Statistical & Phrase Based Machine Translation

Rebecca Knowles

JOHNS HOPKINS UNIVERSITY

Today

- Word-based and phrase-based models
- A *very fast* introduction to PBMT/SMT
  - Entire courses can/have been taught on this topic
  - We'll cover several topics at a high level
  - See textbook: http://www.statmt.org/book/
- These and similar approaches are among the dominant MT paradigms of the last ~20 years

Outline

- Language models
  - *n*-grams
- Translation models
  - Word-level
  - Alignments
  - EM Algorithm
  - Phrases
- Decoding

From Data to Models

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

- Collect data
- Preprocess data
- Design model(s)/features
- Train model(s)
- Tune model(s)
- Evaluate model(s)
- Run model(s)

Review (from reading)

The next several slides draw on the reading:

A Statistical MT Tutorial Workbook

(Knight, 1999)

Basic Probability

- P(e): *A priori* probability (chance that e happens)
- P($f|e$): *Conditional* probability (chance of f given e)
- P(e, f): *Joint* probability (chance of e and f both happening)

Basic Probability

- P(e): What is the chance of e being an English sentence?
- P(f|e): What is the chance of the translator producing f given the English sentence e?

Review (from reading)

Put these words in the correct English order:

have programming a seen never I language better

Review (from reading)

Put these words in the correct English order:

have programming a seen never I language better

I have never seen a better programming language

- What kind of knowledge are you applying here?
- Do you think a machine could do this job?
- Can you think of a way to automatically test how well a machine is doing, without a lot of human checking?

Language Models (LM)

$P(e)$

P(the cat purrs) > P(purrs cat the)

Intuition

Decipher this student's late note:

(one word per column)

LIE MY GUY BYE	CHARM SOLD CLAM BEAM ALARM DREAM	CODE CIRCLE SHUTE CLOCK VISIT	SOIL DID RAID RISK MUST	ROUT HOT RIOT NOT	WAKE MAKE SALE WANE	HE I'LL ME BE SEE	US IS UGH UP	THIS TAXIS THAI TEAR	MOVING HAVING READING MORNING LOVING

Adapted from: http://nacloweb.org/resources/problems/sample/Handwriting-solution.pdf

Intuition

Decipher this student's late note:

(one word per column)

LIE	CHARM	CODE	SOIL	ROUT	**WAKE**	HE	US	**THIS**	MOVING
MY	SOLD	CIRCLE	**DID**	HOT	MAKE	I'LL	IS	TAXIS	HAVING
GUY	CLAM	SHUTE	RAID	RIOT	SALE	**ME**	UGH	THAI	READING
BYE	BEAM	**CLOCK**	RISK	**NOT**	WANE	BE	**UP**	TEAR	**MORNING**
	ALARM	VISIT	MUST			SEE			LOVING
	DREAM								

Adapted from: http://nacloweb.org/resources/problems/sample/Handwriting-solution.pdf

n-grams

A sequence of n *words is an* **n-gram.**

unigram: “word”

bigram: “a bigram”

trigram: “read a trigram”

4-gram: “this is an example”

...

Trigram Language Model

b(z|x, y) = count("x y z")/count("x y")

P(the cat purrs) ≅ b(the|BOS, BOS)*
b(cat | the, BOS)*
b(purrs | the, cat)*
b(EOS | cat, purrs)

We can estimate counts from a large monolingual corpus!

What am I glossing over?

- Smoothing (see A Statistical MT Tutorial Workbook (Knight, 1999) section 11)
- Markov assumption
- Chain rule of probability
- Neural LMs
- See these slides for much more information: http://statmt.org/book/slides/07-language-models.pdf

Translation Models

$$P(f|e)$$

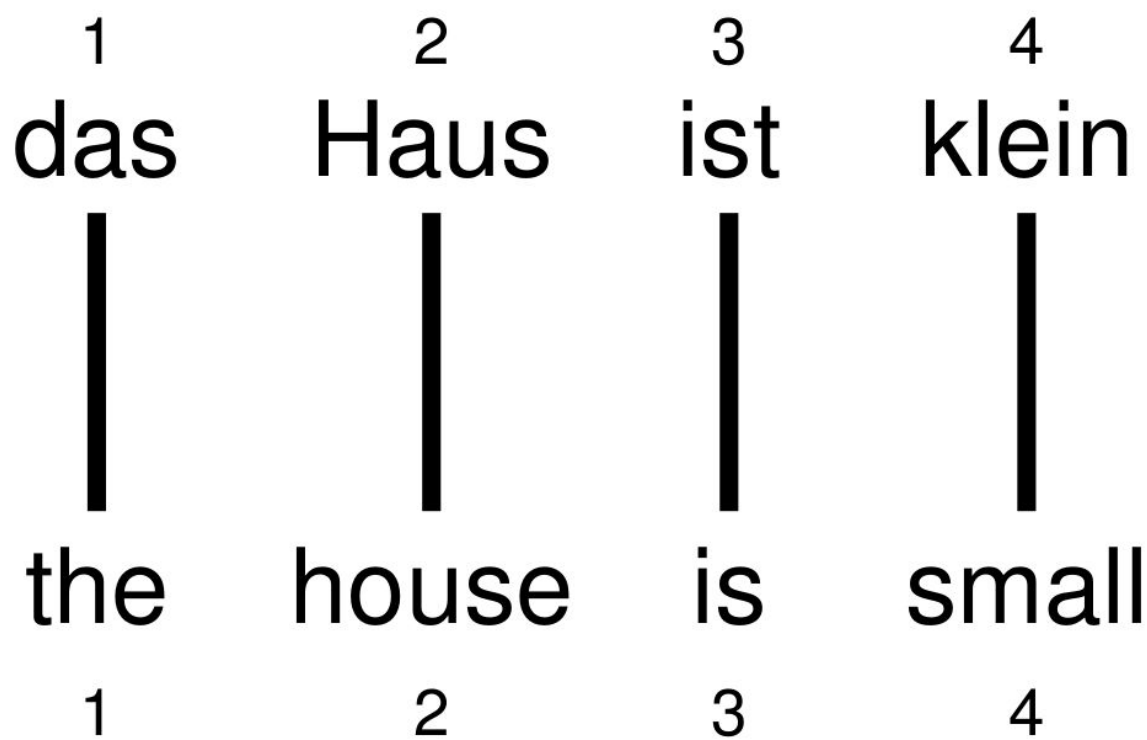
$$P(\text{Haus}|\text{house}) > P(\text{Haus}|\text{building}) > P(\text{Haus}|\text{the})$$

$P(f|e)$

- How can we find out $P(f|e)$?

P(*f*|*e*)

- How can we find out P(*f*|*e*)? *Alignments*.



Koehn (2017): http://mt-class.org/jhu/slides/lecture-ibm-model1.pdf

Alignments

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

IBM Model 1 (1993)

- Lexical Translation Model
- Word Alignment Model
- http://www.aclweb.org/anthology/J93-2003

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

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

Simplified IBM 1

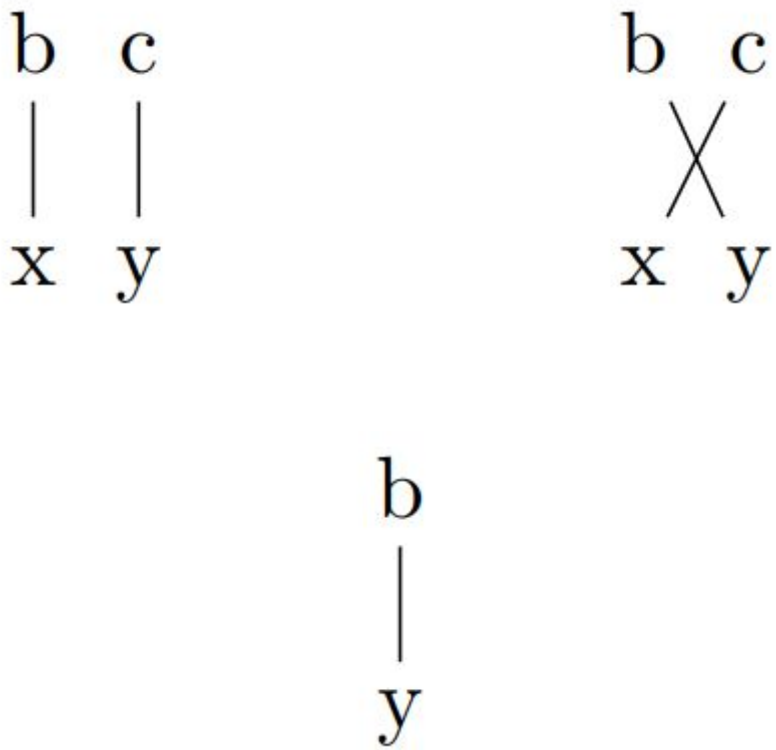
- We'll work through an example with a simplified version of IBM Model 1
- Figures and examples are drawn from A Statistical MT Tutorial Workbook, Section 27, (Knight, 1999)
- **Simplifying assumption:** each source word must translate to exactly one target word

Data

Two sentence pairs:

English	French
b c	x y
b	y

Possible Alignments



EM Algorithm

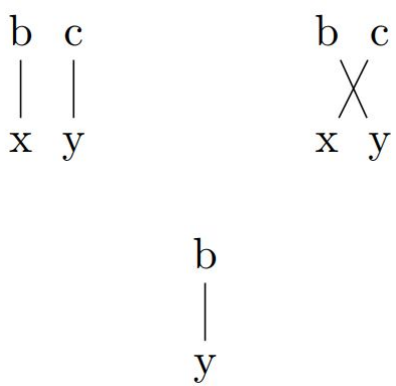
A. Initialize model parameters.
B. **E**xpectation: Assign probabilities to data.
C. **M**aximization: Re-estimate model from data.
D. Repeat B and C until convergence.

This is what we might describe as “training” or “learning” the model’s parameters.

EM Step 1: Initialize

Set parameter values uniformly.

All translations have an equal chance of happening.



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

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

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

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

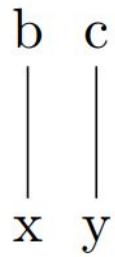
Step 2: Compute P(*a*,*f*|*e*)

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

Remember:

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

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



$P(a, f|e) =$?

$P(a, f|e) =$?

$P(a, f|e) = 1/2$

Step 2: Compute P(*a*,*f*|*e*)

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

Remember:

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

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

$$P(a, f|e) = 1/2 * 1/2 = 1/4$$

$$P(a, f|e) = 1/2 * 1/2 = 1/4$$

$$P(a, f|e) = 1/2$$

Step 3: Compute P(*a*|*e*,*f*)

We can compute P(*a*|*e*,*f*) using P(*a*,*f*|*e*):

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

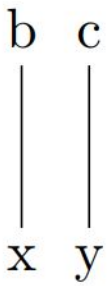
Step 3: Compute P(*a*|*e*,*f*)

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

$P(a,f|e) = 1/2 * 1/2 = 1/4$

$P(a,f|e) = 1/2 * 1/2 = 1/4$

$P(a,f|e) = 1/2$



$P(a|e,f) = \frac{1/4}{2/4} = 1/2$

$P(a|e,f) = \frac{1/4}{2/4} = 1/2$

$P(a|e,f) = \frac{1/2}{1/2} = 1$

Step 4: Collect Counts

$$P(a|e,f) = \frac{1/4}{2/4} = 1/2$$

$$P(a|e,f) = \frac{1/4}{2/4} = 1/2$$

$$P(a|e,f) = \frac{1/2}{1/2} = 1$$

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

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

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

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

Step 5: Normalize

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

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

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

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

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

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

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

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

Repeat Step 2

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

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

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

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

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

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

$P(a, f|e) =$?

$P(a, f|e) =$?

$P(a, f|e) = 3/4$

Repeat Step 2

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

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

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

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

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

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

$$P(a, f|e) = 1/4 * 1/2 = 1/8$$

$$P(a, f|e) = 3/2 * 1/2 = 3/8$$

$$P(a, f|e) = 3/4$$

Repeat Step 3: Compute P(*a*|*e*,*f*)

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

$P(a,f|e) = 1/4 * 1/2 = 1/8$

$P(a,f|e) = 3/2 * 1/2 = 3/8$

$P(a,f|e) = 3/4$

$P(a|e,f) = \frac{1/8}{4/8} = 1/4$

$P(a|e,f) = \frac{3/8}{4/8} = 3/4$

$P(a|e,f) = 1$

Repeat Steps 4 & 5

Collect counts:

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

Normalize counts:

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

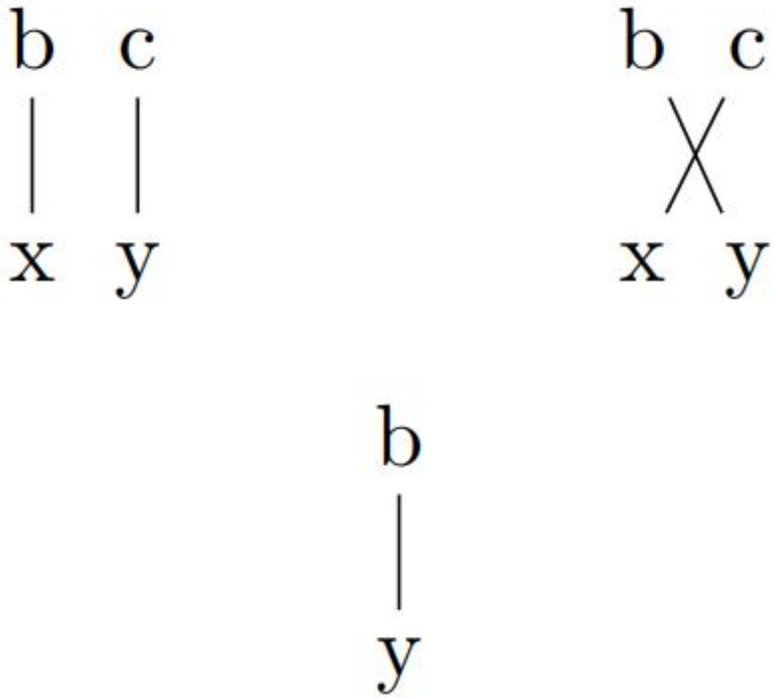
What is happening to t()?

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

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

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

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 would happen to t()...

...if we repeated steps 2-5 many many times?

Repeat Steps 2-5 Many Times

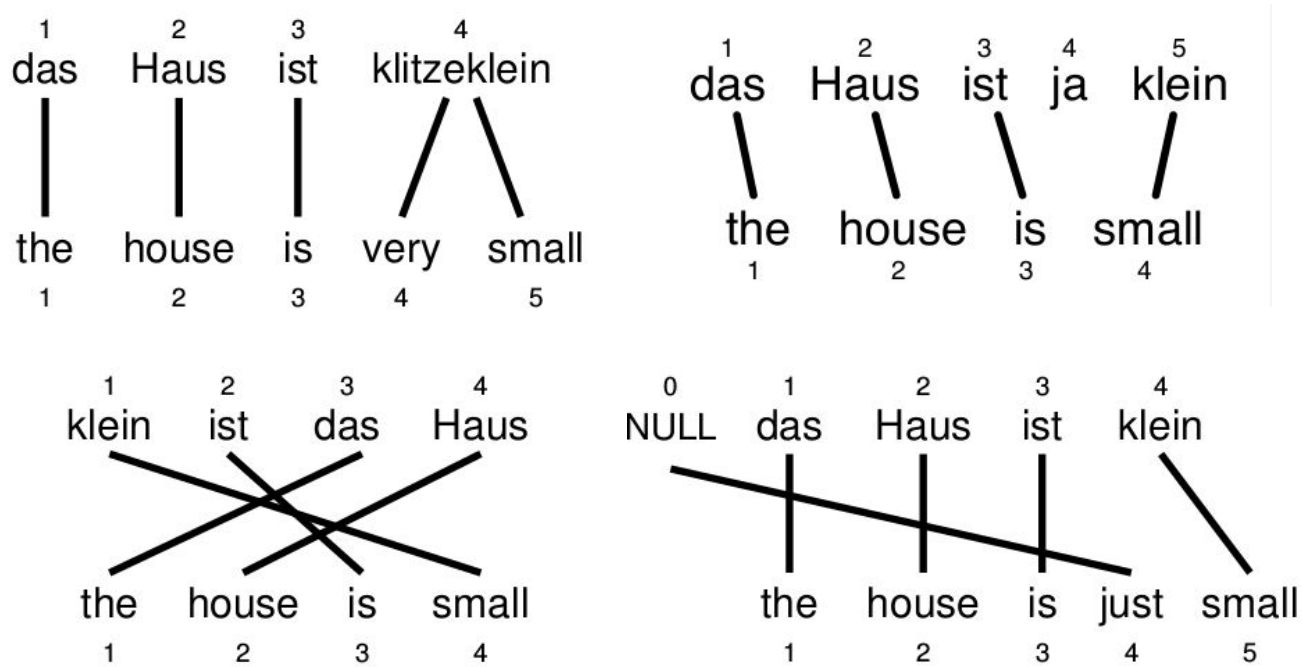
$$t(x|b) = 0.0001$$
$$t(y|b) = 0.9999$$
$$t(x|c) = 0.9999$$
$$t(y|c) = 0.0001$$

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

What am I glossing over?

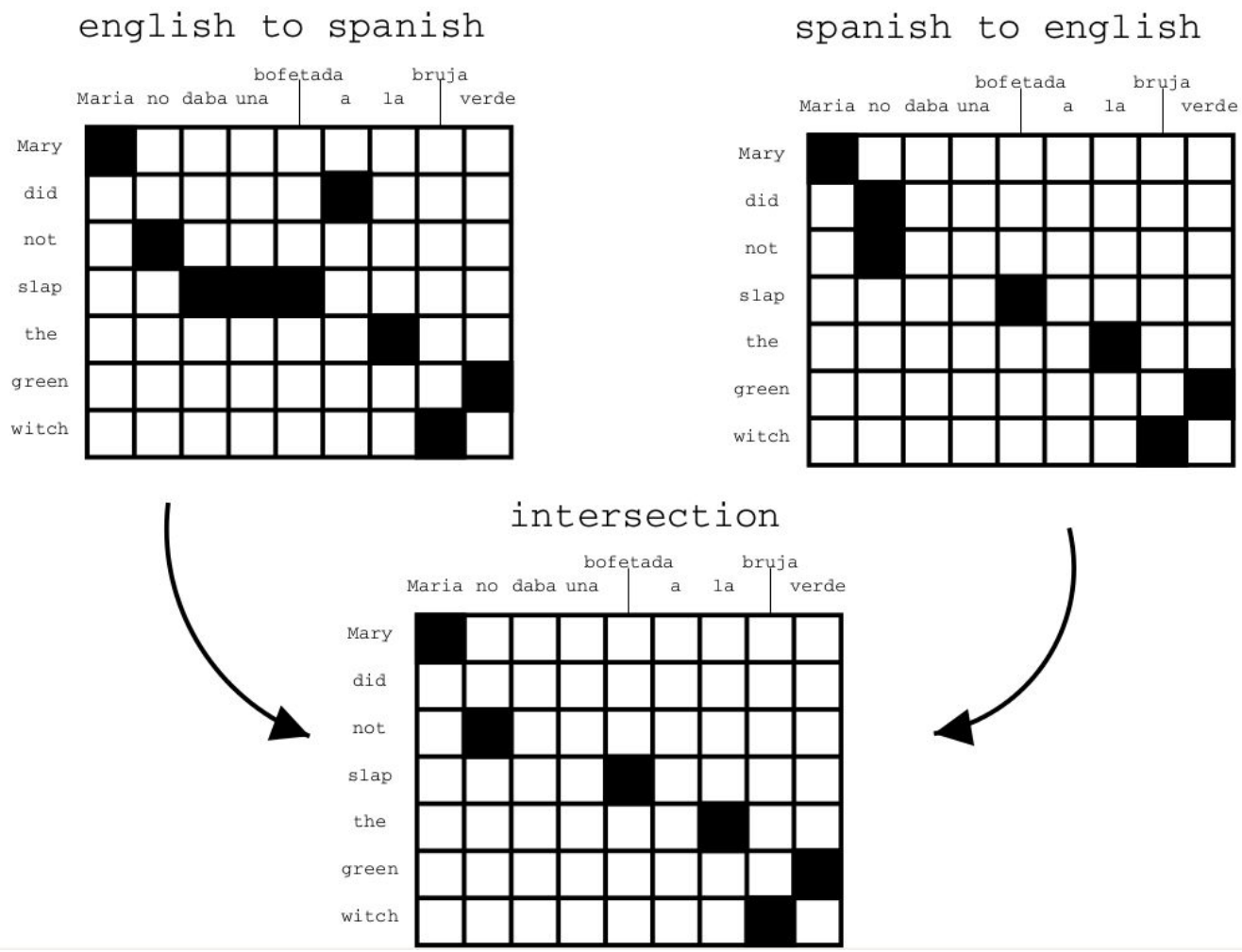
- Other IBM models, less strict alignments: [A Statistical MT Tutorial Workbook](#) (Knight, 1999)



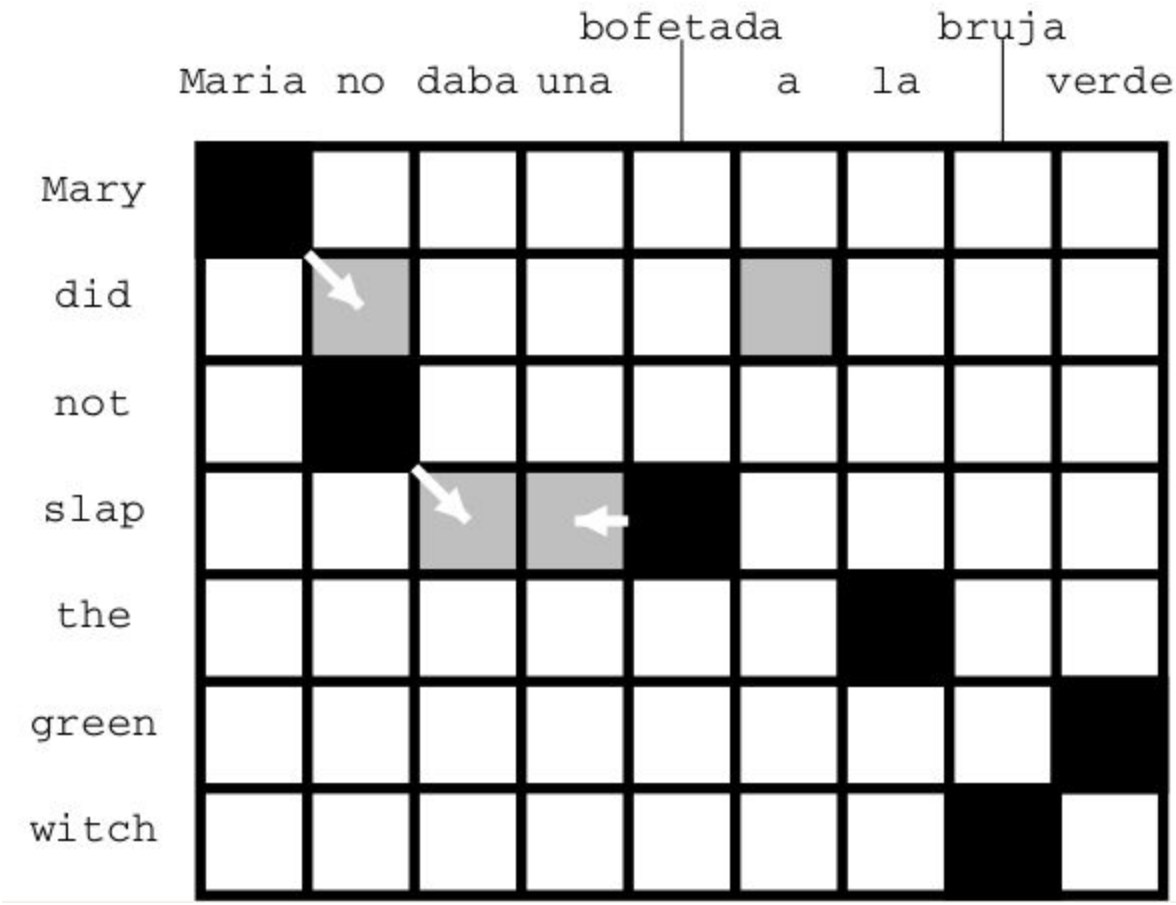
Why is Model 1 insufficient?

- Why won't this produce great translations?
- What might we also want to have in such a model?

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

What am I glossing over?

- Lots and lots of approaches to phrase extraction!
- http://mt-class.org/jhu/slides/lecture-phrase-based-models.pdf

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

We may have many translation options for a word or phrase. Decoding is NP-complete (can verify solutions in polynomial time, but can't locate solutions efficiently). We use heuristics to limit the search space.

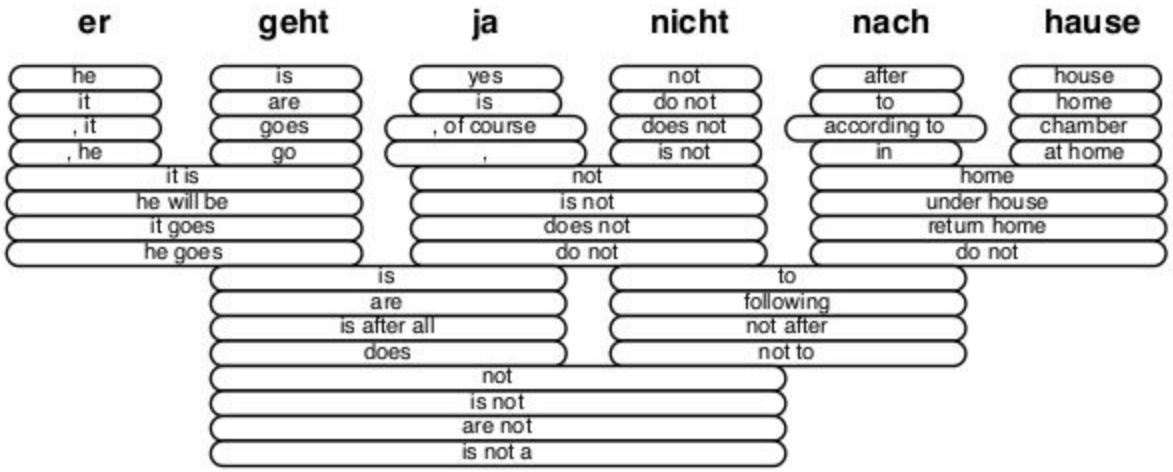
The role of the decoder is to:

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

How can this go wrong?

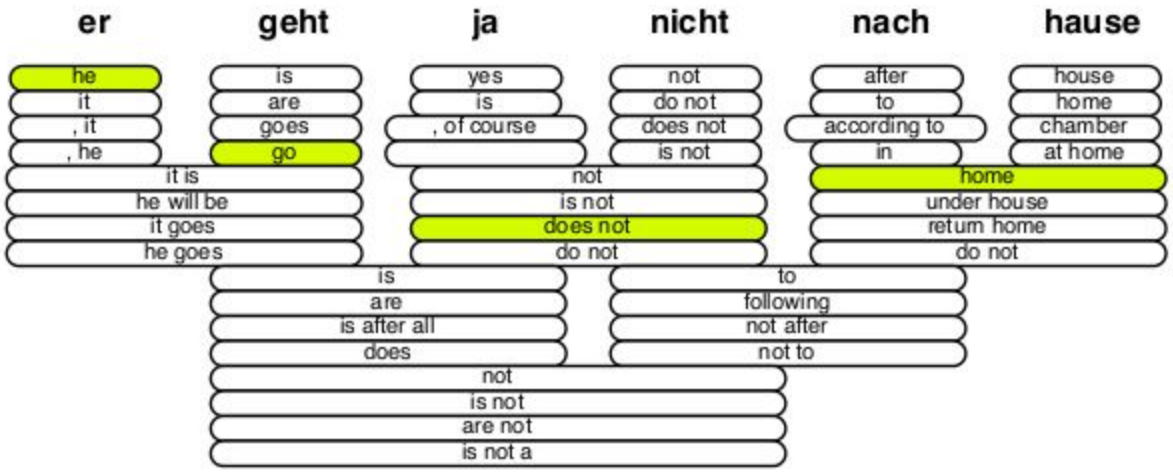
- Search doesn't find the best translation
  - Need to fix the search
- The best translation found is not good
  - Need to fix the model

Example



Koehn, MT class slides

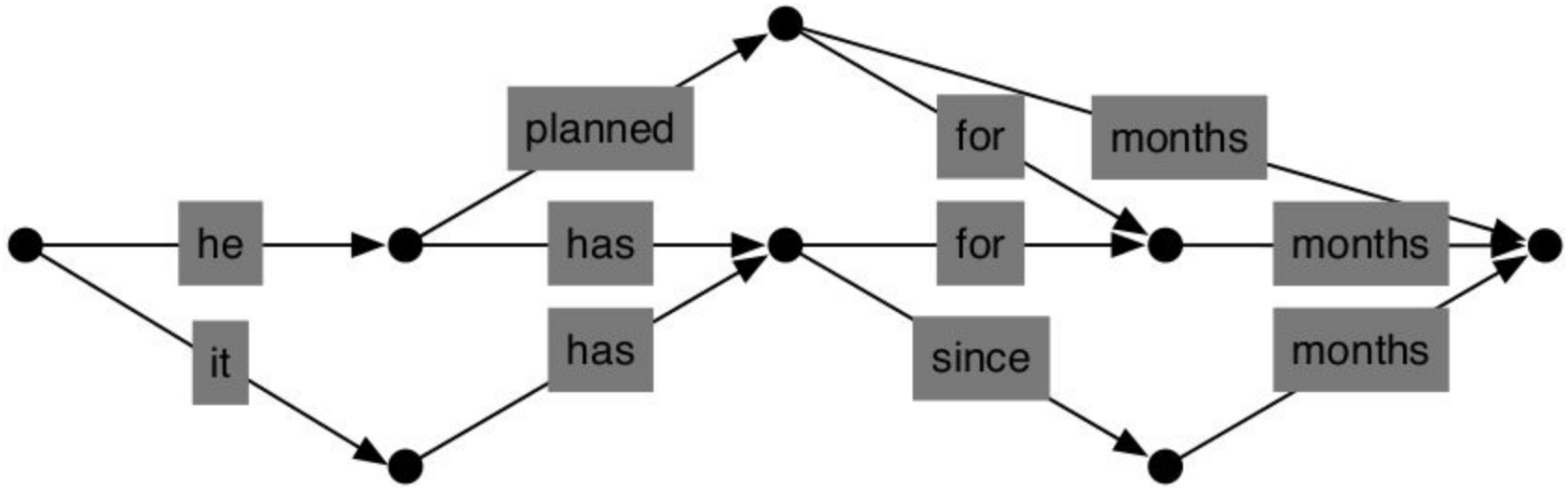
Example



Koehn, MT class slides

Search Graph

Er hat seit Monaten geplant.



Koehn CAT slides (2016)

What am I glossing over?

Most of the of details of decoding!

Try these resources:

Textbook slides: http://www.statmt.org/book/slides/06-decoding.pdf

Demo: http://mt-class.org/jhu/stack-decoder/

Phrase-Based SMT Overview

- Translate phrases (not just words)
- Combine information from language models, translation models, alignments, possibly more
- Need to restrict search space with heuristics (distortion limits, recombining similar hypotheses, limiting number of options to consider, etc.)