Phrase Alignment

Monday, February 23, 2015

Plan for Today:

- Wrap up word alignment
- Phrase tables

Slides from Dave Kauchak, Philipp Koehn and Dan Klein (via Luke Zettlemoyer)

Implementation Details

Each foreign word can be aligned to any of the English words (or NULL)

$(|E|+1)^{|F|}$

Repeat:

E-step

- Enumerate all possible alignments
- Calculate how probable the alignments are under the current model (i.e. p(f|e))

M-step

- Recalculate p(f|e) using counts from **all** alignments, **weighted** by how probable they are

Without the Alignments

p(f -> e): probability that f is aligned to e *in this pair*

a b c

y z

Of all things that y could align to, how likely is it to be a:

$$\frac{p(y \mid a)}{p(y \mid a) + p(y \mid b) + p(y \mid c)}$$

Without the Alignments

Input: corpus of English/Foreign sentence pairs along with alignment

```
for (E, F) in corpus:
    for e in E:
        for f in F:
            p(f -> e) = p(f | e) / ( sum_(e in E) p( f | e) )
            count(e,f) += p( f -> e)
            count(e) += p(f -> e)

for all (e,f) in count:
    p(f|e) = count(e,f) / count(e)
```

Getting better alignments…

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} \prod_{i=1}^{m} p(a_i) \times p(e_i \mid f_{a_i})$$

Model 2 $$= \sum_{\mathbf{a} \in [0,n]^m} \prod_{i=1}^{m} p(a_i \mid i, m, n) \times p(e_i \mid f_{a_i})$$

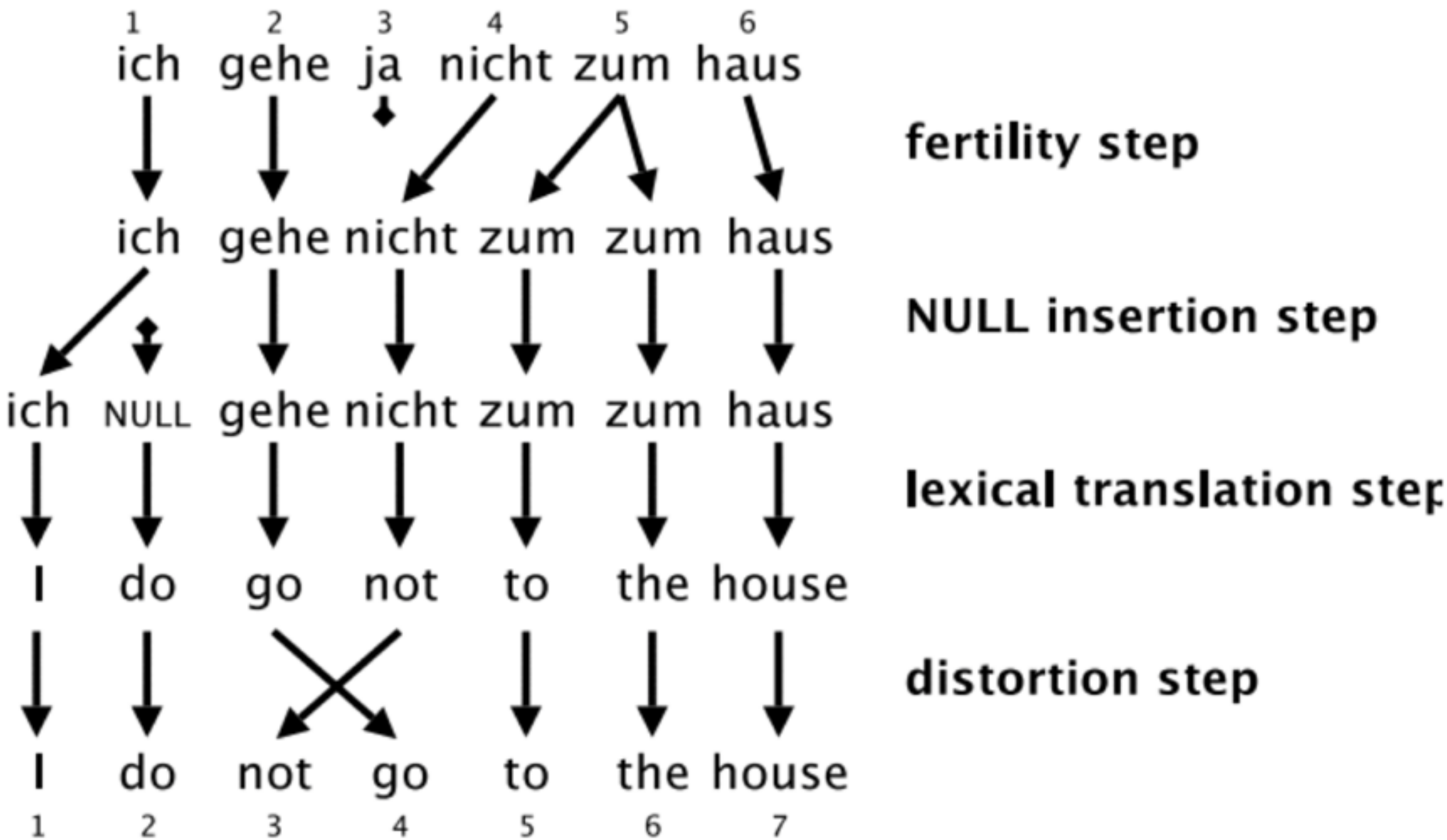
HMM $$= \sum_{\mathbf{a} \in [0,n]^m} \prod_{i=1}^{m} p(a_i \mid a_{i-1}) \times p(e_i \mid f_{a_i})$$

We'll hear more about this method from team HMM!

Fertility Models

- The models we have considered so far have been efficient
- This efficiency has come at a modeling cost:
  - What is to stop the model from "translating" a word 0, 1, 2, or 100 times?
- We introduce *fertility models* to deal with this

IBM Model 3/4/5



Fertility

- Fertility: the number of English words generated by a foreign word
- Modeled by categorical distribution $n(\phi \mid f)$
- Examples:

Unabhaengigkeitserklaerung

0	0.00008
1	0.1
2	0.0002
3	**0.8**
4	0.009
5	0

zum = (zu + dem)

0	0.01
1	0
2	**0.9**
3	0.0009
4	0.0001
5	0

Haus

0	0.01
1	**0.92**
2	0.07
3	0
4	0
5	0

Fertility

$$p(\mathbf{e} \mid \mathbf{f}, m) = \sum_{\mathbf{a} \in [0,n]^m} p(\mathbf{a} \mid \mathbf{f}, m) \times \prod_{i=1}^{m} p(e_i \mid f_{a_i})$$

- Fertility models mean that we can no longer exploit conditional independencies to write $p(\mathbf{a} \mid \mathrm{f}, m)$ as a series of local alignment decisions.
- The solution is beyond our scope — practical solution involves initializing with IBM-Model 2

Lexical Translation

- IBM Models 1-5 [Brown et al., 1993]
  - Model 1: lexical translation, uniform alignment
  - Model 2: absolute position model
  - Model 3: fertility
  - Model 4: relative position model (jumps in target string)
  - Model 5: non-deficient model
- HMM translation model [Vogel et al., 1996]
  - Relative position model (jumps in source string)
- Latent variables are more useful these days than the translations
- Widely used Giza++ toolkit

When lexical translation fails…

Translational Equivalence

Er hat die Prüfung ***bestanden****, jedoch nur knapp*

He **insisted on** the test, but just barely.

He **passed** the test, but just barely.

How do lexical translation models deal with contextual information?

Translational Equivalence

Er hat die Prüfung ***bestanden****, jedoch nur knapp*

He insisted on the test, but just barely.

He passed the test, but just barely.

F	E	prob
bestanden	**insisted**	0.06
	were	0.06
	existed	0.04
	was	0.04
	been	0.04
	passed	0.03
	consist	0.01

Translational Equivalence

Er hat die Prüfung ***bestanden****, jedoch nur knapp*

He **insisted on** the test, but just barely.

He **passed** the test, but just barely.

Lexical Translation

What is wrong with this?

How can we improve this?

Translation model

- What are the atomic units?
  - Lexical translation: **words**
  - Phrase-based translation: **phrases**
- **Standard model used by Google, Microsoft ...**
- Benefits
  - many-to-many translation
  - use of local context in translation
- Downsides
  - Where do phrases comes from?

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

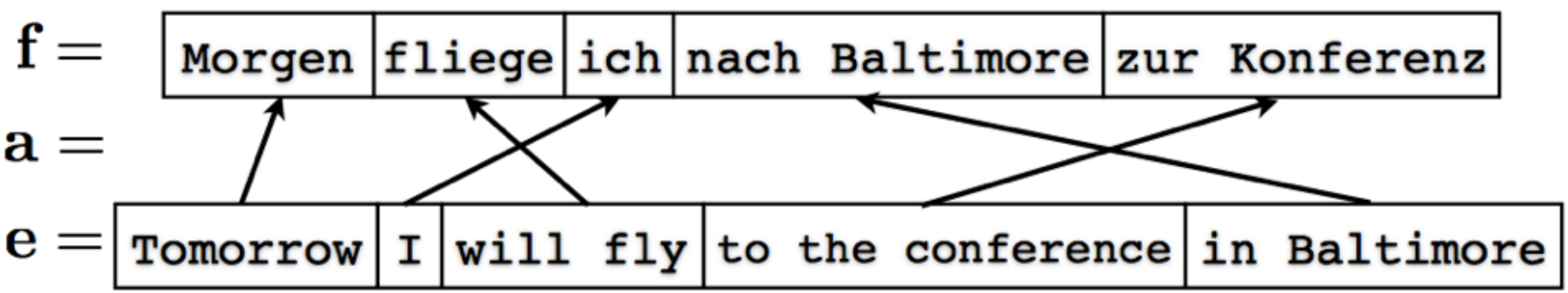
$\mathbf{f} =$ `Morgen fliege ich nach Baltimore zur Konferenz`

$\mathbf{e} =$ `Tomorrow I will fly to the conference in Baltimore`

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

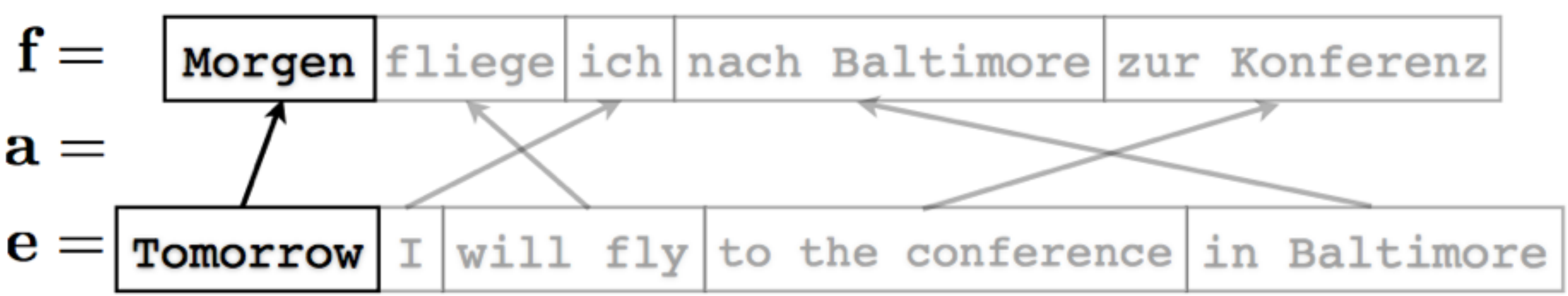
$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

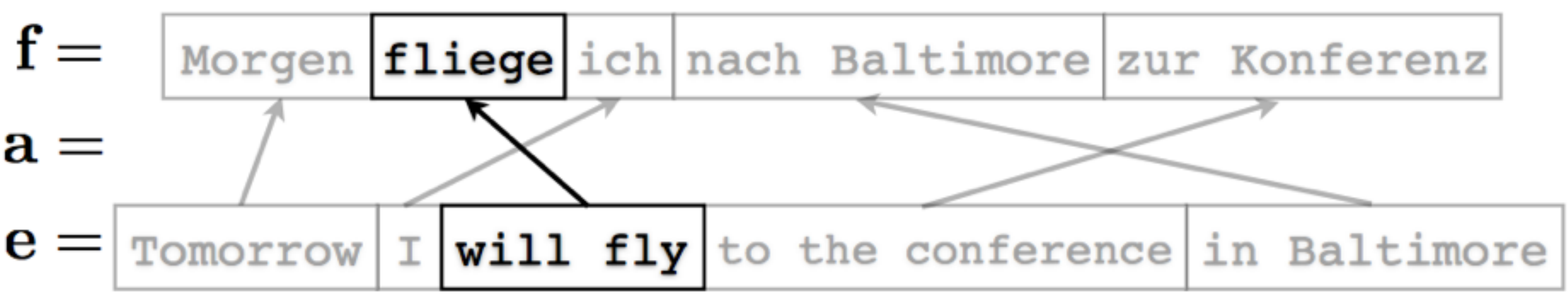


p(Morgen|Tomorrow)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

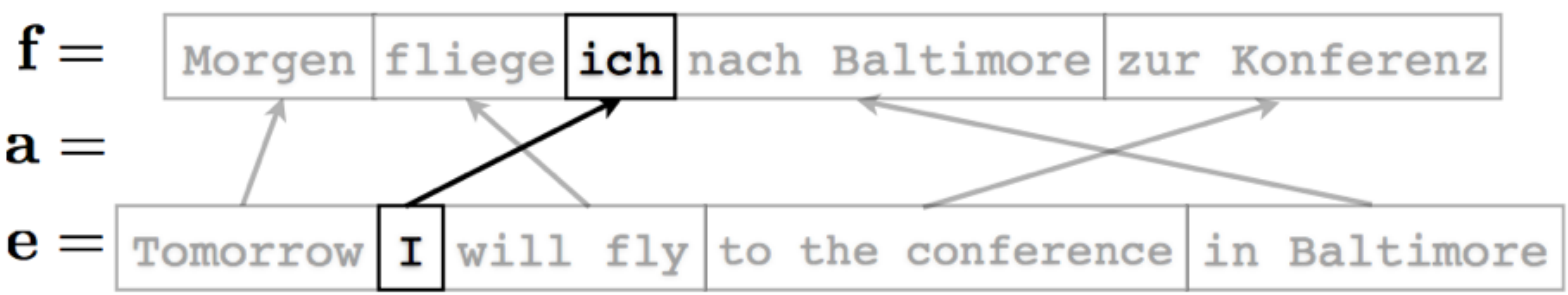


p(Morgen|Tomorrow) x *p*(fliege|will fly)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$



p(Morgen|Tomorrow) x *p*(fliege|will fly) x *p*(ich|I)

Translation model

- With a **latent variable**, we introduce a decomposition into **phrases** which translate **independently**:

$$p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

Marginalize to get p(f|e):

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a} \in \mathcal{A}} p(\mathbf{a}) \prod_{\langle \bar{\mathbf{e}}, \bar{\mathbf{f}} \rangle \in \mathbf{a}} p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$$

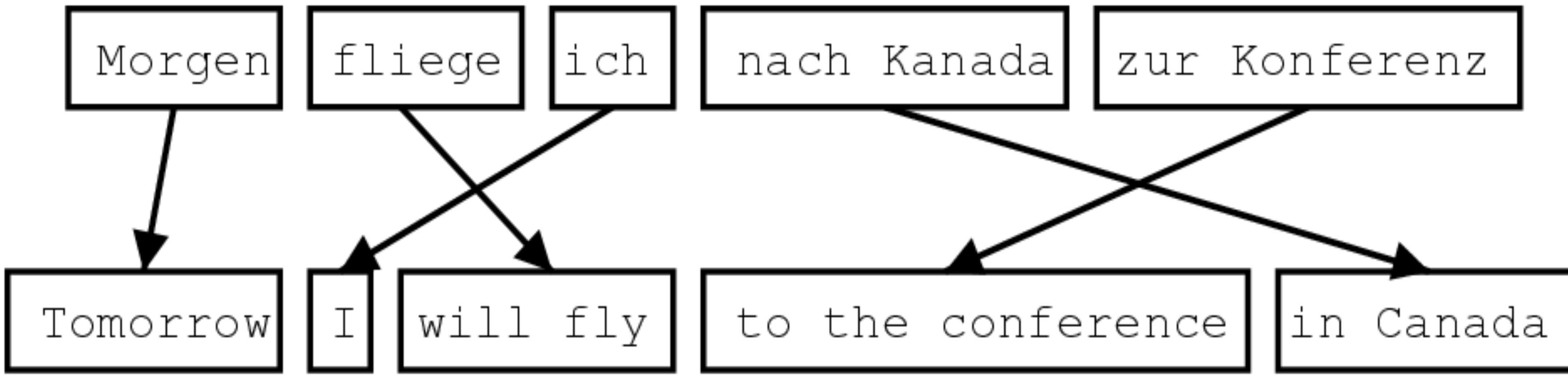
Phrases

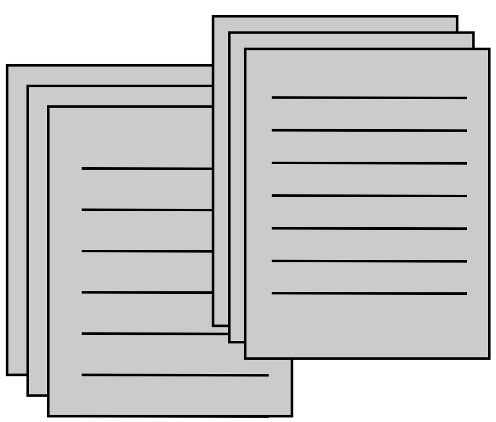
- Contiguous strings of words
- Phrases are not necessarily syntactic constituents
- Usually have maximum limits
- Phrases subsume words (individual words are phrases of length 1)

Phrase Tables

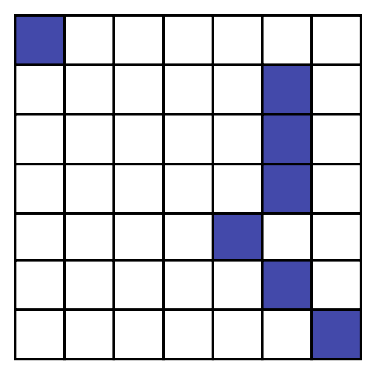
$\bar{\mathbf{f}}$	$\bar{\mathbf{e}}$	$p(\bar{\mathbf{f}} \mid \bar{\mathbf{e}})$
das Thema	the issue	0.41
	the point	0.72
	the subject	0.47
	the thema	0.99
es gibt	there is	0.96
	there are	0.72
morgen	tomorrow	0.9
fliege ich	will I fly	0.63
	will fly	0.17
	I will fly	0.13

Phrase-Based Systems





Sentence-aligned corpus



Word alignments

cat ||| chat ||| 0.9
the cat ||| le chat ||| 0.8
dog ||| chien ||| 0.8
house ||| maison ||| 0.6
my house ||| ma maison ||| 0.9
language ||| langue ||| 0.9
…

Phrase table (translation model)

Phrase Translation Tables

- Defines the space of possible translations
  - each entry has an associated “probability”
- One learned example, for “den Vorschlag” from Europarl data

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	...	...

  - This table is noisy, has errors, and the entries do not necessarily match our linguistic intuitions about consistency….

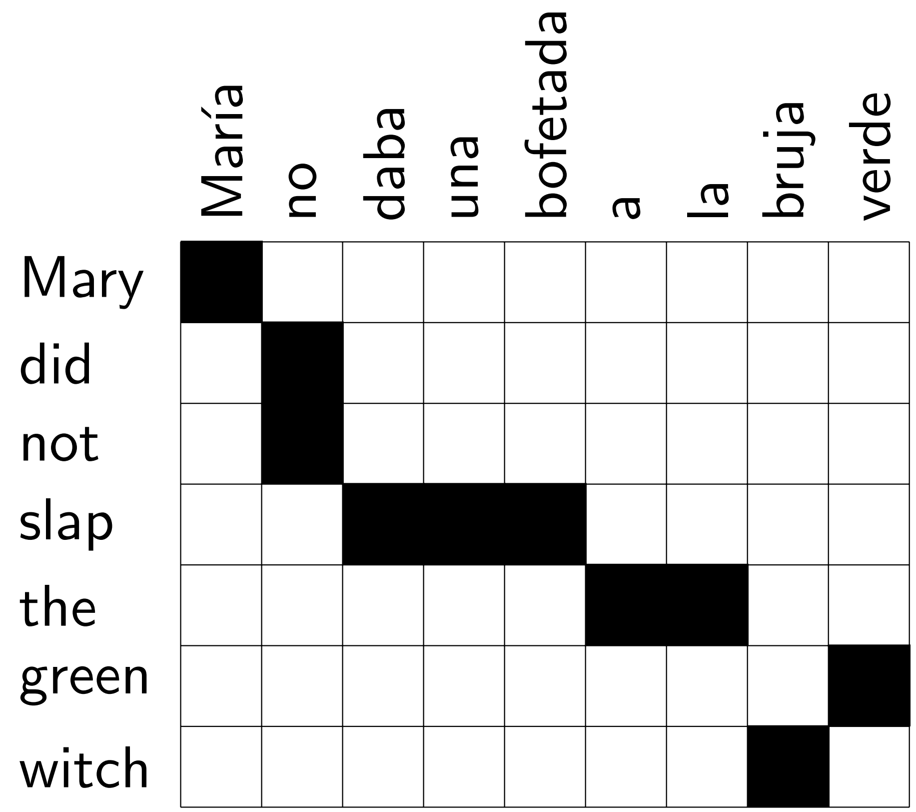
Phrase-Based Decoding

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

Decoder design is important: [Koehn et al. 03]

Extracting Phrases

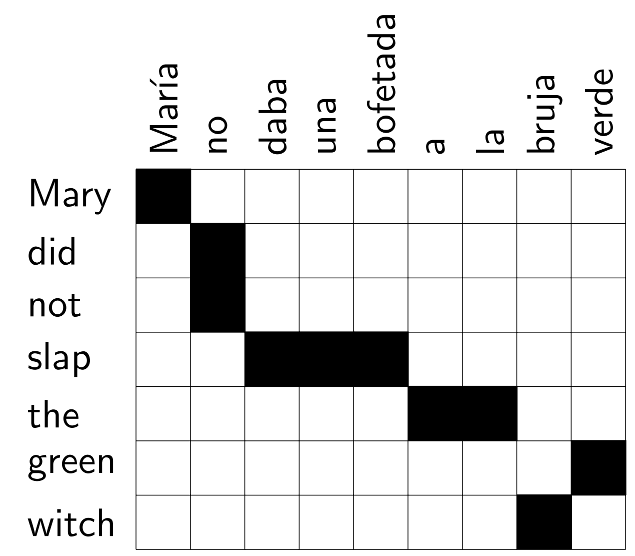
- We will use word alignments to find phrases

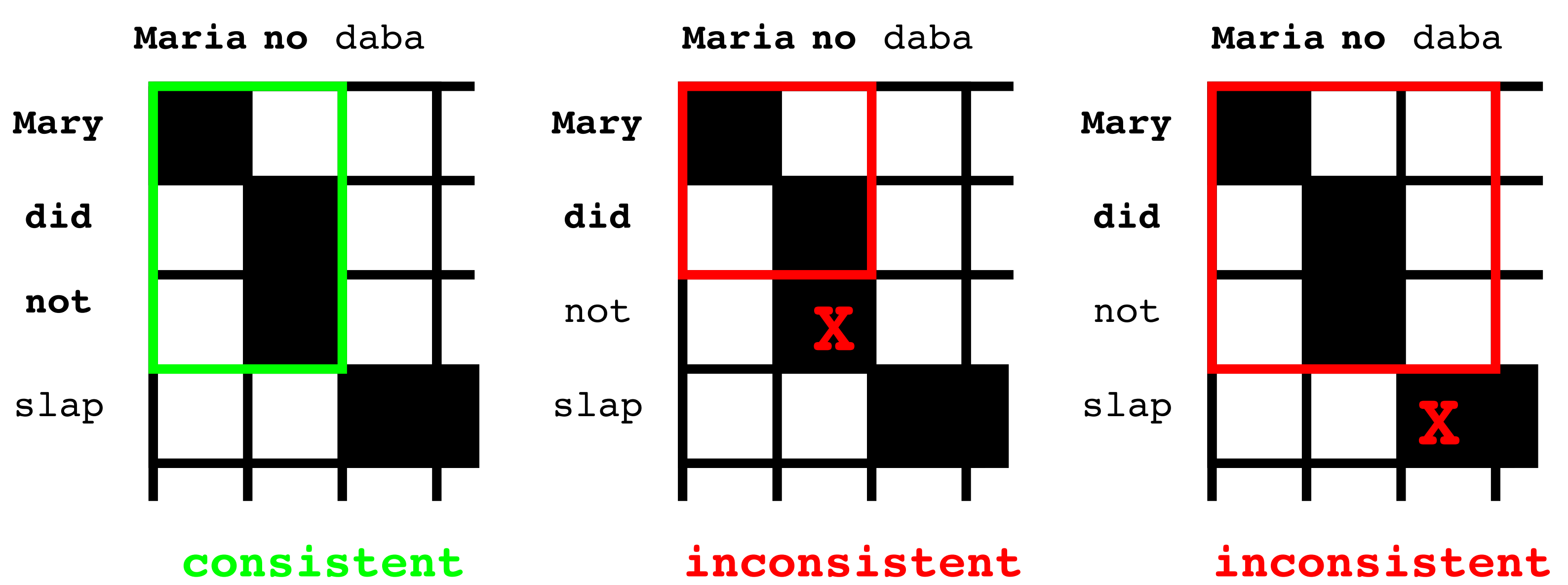


- Question: what is the best set of phrases?

Extracting Phrases

- Phrase alignment must
  - Contain at least one alignment edge
  - Contain all alignments for phrase pair





- Extract all such phrase pairs!

Phrase Pair Extraction Example

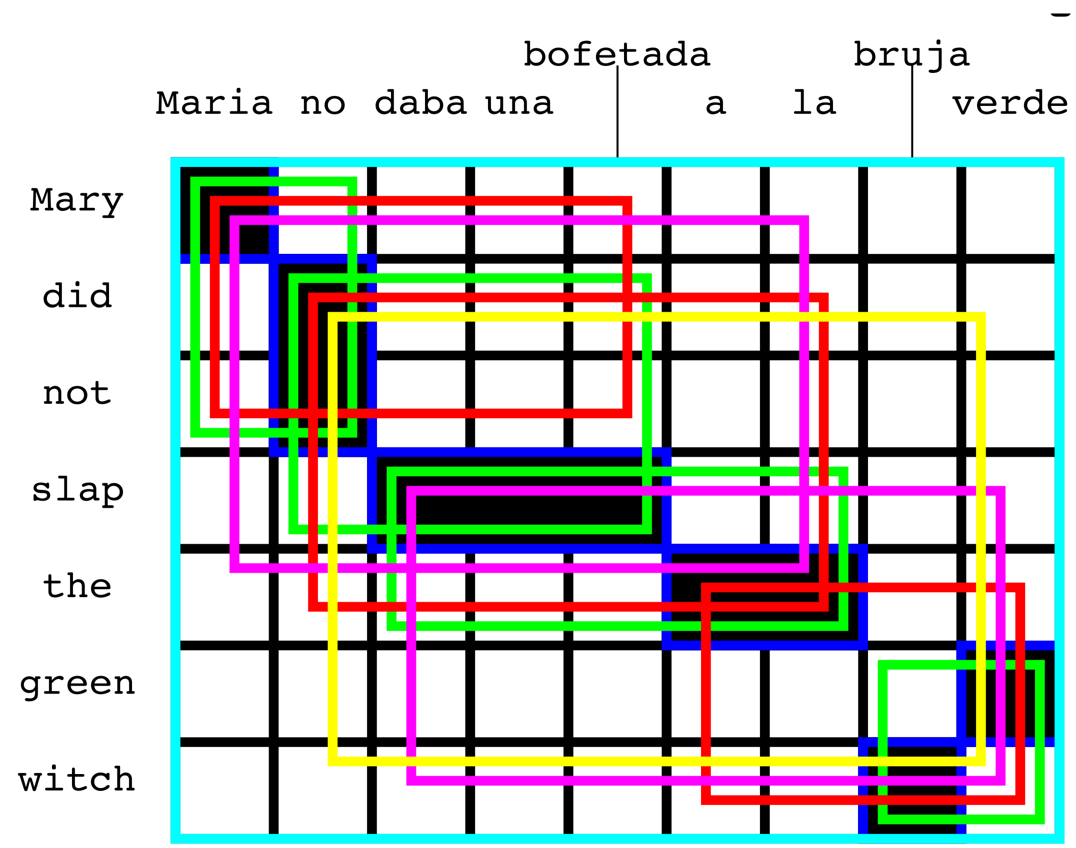
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

(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)

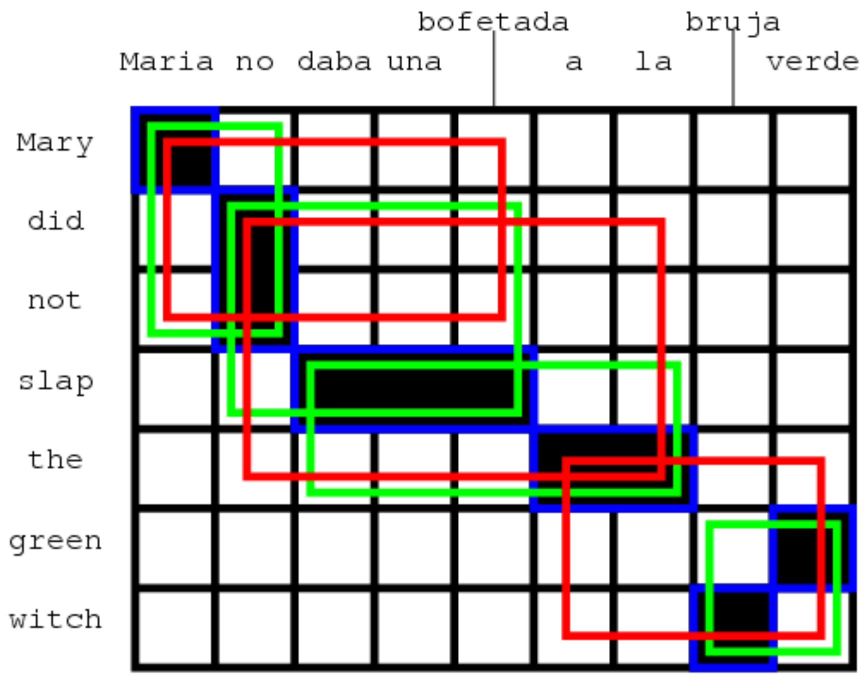
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch)

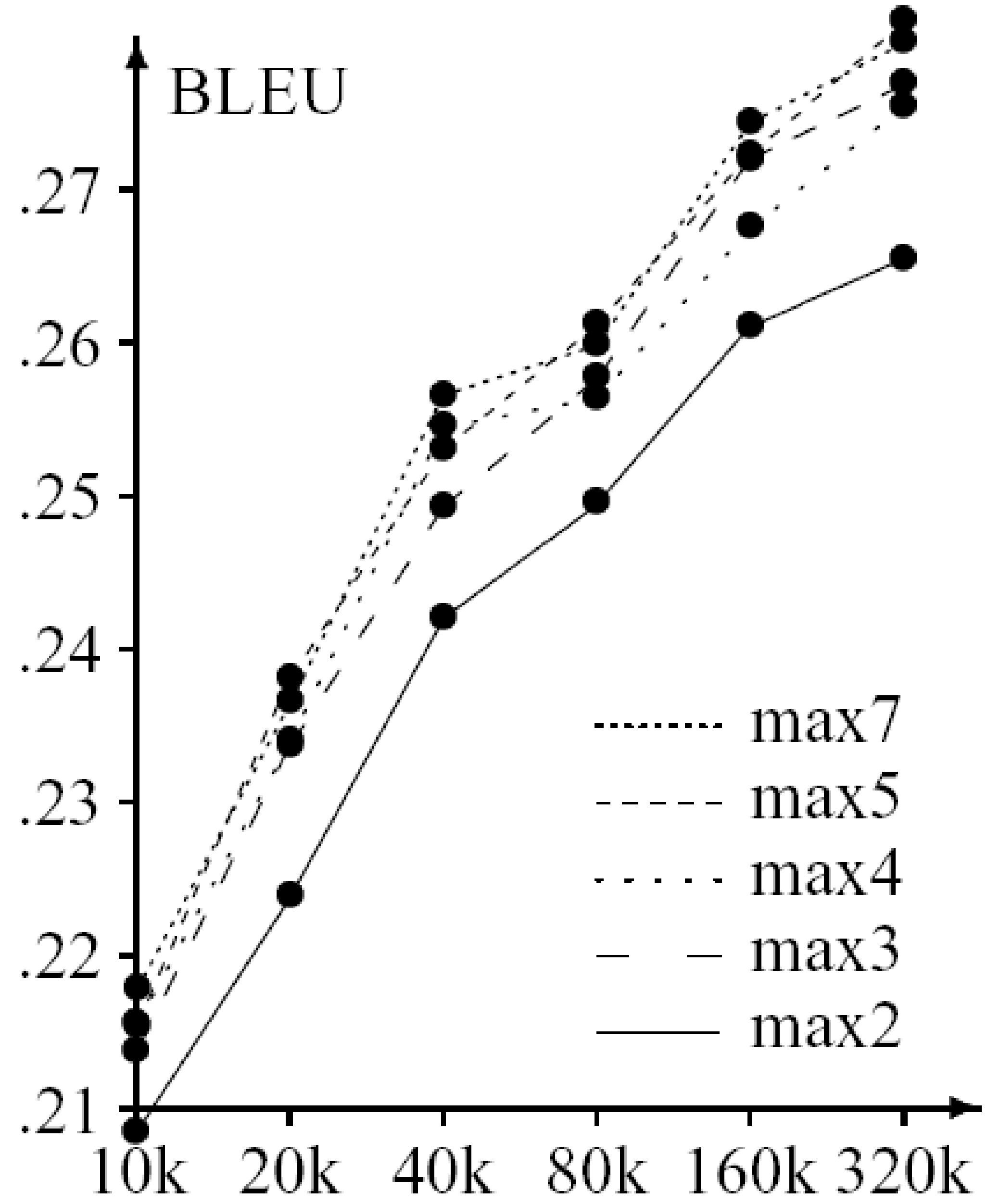
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)



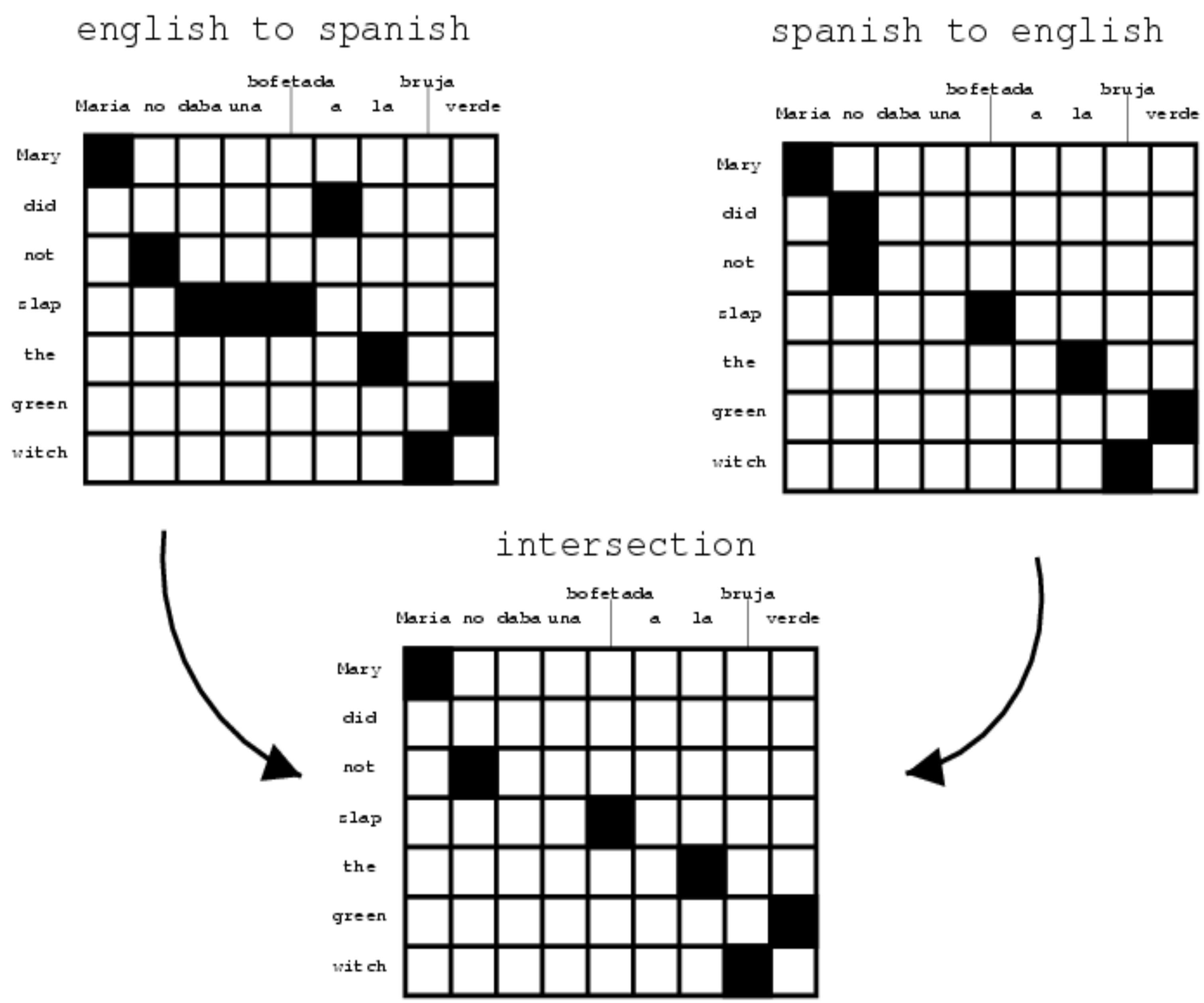
Phrase Size

- Phrases do help
  - But they don't need to be long
  - Why should this be?



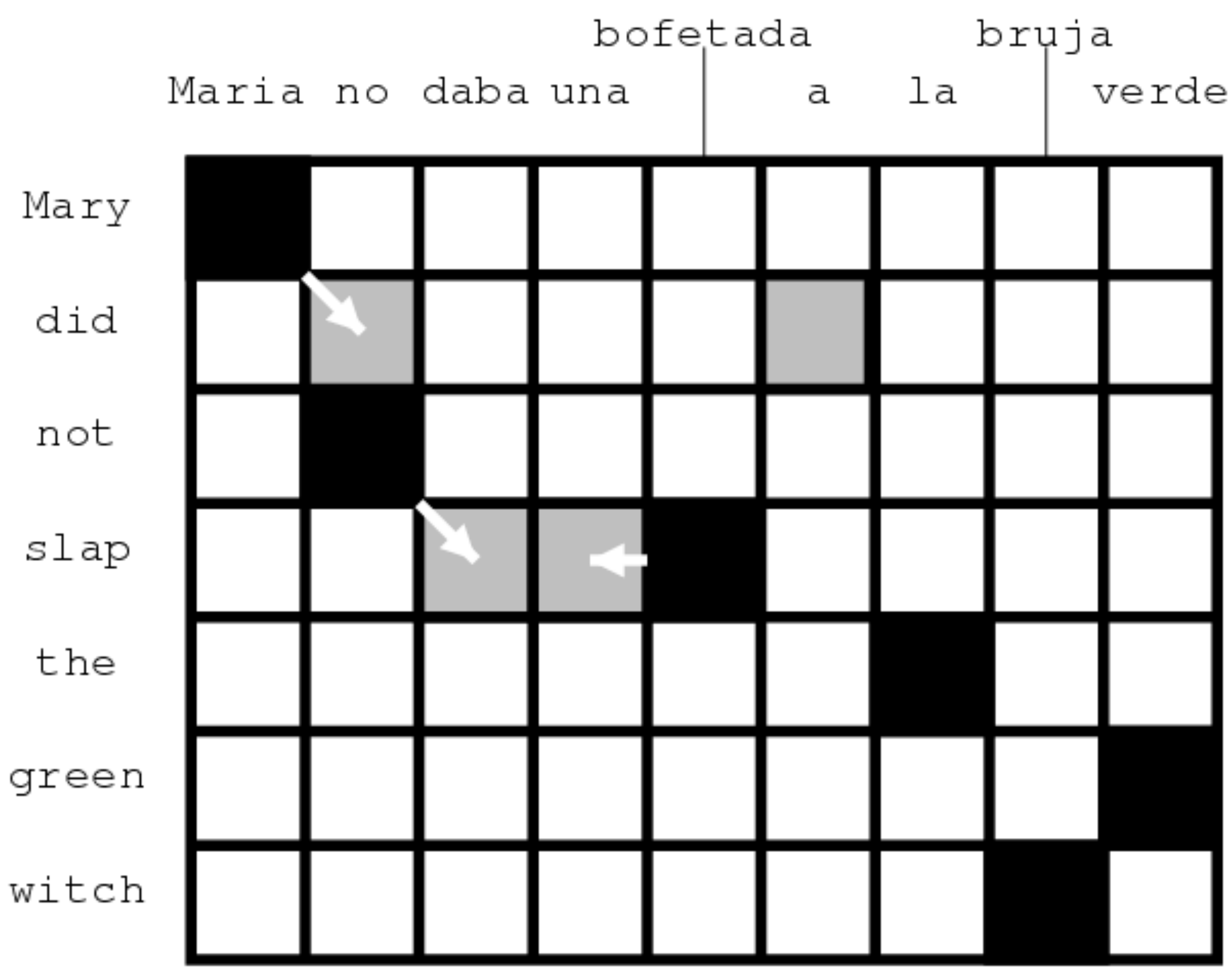


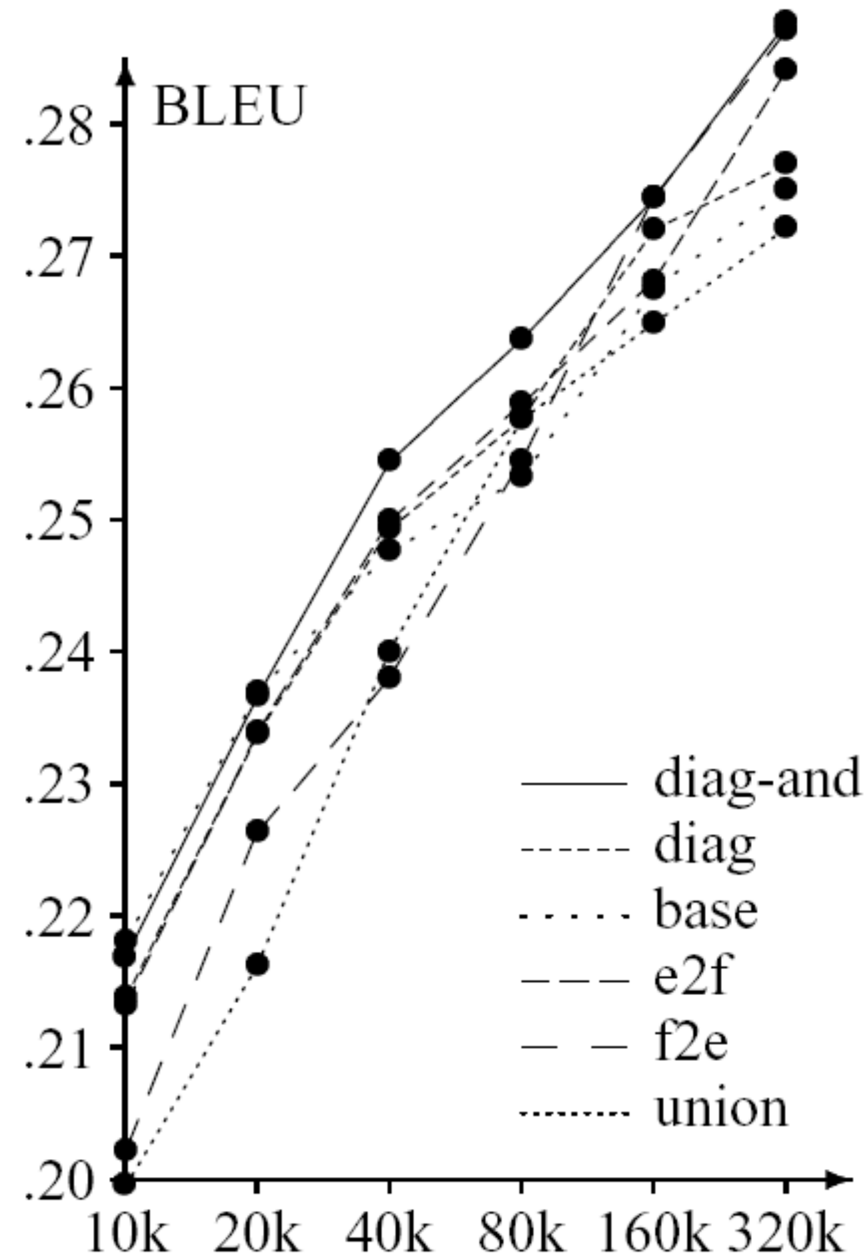
Bidirectional Alignment



We'll hear more about this method from team Combination!

Alignment Heuristics





Looking Forward

Midterm

Available in class Monday, March 2

Due back by start of class Monday, March 9

- If you finish earlier, return to Prof. Medero as soon as you're done!

75 minute take-home exam

- Closed book & notes
- Honor code applies

One option for exam time: No class on Wednesday, March 4.

After the exam…

Week of March 9: The p(e) in our translation equations

After spring break: Decoding and evaluation