Phrase-based Models, Phrase Extraction

Mark Fishel

(adapted from slides by Barry Haddow)

(and before that from ones by Philipp Koehn)

(and before ...)

10 September 2013

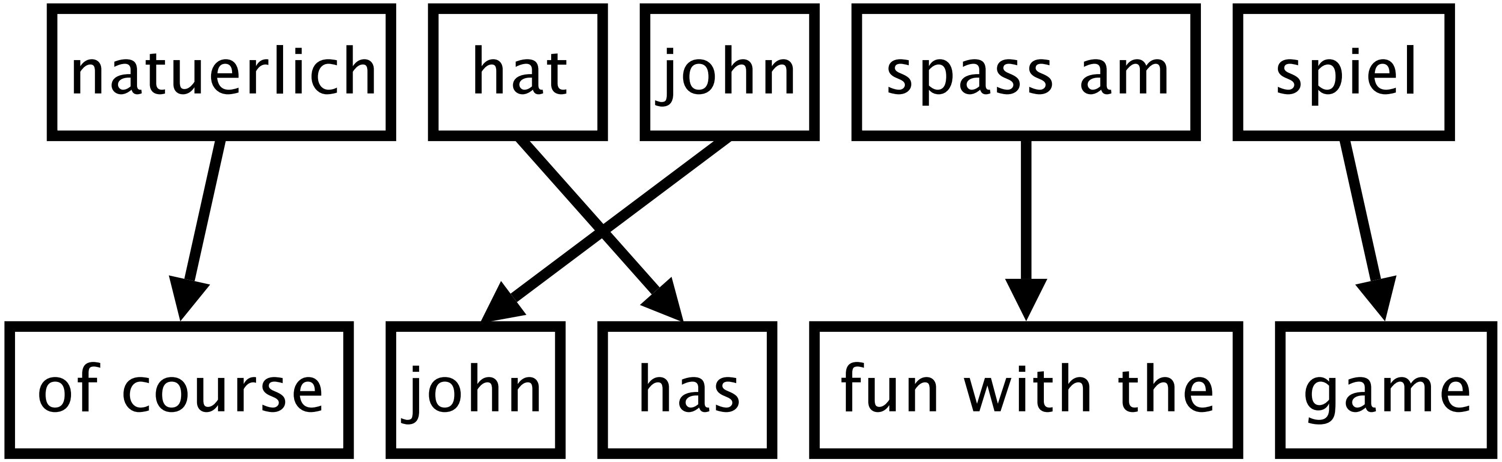
Outline

- From word-based to phrase-based
- Creating the phrase table
- The log-linear model and the “standard” phrase-based MT features

Motivation

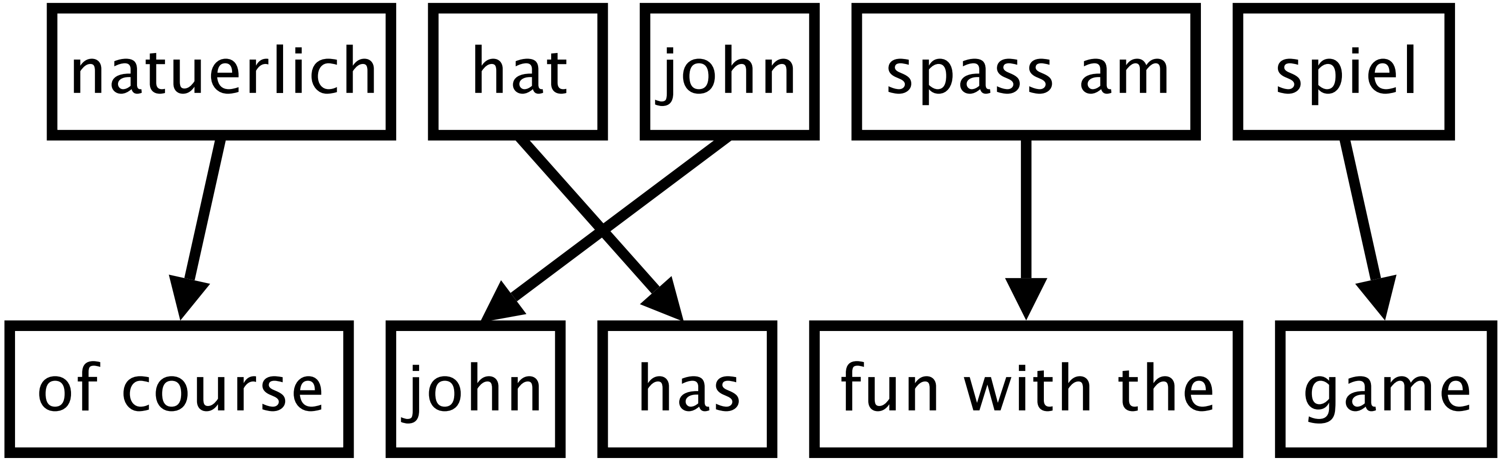
- Phrase-Based Models translate *phrases* instead of *words* as atomic units
- Advantages:
  - many-to-many translation can handle non-compositional phrases
  - use of local context in translation
  - the more data, the longer phrases can be learned
- "Standard Model", used by Google Translate and others

Phrase-Based Model



- Foreign input is segmented in phrases
- Each phrase is translated into English
- Phrases are reordered

Phrase-Based Model



- Simpler generative story
- No fertility or NULL-insertion

Big Picture

$$\begin{aligned}\hat{\mathbf{e}} &= \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\, p_{\text{TM}}(\mathbf{f}|\mathbf{e})\, p_{\text{LM}}(\mathbf{e})\end{aligned}$$

p_{TM} can be estimated in different ways:

- Word-based IBM model:

$$p_{\text{TM}}(\mathbf{f}|\mathbf{e}) = p_{lex}(\mathbf{f}|\mathbf{e}) \cdot p_{dist}(\mathbf{f}|\mathbf{e}) \cdot p_{fert}(\mathbf{f}|\mathbf{e}) \cdots$$

- Let us define how it is estimated in phrase-based translation

Phrase Translation Table

- Main knowledge source: table with phrase translations and their probabilities

- Example: phrase translations for natuerlich

Translation	**Probability** $\phi(\bar{e}\|\bar{f})$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Real Example

- Phrase translations for den Vorschlag learned from the Europarl corpus:

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

  - lexical variation (proposal vs suggestions)
  - morphological variation (proposal vs proposals)
  - included function words (the, a, ...)
  - noise (it)

Linguistic Phrases?

- Model is not limited to linguistic phrases
  (syntactic sub-tree spans: noun phrases, verb phrases, prep. phrases, . . .)

- Example non-linguistic phrase pair

$$\text{spass am} \rightarrow \text{fun with the}$$

- Prior noun often helps with translation of preposition

- Experiments show that limitation to linguistic phrases hurts quality

Learning a Phrase Translation Table

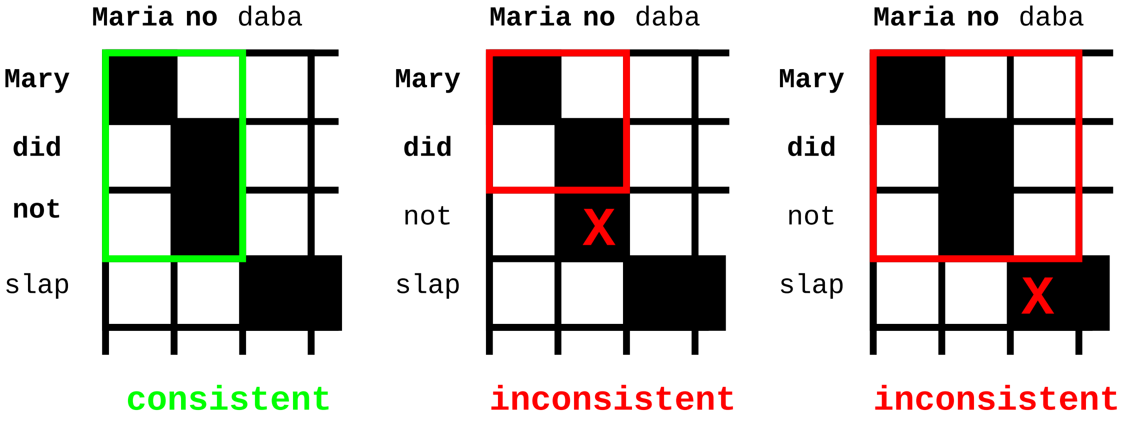
- Task: learn the model from a parallel corpus

- Three stages:
  - word alignment: using IBM models or other method
  - extraction of phrase pairs
  - scoring phrase pairs

Word Alignment

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

Phrase Extraction Criteria



- Phrase alignment has to *contain all alignment points* for all covered words
- Phrase alignment has to *contain at least one alignment point*

Phrase Extraction Criteria, Formalised

A phrase pair $(\overline{e}, \overline{f})$ is *consistent* with an alignment A if and only if:

1. No English words in the phrase pair are aligned to words outside it.

$$\forall e_i \in \overline{e}, (e_i, f_j) \in A \Rightarrow f_j \in \overline{f}$$

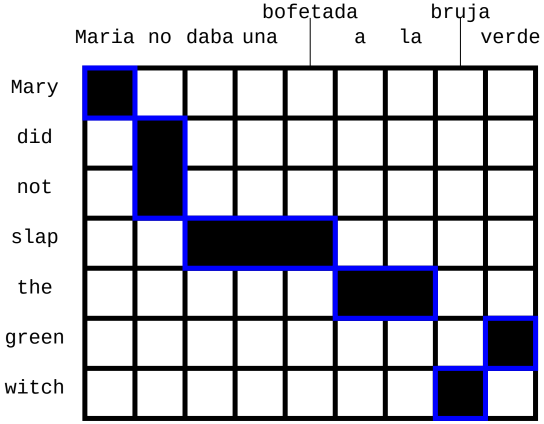
2. No Foreign words in the phrase pair are aligned to words outside it.

$$\forall f_j \in \overline{f}, (e_i, f_j) \in A \Rightarrow e_i \in \overline{e}$$

3. The phrase pair contains at least one alignment point.

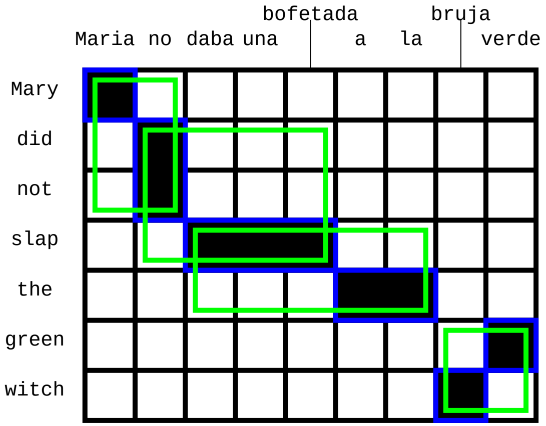
$$\exists e_i \in \bar{e}, f_j \in \bar{f} \; s.t. \; (e_i, f_j) \in A$$

Word alignment induced phrases



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

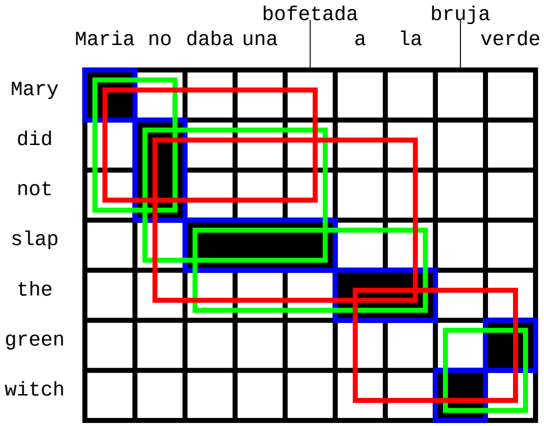
Word alignment induced phrases



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

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

Word alignment induced phrases

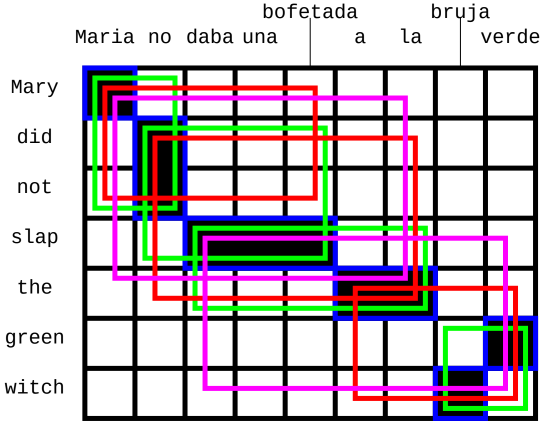


(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

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

(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Word alignment induced phrases



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

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),

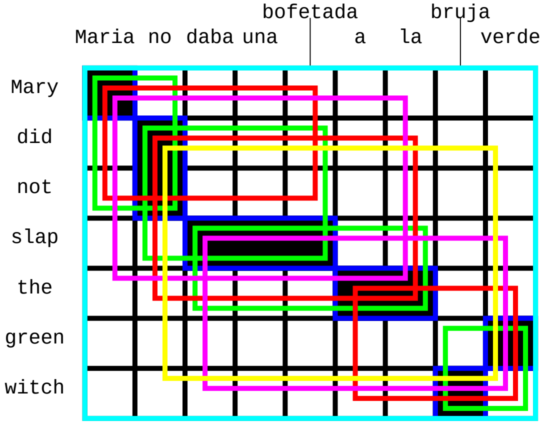
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),

(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),

(Maria no daba una bofetada a la, Mary did not slap the),

(daba una bofetada a la bruja verde, slap the green witch)

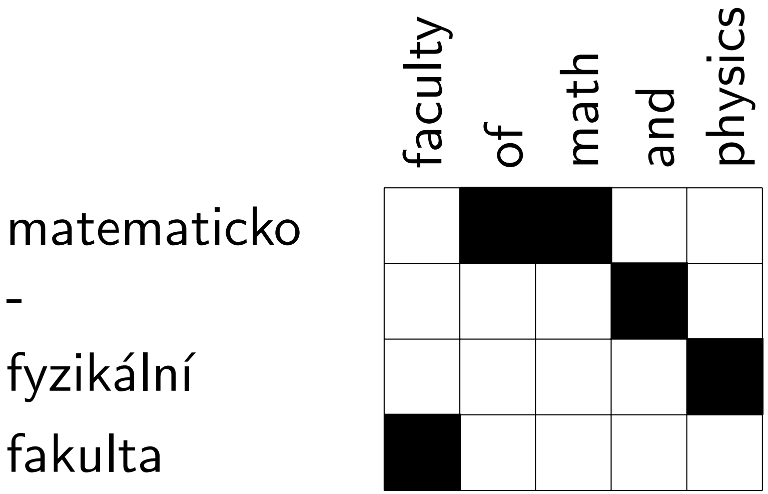
Word alignment induced phrases



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

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

Word alignment can be wrong



	faculty	of	math	and	physics
matematicko		■	■		
-				■	
fyzikální					■
fakulta	■				

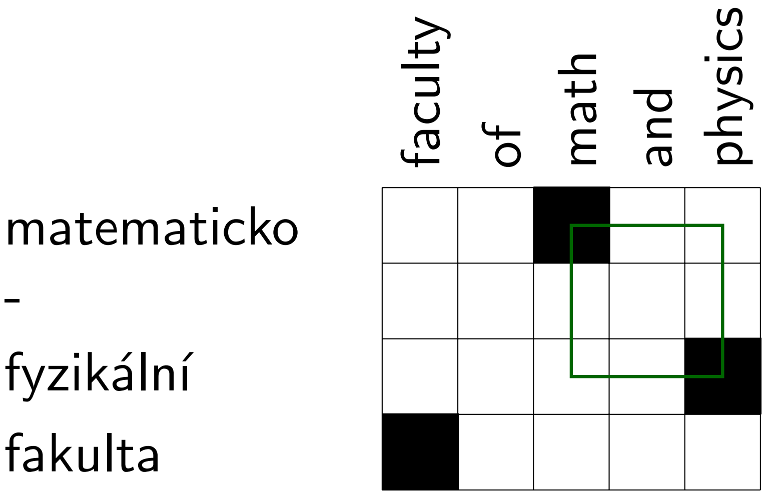
(fakulta, faculty), (matematicko - fyzikální, of math and physics), (fyzikální, physics),

(matematicko, of math), (-, and), (- fyzikální, and physics), etc.

Word alignment can be wrong

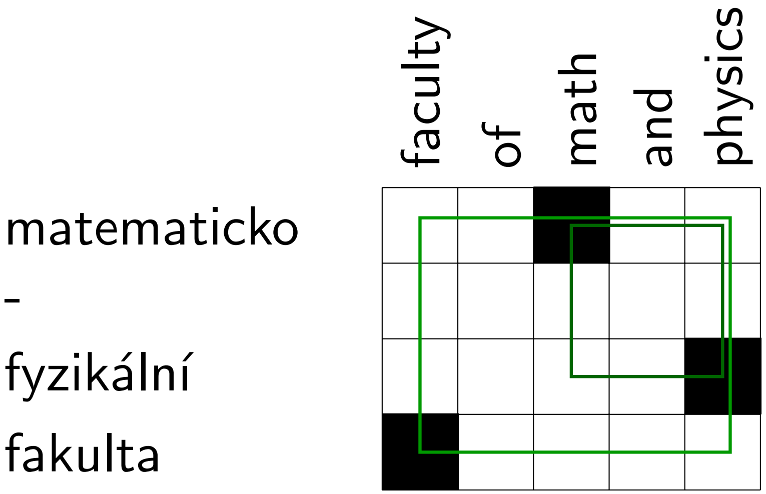
	faculty	of	math	and	physics
matematicko			■		
-					
fyzikální					■
fakulta	■				

Word alignment can be wrong



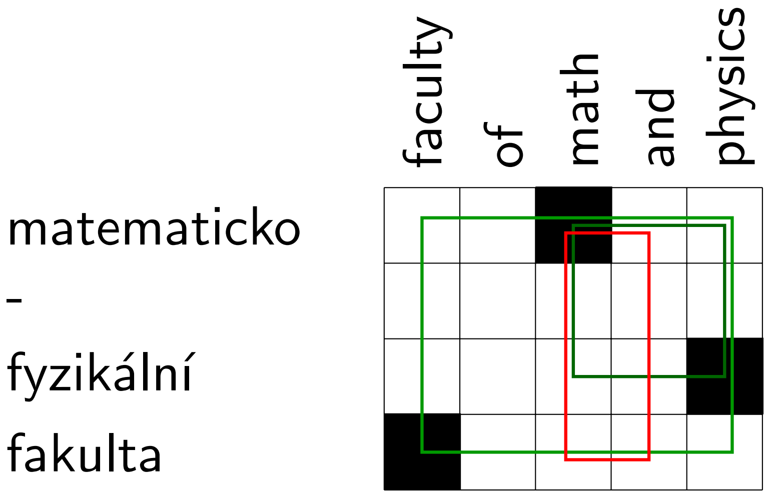
	faculty	of	math	and	physics
matematicko			■		
-					
fyzikální					■
fakulta	■				

Word alignment can be wrong

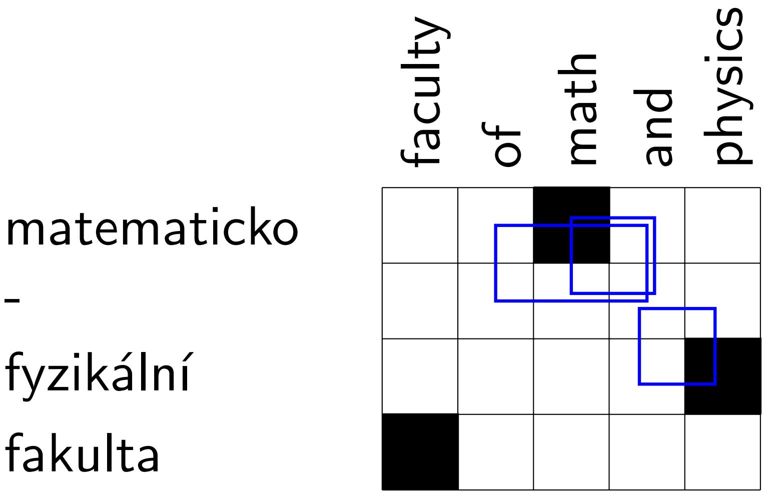


	faculty	of	math	and	physics
matematicko			■		
-					
fyzikální					■
fakulta	■				

Word alignment can be wrong



Word alignment can be wrong



Are these allowed?

Word alignment can be wrong

	the	faculty	of	math	and	physics
matematicko				■		
-						
fyzikální						■
fakulta		■				

This allows for extracting both (fakulta, faculty) and (fakulta, the faculty) = more flexible

Word alignment vs phrase extraction

A word alignment **restricts** the extracted phrase pairs to it

- A dense word alignment
  - will result in fewer phrase pairs
  - might bring wrong phrases in case of wrong or uncertain word alignment points

- A sparse word alignment
  - will result in more phrases, many of which might be noisy
  - but is preferred to an uncertain dense word alignment, since it can also include the correct phrase pairs
  - scoring can bring up the correct phrase pairs

Scoring Phrase Translations

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$

Scoring Translations

- Searching for the most likely output

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f}) = \text{argmax}_{\mathbf{e}}\, p_{\text{TM}}(\mathbf{f}|\mathbf{e})\; p_{\text{LM}}(\mathbf{e})$$

  - translation model $p_{\text{TM}}(\mathbf{f}|\mathbf{e})$, language model $p_{\text{LM}}(\mathbf{e})$

- Decomposition of the translation model

$$p_{\text{TM}}(\mathbf{f}|\mathbf{e}) = p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)$$

Log-linear Model

$$\frac{\partial \ln f(x)}{\partial x} = \frac{1}{x} \cdot \frac{\partial f(x)}{\partial x} \implies \text{argmax}_x f(x) = \text{argmax}_x \ln f(x)$$

which means that we can replace

$$score(\mathbf{e}) = p_{\mathrm{TM}}(\mathbf{f}|\mathbf{e}) \cdot p_{\mathrm{LM}}(\mathbf{e})$$

with

$$logscore(\mathbf{e}) = \ln score(\mathbf{e}) = \ln p_{\mathrm{TM}}(\mathbf{f}|\mathbf{e}) + \ln p_{\mathrm{LM}}(\mathbf{e})$$

and search for its maximum instead

Weighted Log-linear Model

We can weight the models differently

$$logscore(\mathbf{e}) = \lambda_{\text{TM}} \log p_{\text{TM}}(\mathbf{f}|\mathbf{e}) + \lambda_{\text{LM}} \log p_{\text{LM}}(\mathbf{e})$$

and actually add any number of models, called *feature functions*:

$$logscore(\mathbf{e}) = \sum_{k=1}^{n} \lambda_k h_k(\mathbf{e}, \mathbf{f})$$

Weights $\vec{\lambda}$ determined automatically via tuning (MERT/PRO/MIRA/. . .)

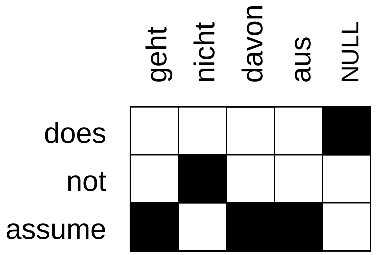
Adding features

- Reverse phrase translation probability (i.e. $p_{\text{TM}}(\mathbf{e}|\mathbf{f})$)
- Lexical translation probabilities (in both directions)
- Phrase count
- Word count
- Distortion cost
- Reordering score

Adding features

- Rare phrase pairs have unreliable phrase translation probability estimates

  → lexical weighting with word translation probabilities



	geht	nicht	davon	aus	NULL
does					■
not		■			
assume	■		■	■	

$$\text{lex}(\bar{e}|\bar{f},a) = \prod_{i=1}^{\text{length}(\bar{e})} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall (i,j) \in a} w(e_i|f_j)$$

Adding features

- Language model has a bias towards short translations

  → word count: $\text{wc}(e) = \log |\mathbf{e}|^{\omega}$

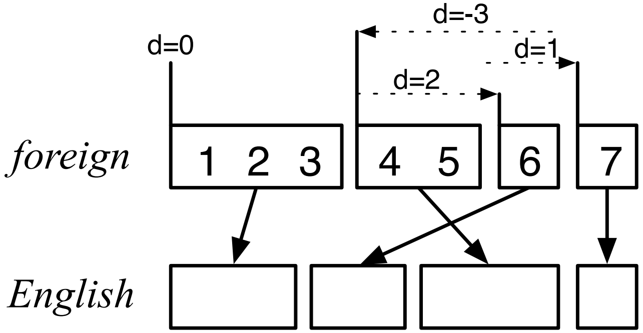
- We may prefer finer or coarser segmentation

  → phrase count $\text{pc}(e) = \log |I|^{\rho}$

- Multiple language models

- Multiple translation models

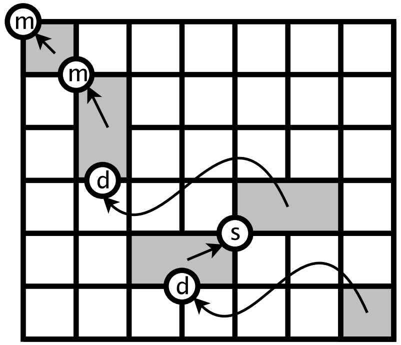
Distance-Based Reordering



phrase	translates	movement	distance
1	1–3	start at beginning	0
2	6	skip over 4–5	+2
3	4–5	move back over 4–6	-3
4	7	skip over 6	+1

Scoring function: $d(x) = \alpha^{|x|}$ — exponential with distance

Lexicalized Reordering



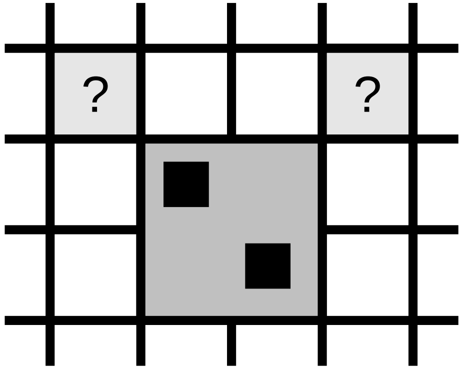
- Distance-based reordering model is weak
  → learn reordering preference for each phrase pair

- Three orientations types: (m) monotone, (s) swap, (d) discontinuous

$$\text{orientation} \in \{m, s, d\}$$

$$p_o(\text{orientation}|\bar{f}, \bar{e})$$

Learning Lexicalized Reordering



- Collect orientation information during phrase pair extraction
  - if word alignment point to the top left exists → **monotone**
  - if a word alignment point to the top right exists→ **swap**
  - if neither a word alignment point to top left nor to the top right exists → neither monotone nor swap → **discontinuous**

Learning Lexicalized Reordering

- Estimation by relative frequency

$$p_o(\text{orientation}) = \frac{\sum_{\bar{f}} \sum_{\bar{e}} count(\text{orientation}, \bar{e}, \bar{f})}{\sum_o \sum_{\bar{f}} \sum_{\bar{e}} count(o, \bar{e}, \bar{f})}$$

- Smoothing with unlexicalized orientation model $p(\text{orientation})$ to avoid zero probabilities for unseen orientations

$$p_o(\text{orientation}|\bar{f}, \bar{e}) = \frac{\sigma\ p(\text{orientation}) + count(\text{orientation}, \bar{e}, \bar{f})}{\sigma + \sum_o count(o, \bar{e}, \bar{f})}$$

In Practice:

Phrase table (phrase probability, lexical weight, reverse probability, reverse weight):

```
Maria no             ||| Mary did not ||| 0.1   0.01  0.12  0.02
no daba una bofetada ||| did not slap ||| 0.23  0.009  0.29  0.08
daba una bofetada    ||| slap         ||| 0.16  0.027  0.04  0.06
...
```

Lexicalized reordering table (monotone/swap/discont for both directions):

```
Maria no             ||| Mary did not ||| 0.2 0.2 0.6 0.6 0.2 0.2
no daba una bofetada ||| did not slap ||| 0.4 0.3 0.3 0.3 0.2 0.5
daba una bofetada    ||| slap         ||| 0.3 0.1 0.6 0.6 0.3 0.1
...
```

EM Training of the Phrase Model

- We presented a heuristic set-up to build phrase translation table (word alignment, phrase extraction, phrase scoring)

- Alternative: align phrase pairs directly with EM algorithm
  - initialization: uniform model, all $\phi(\bar{e}, \bar{f})$ are the same
  - expectation step:
    - estimate likelihood of all possible phrase alignments for all sentence pairs
  - maximization step:
    - collect counts for phrase pairs $(\bar{e}, \bar{f})$, weighted by alignment probability
    - update phrase translation probabilties $p(\bar{e}, \bar{f})$

- However: method easily overfits (learns very large phrase pairs, spanning entire sentences)

Summary

- Phrase Model
- Training the model
  - word alignment
  - phrase pair extraction
  - phrase pair scoring
- Log linear model
  - sub-models as feature functions
  - lexical weighting
- Lexicalized reordering model
- EM training of the phrase model