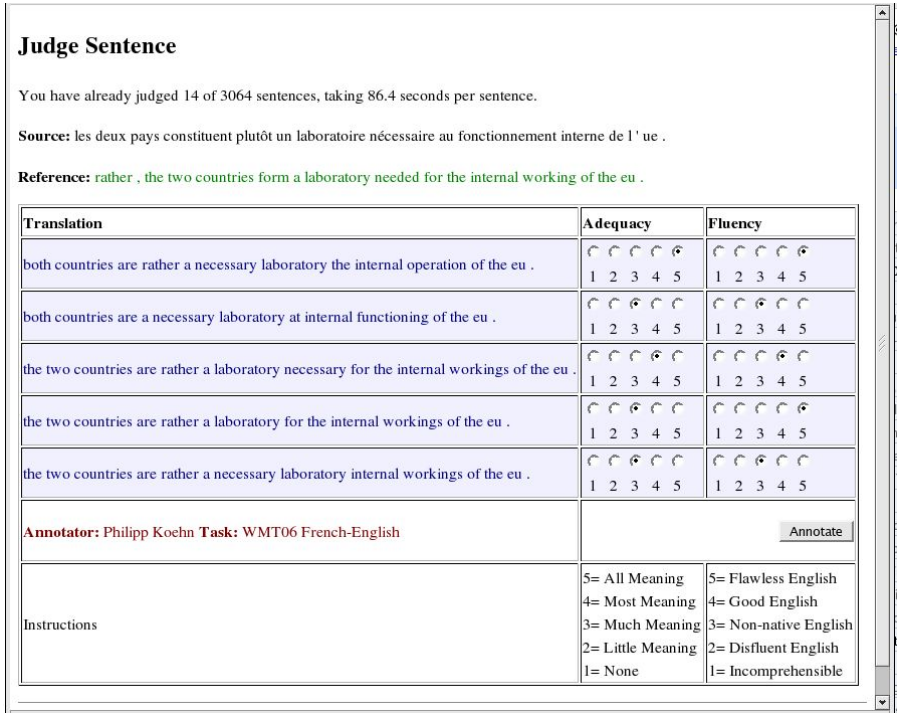


Koehn CAT slides (2016)

Evaluating Machine Translation

Human evaluations:

- Test set (source, human reference translations, MT output)
- Humans judge the quality of MT output (in one of several possible ways)



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Koehn (2017), http://mt-class.org/jhu/slides/lecture-evaluation.pdf

Evaluating Machine Translation

Automatic evaluations:

- Test set (source, human reference translations, MT output)
- Aim to mimic (correlate with) human evaluations

Many metrics:

- TER (Translation Error/Edit Rate)
- HTER (Human-Targeted Translation Edit Rate)
- BLEU (Bilingual Evaluation Understudy)
- METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Computer Aided Translation

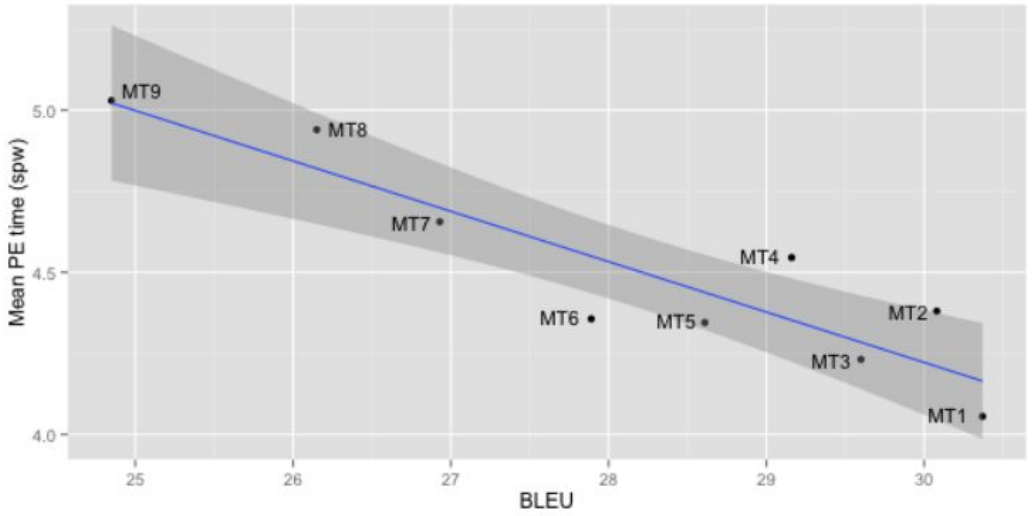
Interactive Translation Prediction

In addition to this, there are more than 18 tailing heaps {a4}located right in the city{/a4}, which has caused serious health impacts":

Zusätzlich zu diesen gibt es

mehr als 18

Post-Editing



Sanchez-Torron & Koehn (2016), http://www.cs.jhu.edu/~phi/publications/machine-translation-quality.pdf

Questions?