Statistical Machine Translation
Part IIIb – Phrase-based Model

Alexander Fraser
CIS, LMU München

2015.11.10 WSD and MT

Where we have been

- We defined the overall problem and talked about evaluation
- We have now covered word alignment
  - IBM Model 1, true Expectation Maximization
  - Briefly mentioned: IBM Model 4, approximate Expectation Maximization
  - Symmetrization Heuristics (such as Grow)
    - Applied to two Viterbi alignments (typically from Model 4)
    - Results in final word alignment

Where we are going

- We will discuss the "traditional" phrase-based model (which noone actually uses, but gives a good intuition)
- Then we will define a high performance translation model (next slide set)
- Finally, we will show how to solve the search problem for this model (= decoding)

Outline

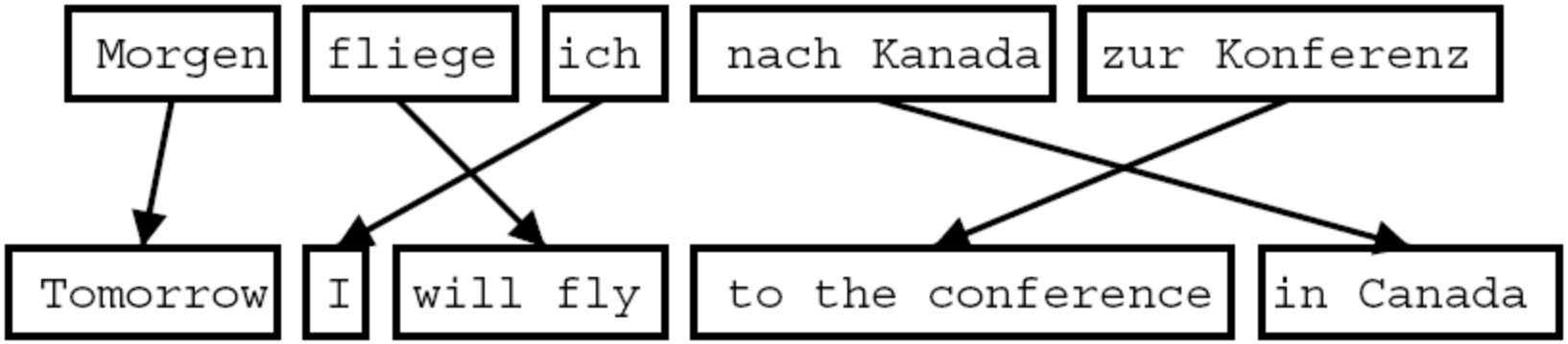
- Phrase-based translation
  - Model
  - Estimating parameters
- Decoding

- We could use IBM Model 4 in the direction p(f|e), together with a language model, p(e), to translate

$$\underset{e}{\text{argmax}}\ P(e \mid f) = \underset{e}{\text{argmax}}\ P(f \mid e)\ P(e)$$

- However, decoding using Model 4 doesn't work well in practice
  - One strong reason is the bad 1-to-N assumption
  - Another problem would be defining the search algorithm
    - If we add additional operations to allow the English words to vary, this will be very expensive
  - Despite these problems, Model 4 decoding was briefly state of the art
- We will now define a better model…

Phrase-based translation

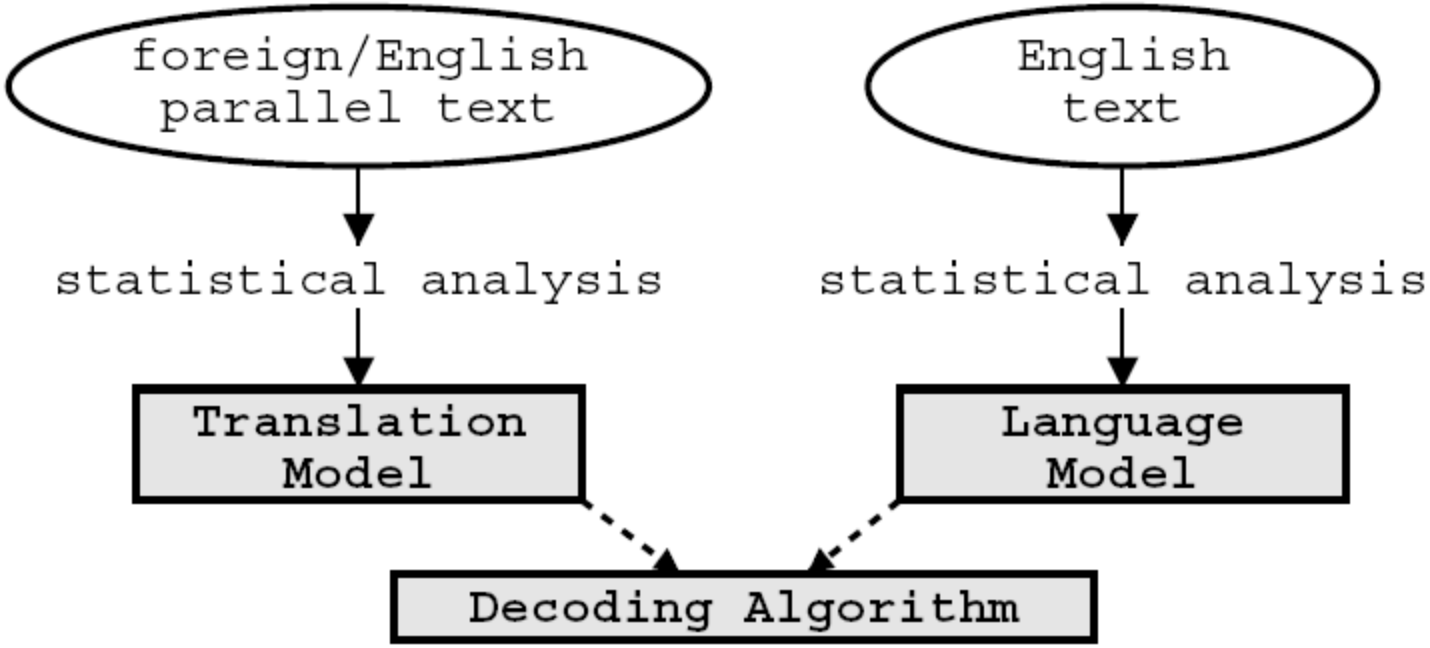


- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Slide from Koehn 2008

Statistical Machine Translation

- Components: Translation model, language model, decoder



Slide from Koehn 2008

Language Model

- Often a trigram language model is used for p(e)
  - P(the man went home) = p(the | START) p(man | START the) p(went | the man) p(home | man went)
- Language models work well for comparing the grammaticality of strings of the same length
  - However, when comparing short strings with long strings they favor short strings
  - For this reason, an important component of the language model is the length bonus
    - This is a constant > 1 multiplied for each English word in the hypothesis
    - It makes longer strings competitive with shorter strings

Phrase-based translation model

- Major components of phrase-based model
  - **phrase translation model** $\phi(\mathbf{f}|\mathbf{e})$
  - **reordering model** d
  - **language model** $p_{\text{LM}}(\mathbf{e})$
- Bayes rule

$$\begin{aligned} \text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \\ &= \text{argmax}_{\mathbf{e}} \phi(\mathbf{f}|\mathbf{e}) p_{\text{LM}}(\mathbf{e}) \omega^{\text{length}(\mathbf{e})} \end{aligned}$$

- Sentence $\mathbf{f}$ is decomposed into I phrases $\bar{f}_1^I = \bar{f}_1, ..., \bar{f}_I$
- Decomposition of $\phi(\mathbf{f}|\mathbf{e})$

$$\phi(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

Modified from Koehn 2008

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

Slide from Koehn 2008

Phrase translation table

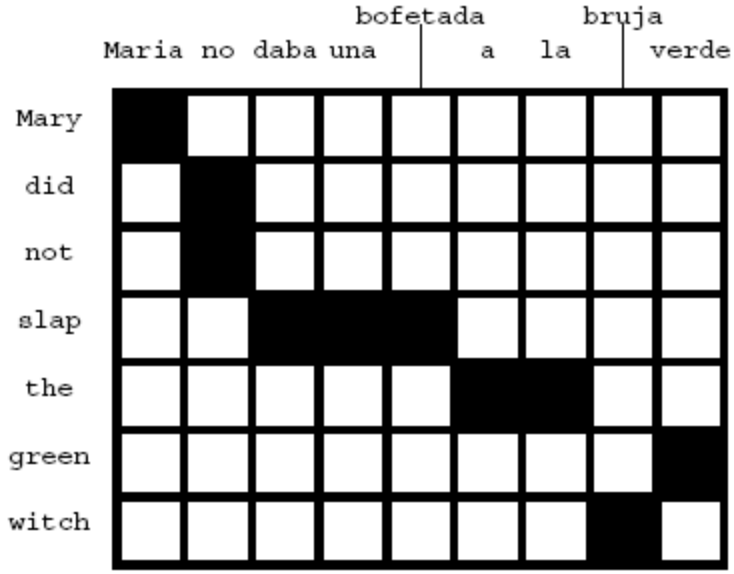
- Phrase translations for *den Vorschlag*

English	ϕ(e|f)	English	ϕ(e|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	...	...

Slide from Koehn 2008

How to learn the phrase translation table?

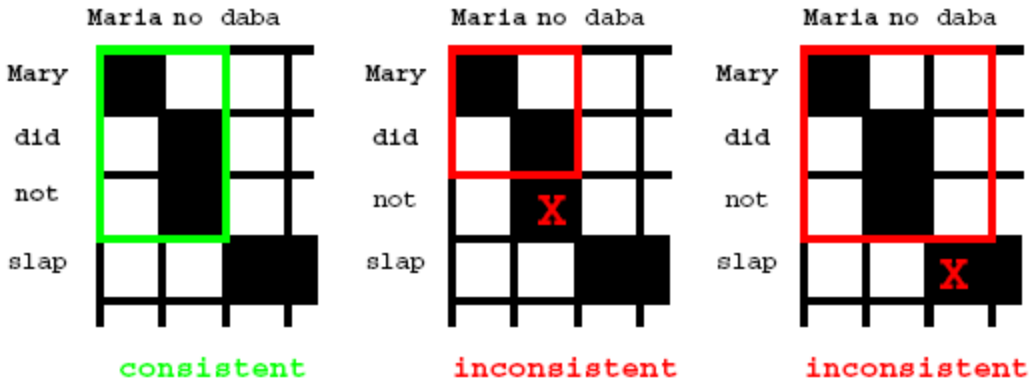
- Start with the *word alignment*:



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

Slide from Koehn 2008

Consistent with word alignment



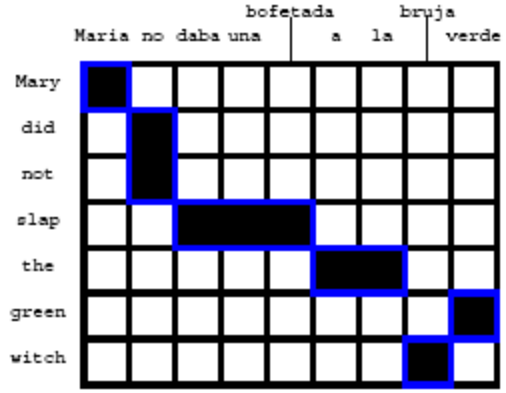
- **Consistent with the word alignment** :=

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

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

Slide from Koehn 2008

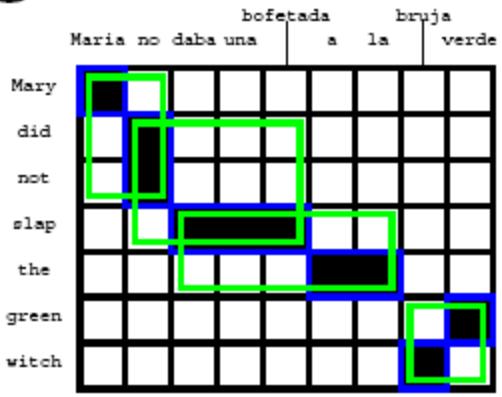
Word alignment induced phrases



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

Slide from Koehn 2008

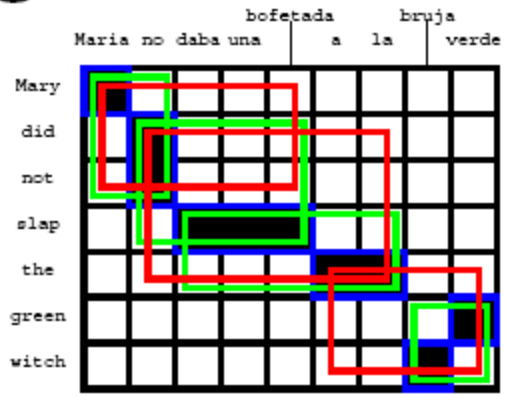
Word alignment induced phrases



(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)

Slide from Koehn 2008

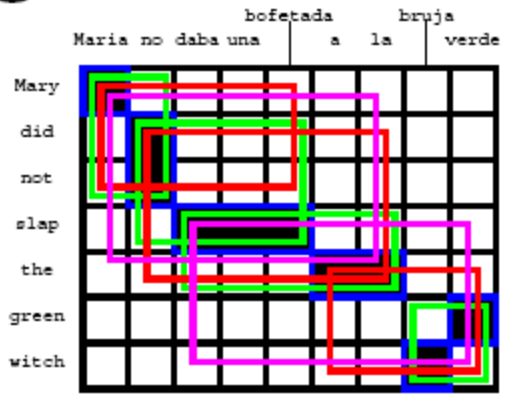
Word alignment induced phrases



(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)

Slide from Koehn 2008

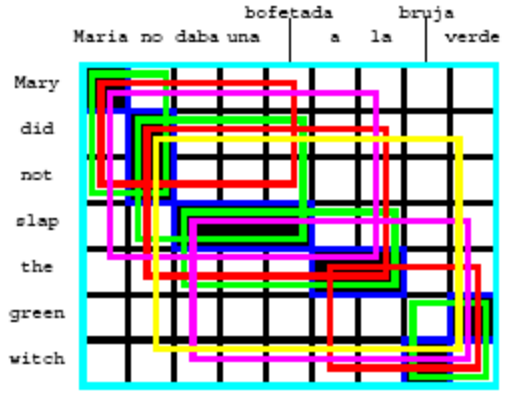
Word alignment induced phrases



(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)

Slide from Koehn 2008

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)

Slide from Koehn 2008

Probability distribution of phrase pairs

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

$\Rightarrow$ Possible *choices*

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

Slide from Koehn 2008

Reordering

- *Monotone* translation
  - do not allow any reordering

  → worse translations

- *Limiting* reordering (to movement over max. number of words) helps

- *Distance-based* reordering cost
  - moving a foreign phrase over n words: cost z^n

- *Lexicalized* reordering model

Slide from Koehn 2008