Statistical Machine Translation
Lecture 3

Beyond IBM Model 1 to Phrase-Based Models

Stephen Clark

(based on slides by Philipp Koehn)

Model 2

- Introduces more realistic assumption for the alignment probabilities:

$$\begin{aligned} a(a_j|j,m,l) &= p(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \\ &\approx p(a_j|j,m,l) \end{aligned}$$

Model 2

- What effect does this have on the estimation process (EM)?
  - very little: similar “trick” to that from Model 1 which allowed efficient summing over an exponential number of alignments still works
- Estimates from Model 1 can be used as initial guesses for Model 2 (Model 1 is a special case of Model 2)
- See Appendix B of the classic paper for a concise description of the various models

Model 3

- First allow a single English word to translate to many French
  - *fertility* probabilities
- Next translate the English words into French words
  - just a generalisation of what we've seen already
- Finally allow the French units to be reordered
  - *distortion* probabilities

Models 4 and 5

- Some technical extensions of Model 3
- Crucially, increased sophistication of Model 3 means that computational trick no longer works for summing over alignments
  - so have to sum over *high probability* alignments
- Search for most probable alignment
  - this is known as the *Viterbi* alignment: the alignment which maximises $p(a|e, f)$
  - no efficient algorithm for finding the Viterbi alignment for Models 3-5
- Obtain the sample for summing over by introducing slight modifications to the high probability alignment

Flaws of Word-Based MT

- Multiple English words for one German word

  one-to-many problem: `Zeitmangel` → `lack of time`

German:	`Zeitmangel`	`erschwert`		`das`	`Problem` `.`
Gloss:	LACK OF TIME	MAKES MORE DIFFICULT		THE	PROBLEM .
Correct translation:	Lack of time makes the problem more difficult.				
MT output:	`Time makes the problem .`				

- Phrasal translation

  non-compositional phrase: `erübrigt sich` → `there is no point in`

German:	`Eine`	`Diskussion`	`erübrigt`	`sich`	`demnach`
Gloss:	A	DISCUSSION	IS MADE UNNECESSARY	ITSELF	THEREFORE
Correct translation:	Therefore, there is no point in a discussion.				
MT output:	`A debate turned therefore .`				

Flaws of Word-Based MT (2)

- Syntactic transformations

reordering, genitive NP: `der Sache` → `for this matter`

German:	`Das`	`ist`	`der`	`Sache`	`nicht`	`angemessen`	`.`
Gloss:	THAT	IS	THE	MATTER	NOT	APPROPRIATE	.

Correct translation: That is not appropriate for this matter .

MT output: `That is the thing is not appropriate .`

object/subject reordering

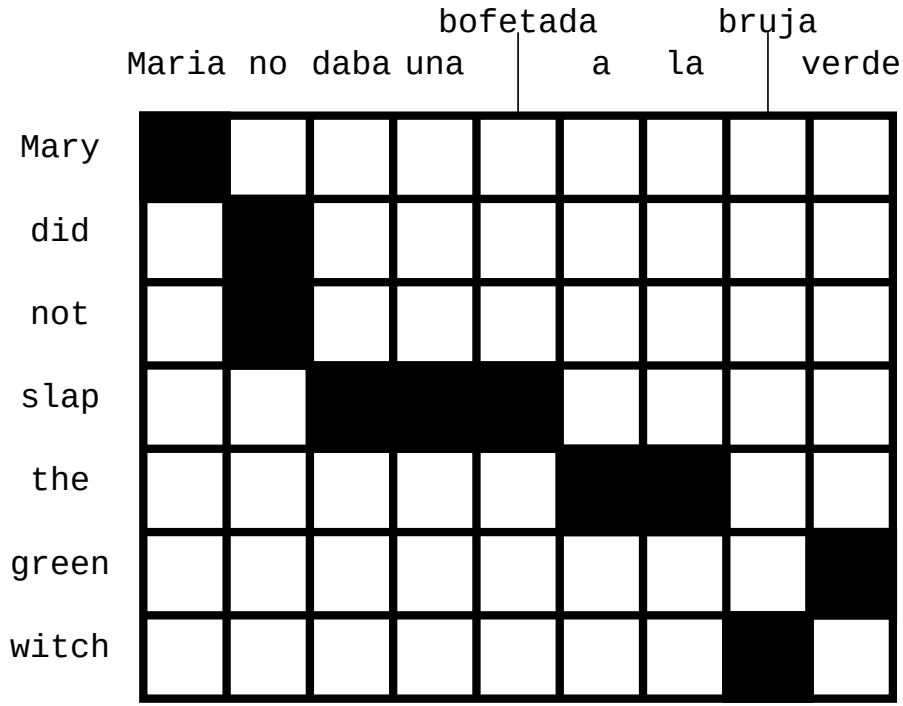
German:	`Den`	`Vorschlag`	`lehnt`	`die`	`Kommission`	`ab`	`.`
Gloss:	THE	PROPOSAL	REJECTS	THE	COMMISSION	OFF	.

Correct translation: The commission rejects the proposal .

MT output: `The proposal rejects the commission .`

Word Alignment

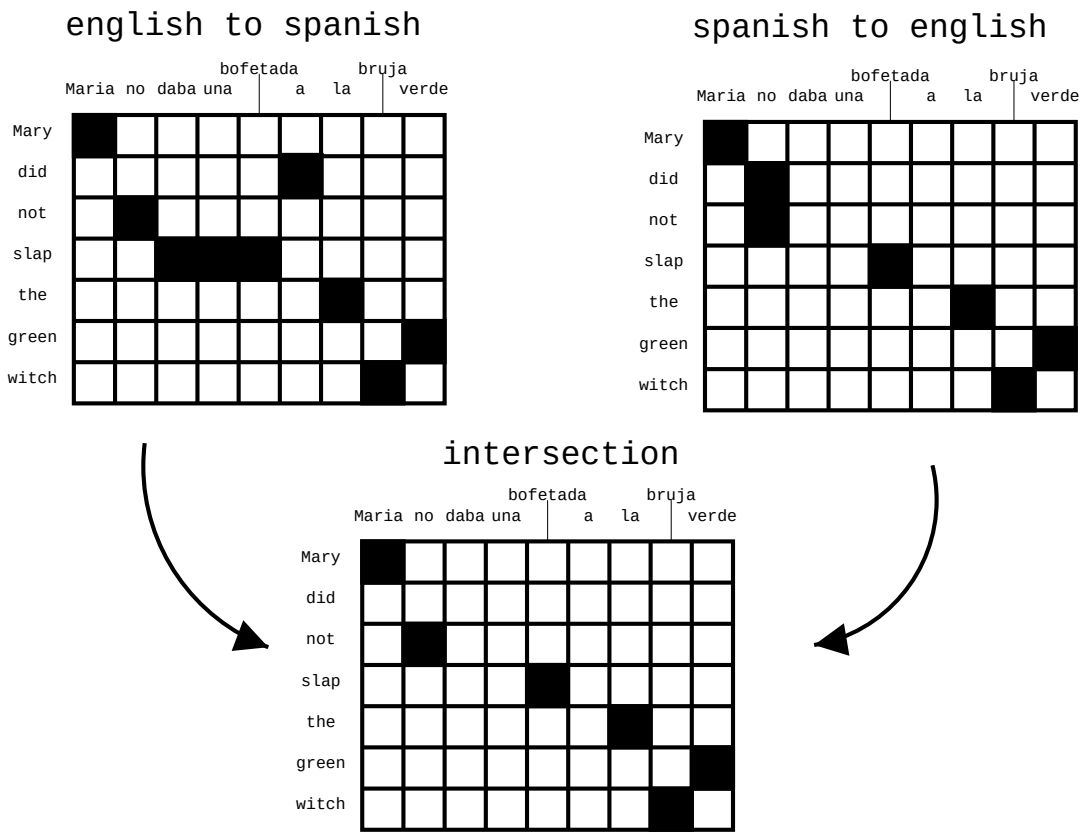
- Notion of word alignments valuable
- Trained humans can achieve high agreement



Word Alignment with IBM Models

- IBM Models create a many-to-one mapping
  - words are aligned using an alignment function
- More generally we need many-to-many mappings

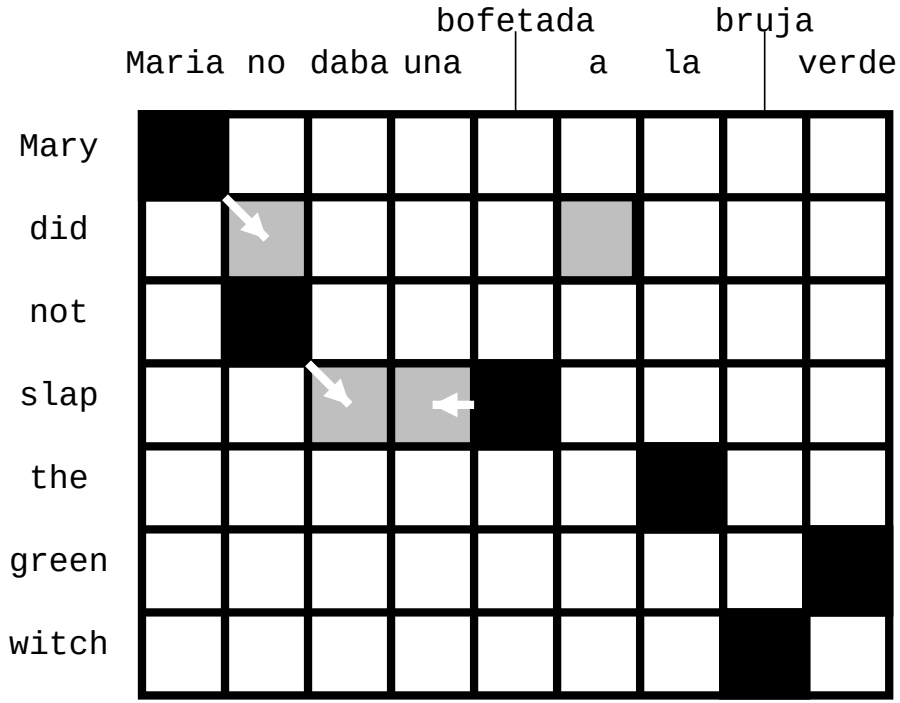
Improved Word Alignments



- Intersection of GIZA++ bidirectional alignments

[GIZA is a freely available alignment tool which implements the IBM models]

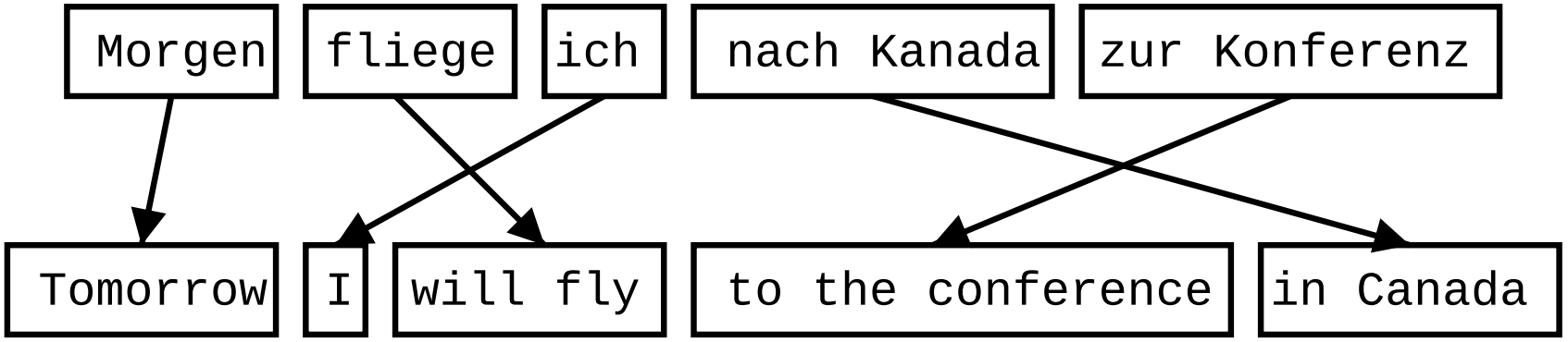
Improved Word Alignments (2)



- Grow additional alignment points

(Och and Ney, A Systematic Comparison of Various Statistical Alignment Models, Comp. Linguistics 2003)

Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, *not necessarily linguistically motivated*
- Each phrase is translated into English
- Phrases are reordered

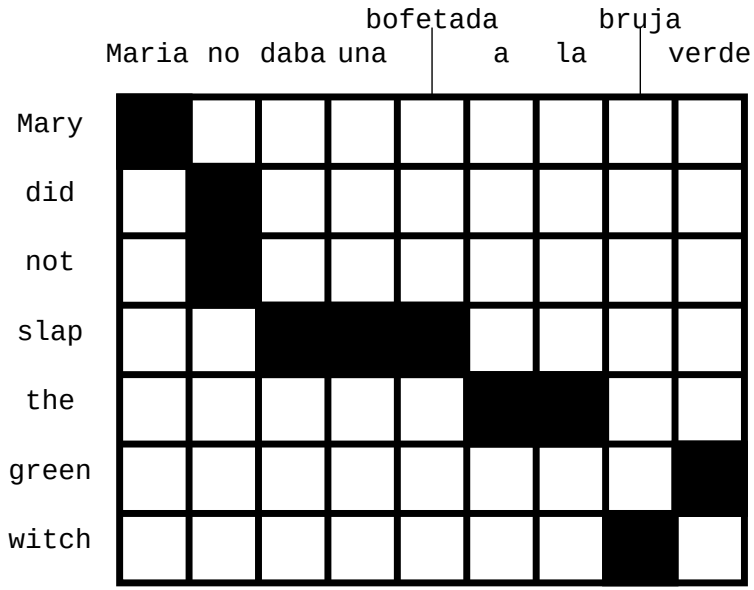
(Koehn, Och and Marcu, Statistical Phrase-Based Translation, NAACL 2003)

Advantages of Phrase-Based Translation

- Many-to-many translation can handle non-compositional phrases
- Use of local context in translation
- The more data, the longer phrases can be learned
- Perhaps conceptually simpler than the later IBM models based on words (no fertility parameters etc)

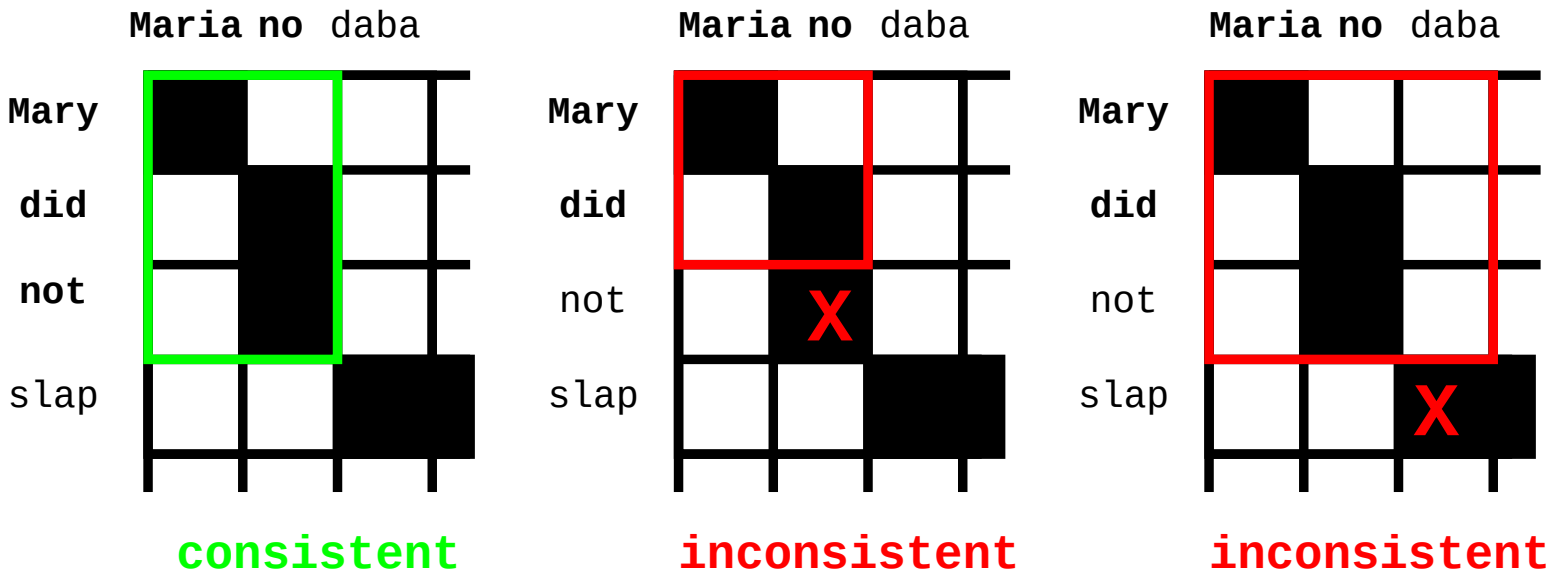
How to Learn the Phrase Translation Table?

- Start with the word alignment:



- Collect all phrase pairs that are consistent with the word alignment

Consistent with Word Alignment



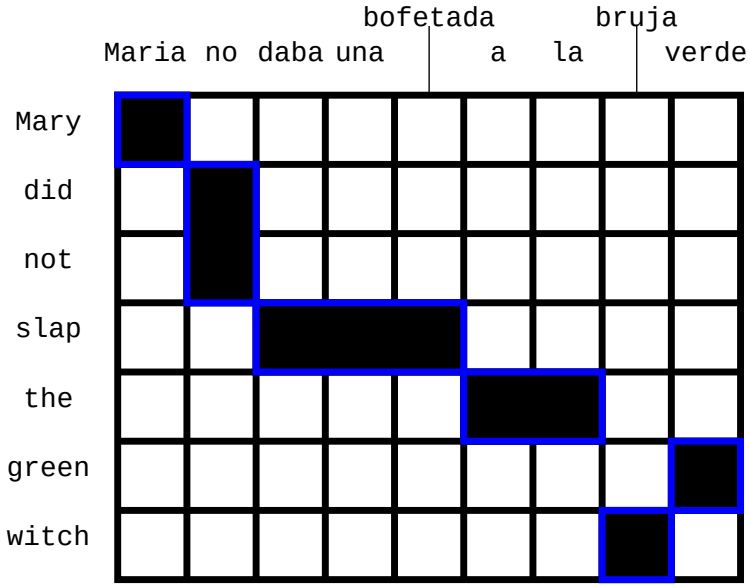
- Consistent with the word alignment :=

  phrase alignment has to contain all alignment points for all covered words

$$(\bar{e}, \bar{f}) \in BP \Leftrightarrow \quad \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$$

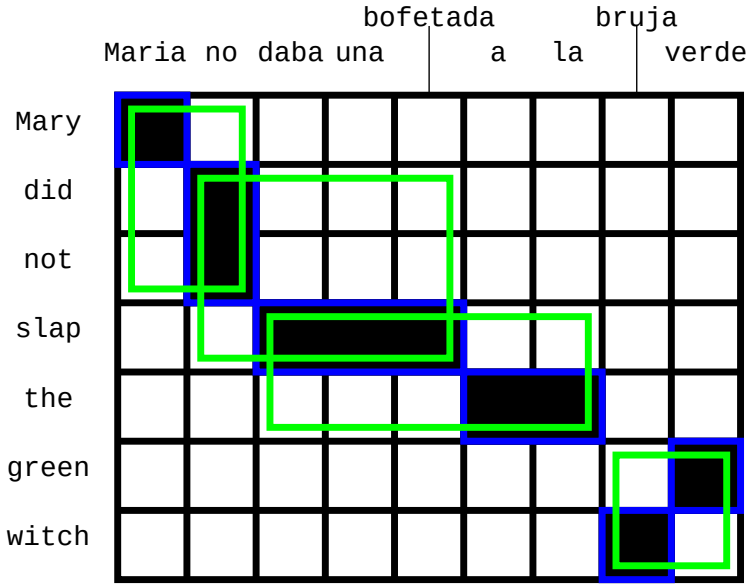
$$\text{AND} \quad \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}$$

Word Alignment Induced Phrases



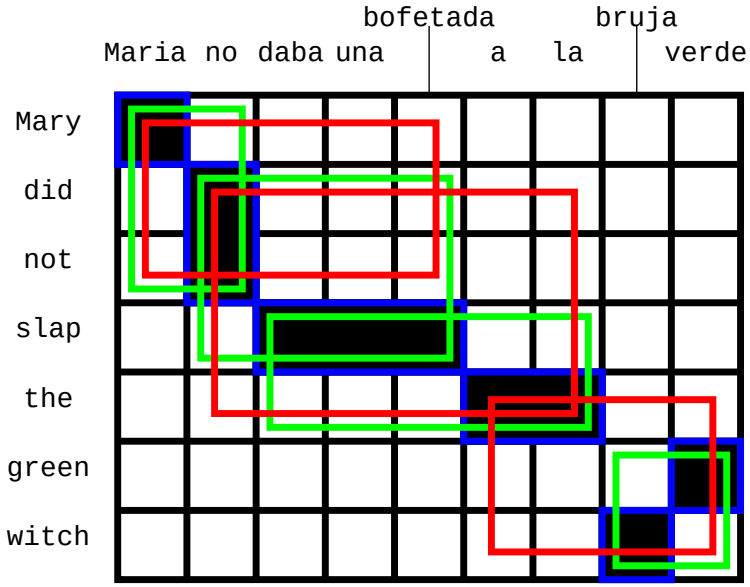
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Word Alignment Induced Phrases (2)



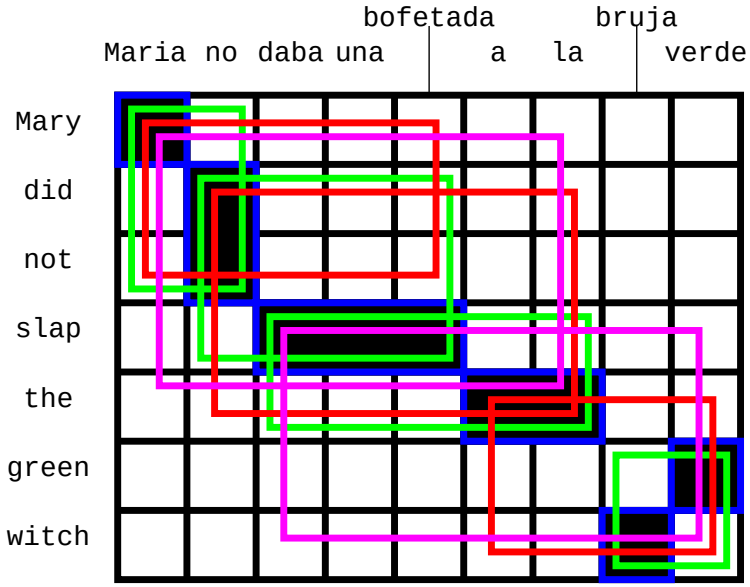
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word Alignment Induced Phrases (3)



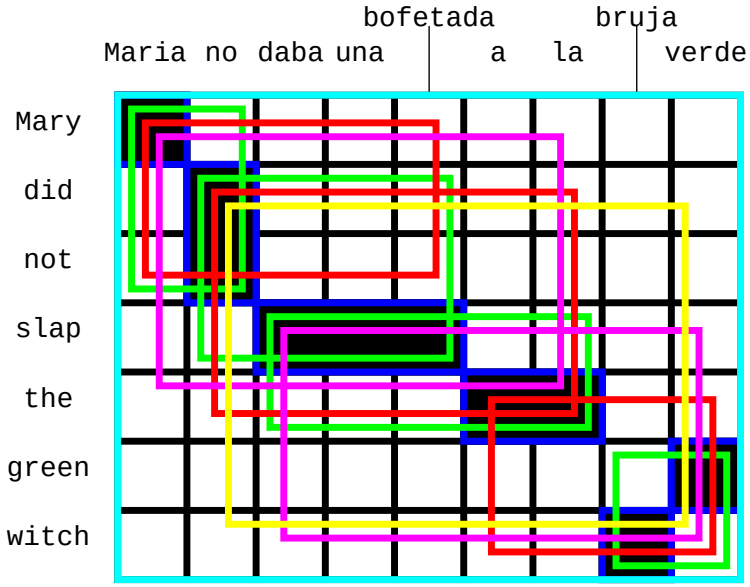
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green), (Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap), (no daba una bofetada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

Word Alignment Induced Phrases (4)



(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 (5)



(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)

Probability Distribution of Phrase Pairs

- We need a probability distribution $\phi(\bar{f}|\bar{e})$ over the collected phrase pairs

$\Rightarrow$ Possible choices

  - relative frequency of collected phrases:
    $$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$$
  - or, conversely $\phi(\bar{e}|\bar{f})$
  - use lexical translation probabilities

Reordering

- Monotone translation
  - do not allow any reordering

  $\longrightarrow$ worse translations
- Distance-based reordering cost
  - moving a foreign phrase over n words: cost ω^n

Phrase-Based Model

- Use noisy channel model as before
- Only the translation model changes:

$$
\begin{aligned}
p(f|e) &= p(\overline{f}_1^I|\overline{e}_1^I) \\
&= \prod_{i=1}^{I} \phi(\overline{f}_i|\overline{e}_i) d(\mathit{start}_i - \mathit{end}_{i-1} - 1)
\end{aligned}
$$

- Sentence pair split into I phrase pairs
- Each segmentation is assumed to be equally likely
- d is a distance-based reordering model

Distance-based Reordering Model

$$d(\mathit{start}_i - \mathit{end}_{i-1} - 1)$$

- start_i is the position of the first word of the foreign phrase that translates to the ith English phrase
- end_i is the position of the last word of the foreign phrase that translates to the ith English phrase
- d isn't estimated from data; just assume an exponentially decaying cost function:

$$d(x) = \alpha^{|x|} \quad \alpha \in [0, 1]$$

Log-Linear Models

- IBM Models provided mathematical justification for factoring components together

$$p_{LM} \times p_{TM} \times p_D$$

- These may be weighted

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

- Many components p_i with weights λ_i

$\Rightarrow \prod_i p_i^{\lambda_i} = exp(\sum_i \lambda_i log(p_i))$

$\Rightarrow log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i log(p_i)$

Set Feature Weights

- Contribution of components p_i determined by weight λ_i
- Methods
  - manual setting of weights: try a few, take best
  - automate this process
- Learn weights
  - set aside a development corpus
  - set the weights, so that optimal translation performance on this development corpus is achieved
  - requires automatic scoring method (e.g., BLEU)

Additional Features

- Word count
  - add fixed factor for each generated word
  - if output is too short $\longrightarrow$ add benefit for each word
- Phrase count
  - add fixed factor for each phrase
  - balances use of longer or shorter phrases