89688: Statistical Machine Translation

Phrase-Based Models and Decoding

May 2020

Roee Aharoni
Computer Science Department
Bar Ilan University

Based in part on slides by Phillip Koehn

More on word-based models

More on word-based models

- So far, we discussed IBM Model1 - ignores word positions

$$p(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) =$$

$$= \prod_{i=1}^{m} \frac{1}{(l+1)} \times t(f_i | e_{a_i}) = \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

More on word-based models

- So far, we discussed IBM Model1 - ignores word positions
- The additional IBM models are:

$$p(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) =$$

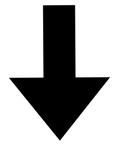
$$= \prod_{i=1}^{m} \frac{1}{(l+1)} \times t(f_i | e_{a_i}) = \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

More on word-based models

- So far, we discussed IBM Model1 - ignores word positions
- The additional IBM models are:
  - Model 2 - Word positions matter! (and Model1 is a special case)

$$p(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) =$$

$$= \prod_{i=1}^{m} \frac{1}{(l+1)} \times t(f_i | e_{a_i}) = \frac{1}{(l+1)^m} \prod_{i=1}^{m} t(f_i | e_{a_i})$$

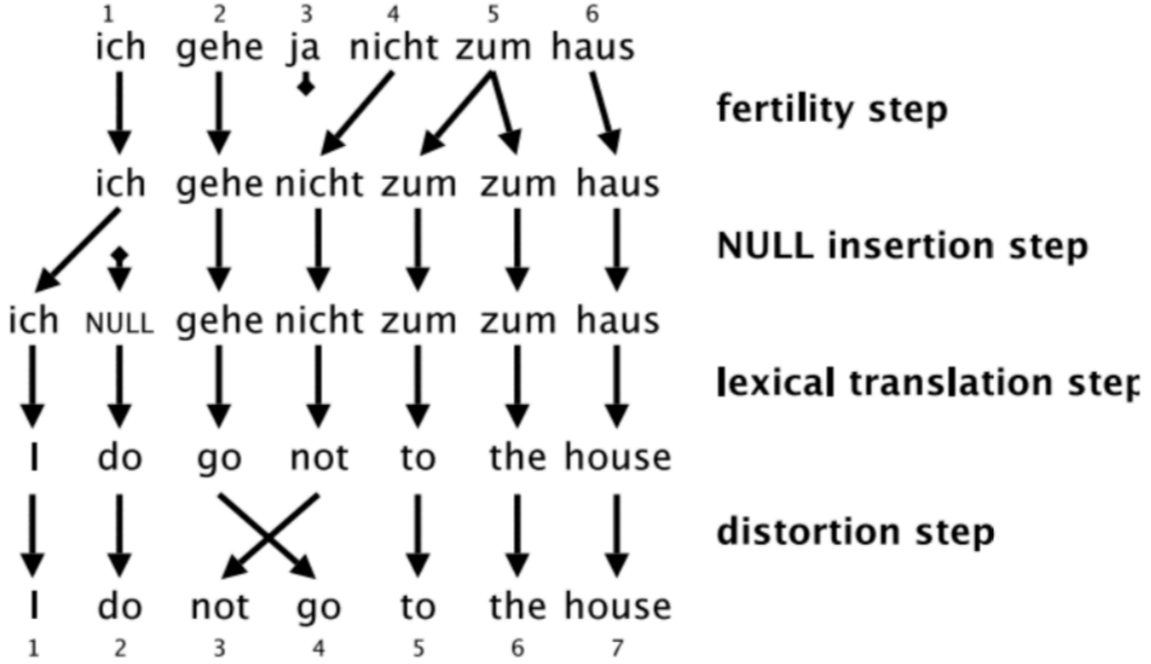


$$p(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) =$$

$$= \prod_{i=1}^{m} q(a_i | i, l, m) t(f_i | e_{a_i})$$

More on word-based models

- So far, we discussed IBM Model1 - ignores word positions
- The additional IBM models are:
  - Model 2 - Word positions matter! (and Model1 is a special case)
  - Model 3 - Introduce *Fertility*



Unabhaengigkeitserklaerung

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

zum = (zu + dem)

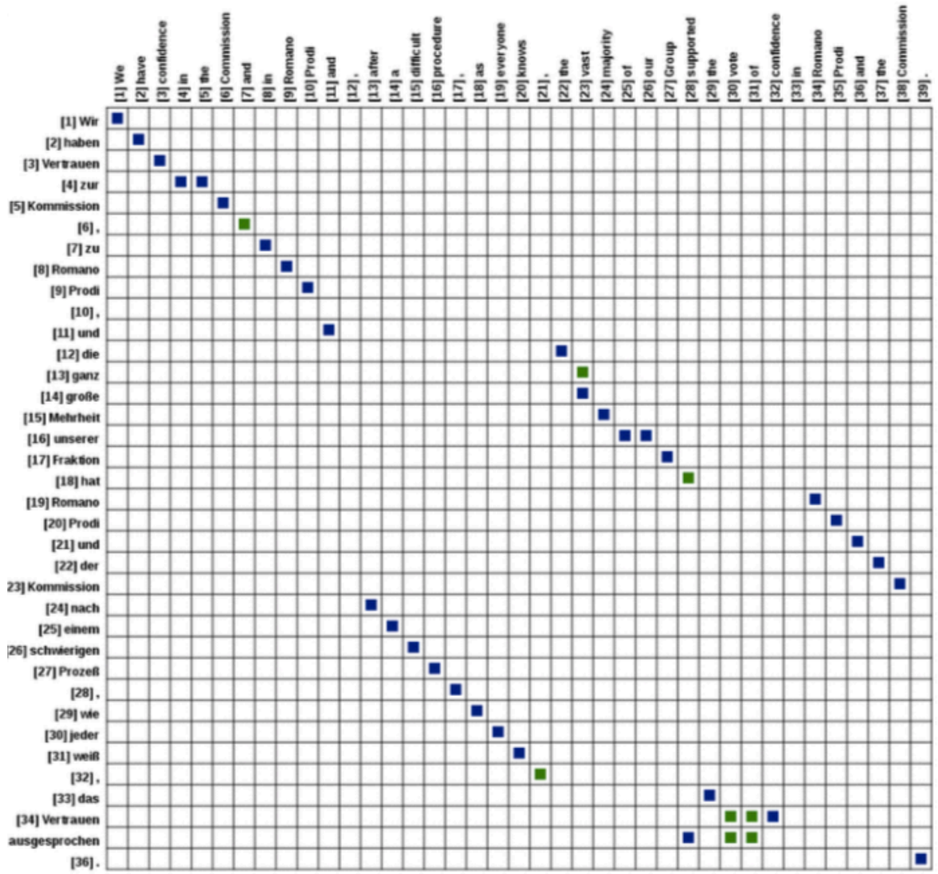
0	0.01
1	0
2	**0.9**
3	0.0009
4	0.0001
5	0

Haus

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

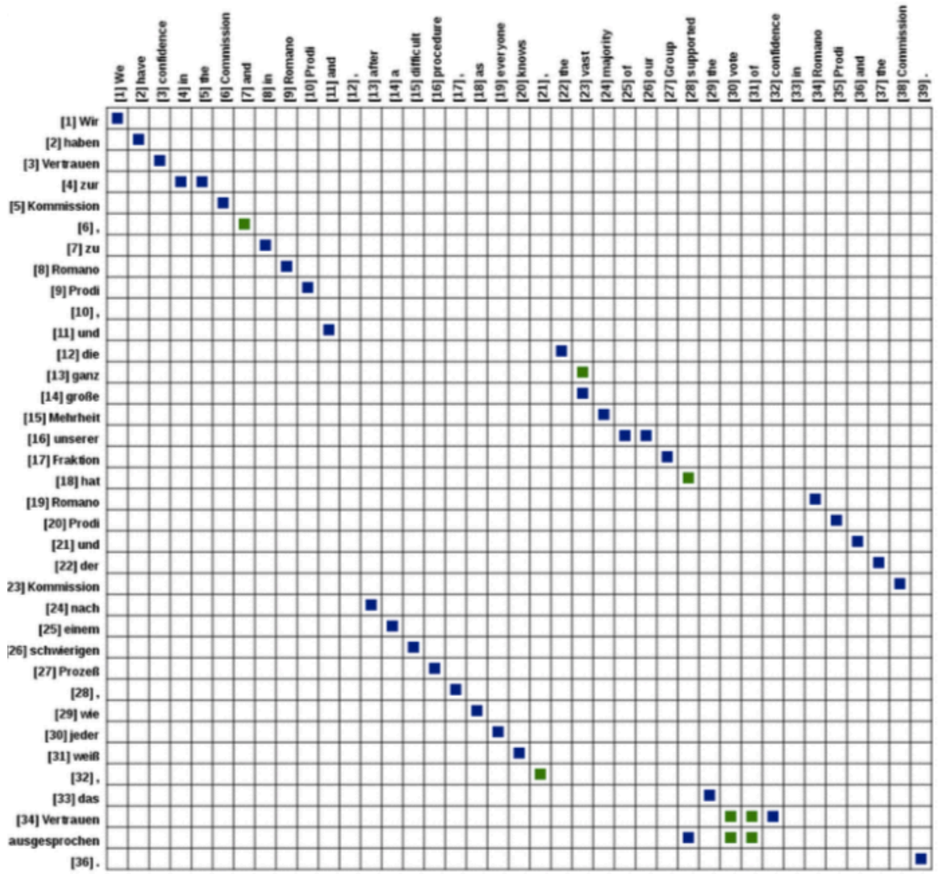
More on word-based models

- So far, we discussed IBM Model1 - ignores word positions
- The additional IBM models are:
  - Model 2 - Word positions matter! (and Model1 is a special case)
  - Model 3 - Introduce *Fertility*
  - Model 4 - *Relative* word positions



More on word-based models

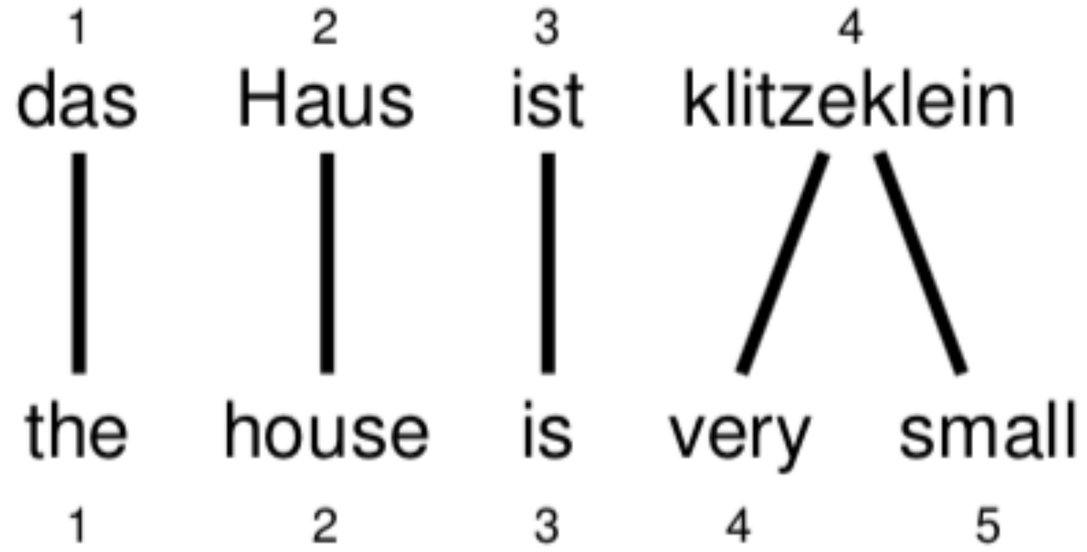
- So far, we discussed IBM Model1 - ignores word positions
- The additional IBM models are:
  - Model 2 - Word positions matter! (and Model1 is a special case)
  - Model 3 - Introduce *Fertility*
  - Model 4 - *Relative* word positions
  - Model 5 - Avoids aligning target words to the same source word



Problems with word-based models

Problems with word-based models

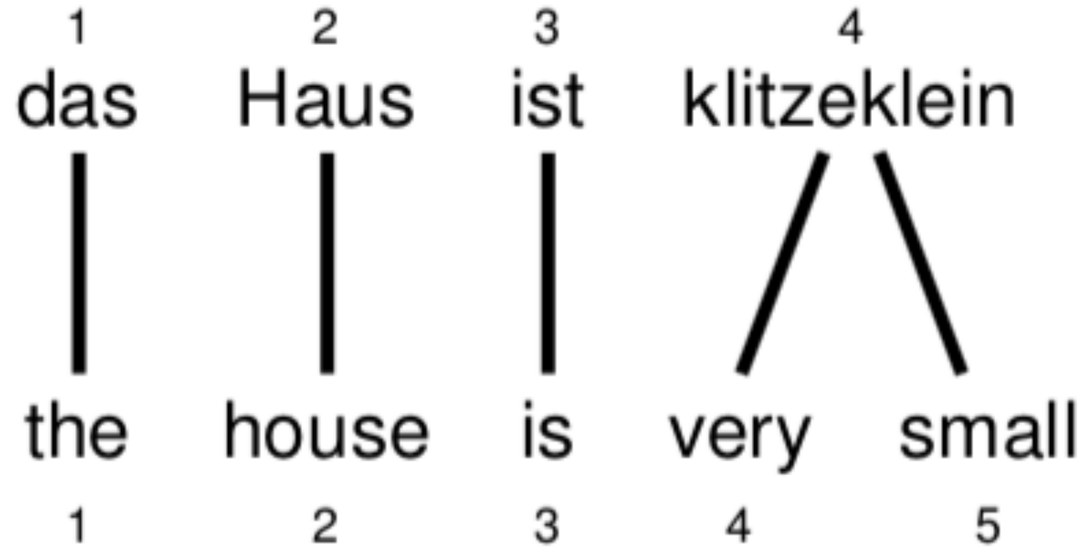
- IBM models create a many-to-one mapping



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Problems with word-based models

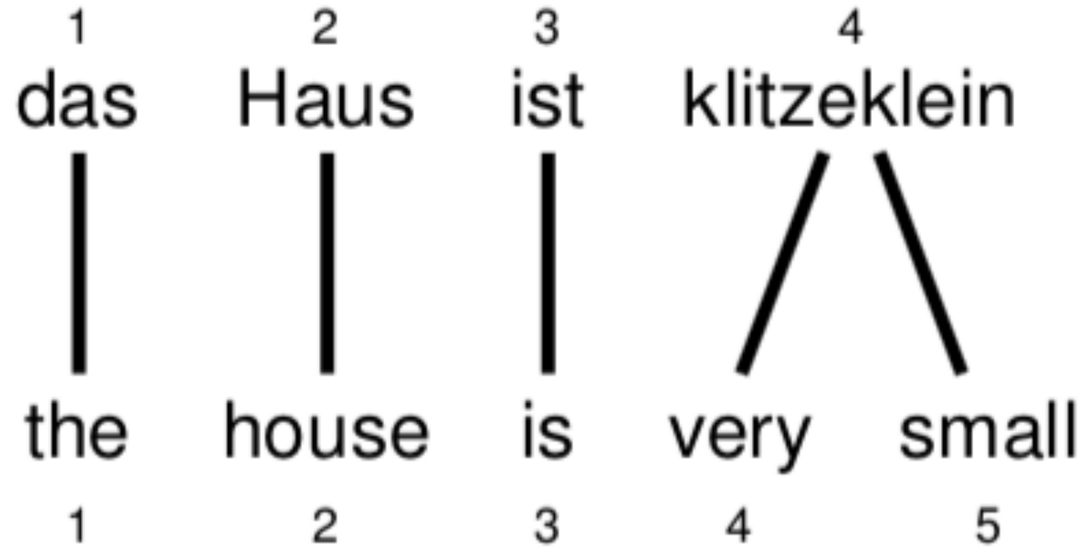
- IBM models create a many-to-one mapping
- One source word can generate several target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Problems with word-based models

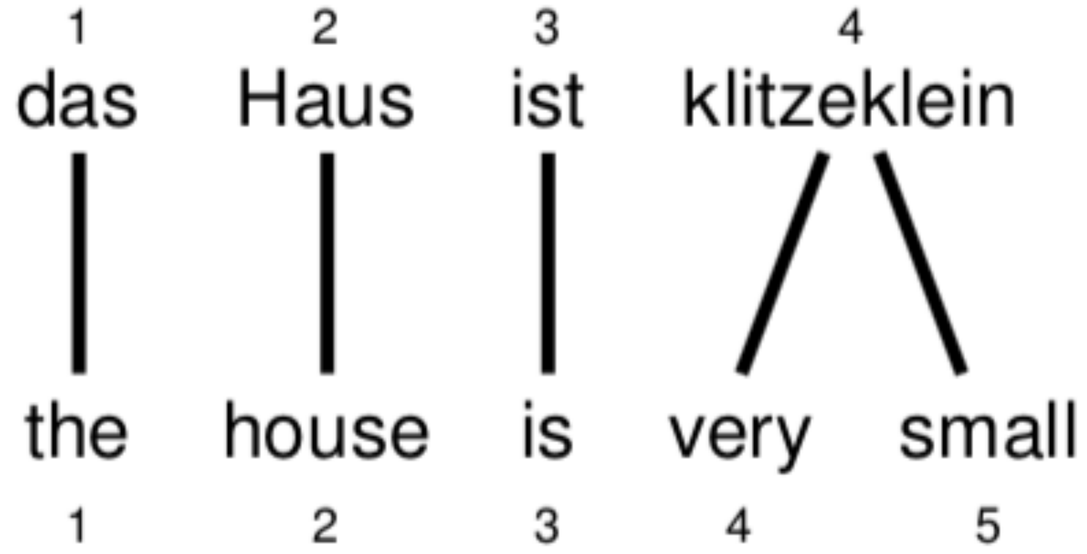
- IBM models create a many-to-one mapping
- One source word can generate several target words
- But we need many-to-many mappings…



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Problems with word-based models

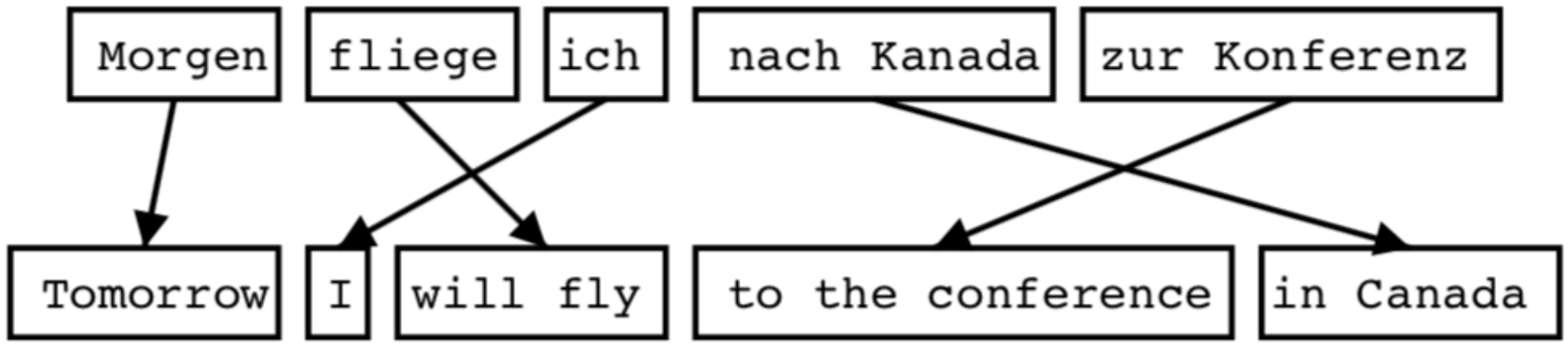
- IBM models create a many-to-one mapping
- One source word can generate several target words
- But we need many-to-many mappings…
  - “Prime Minister"



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

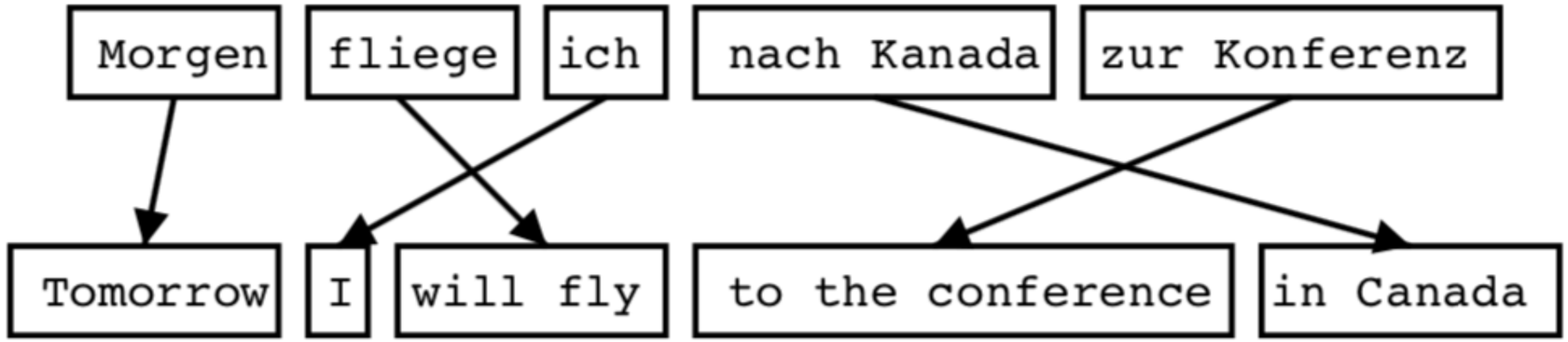
Phrase-Based Translation

Phrase-Based Translation



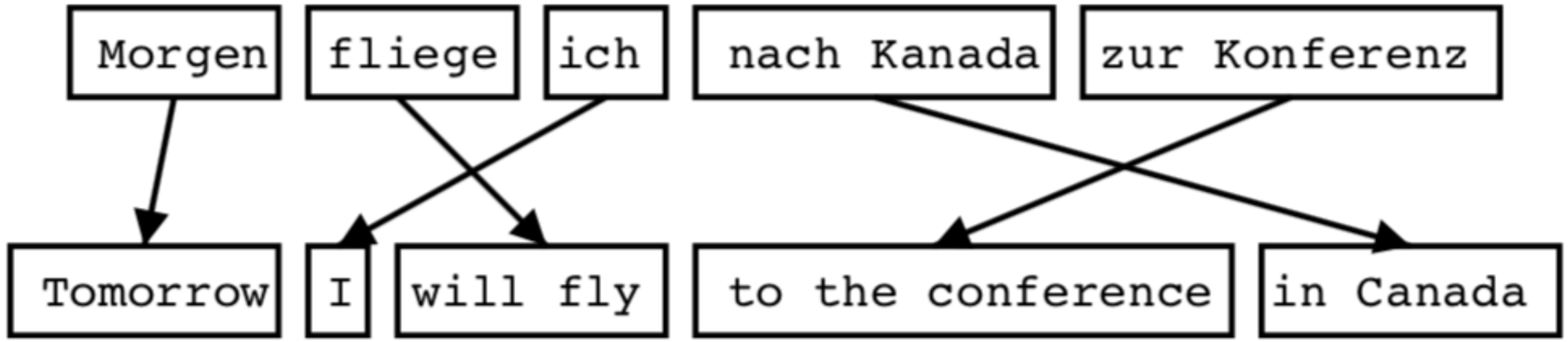
Phrase-Based Translation

- Koehn, Och & Marcu (2003)



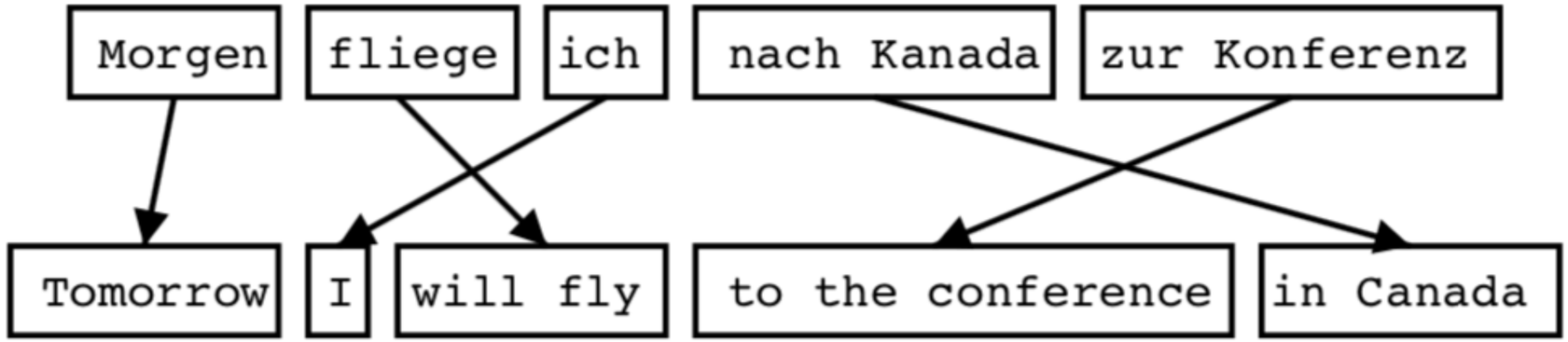
Phrase-Based Translation

- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases



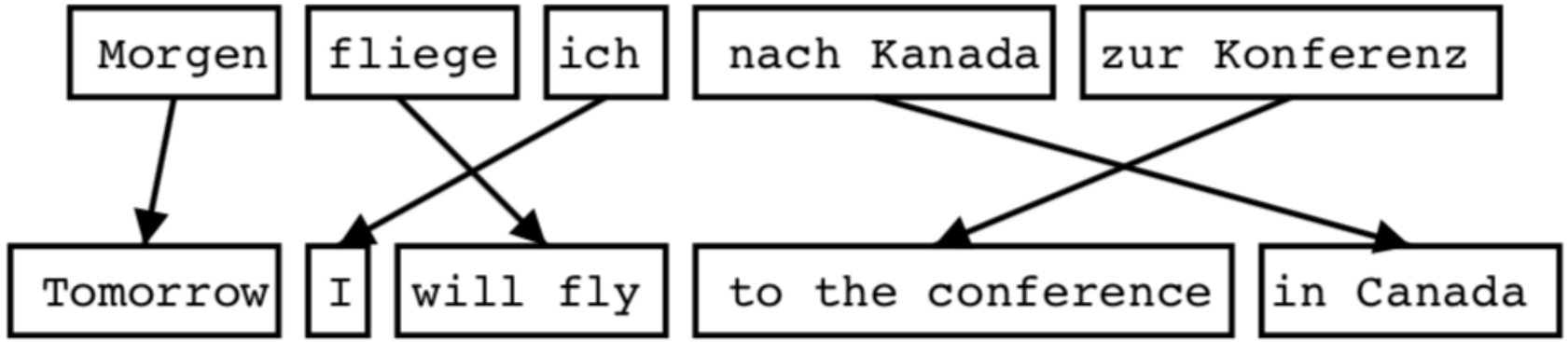
Phrase-Based Translation

- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases
  - Each phrase is translated into another phrase



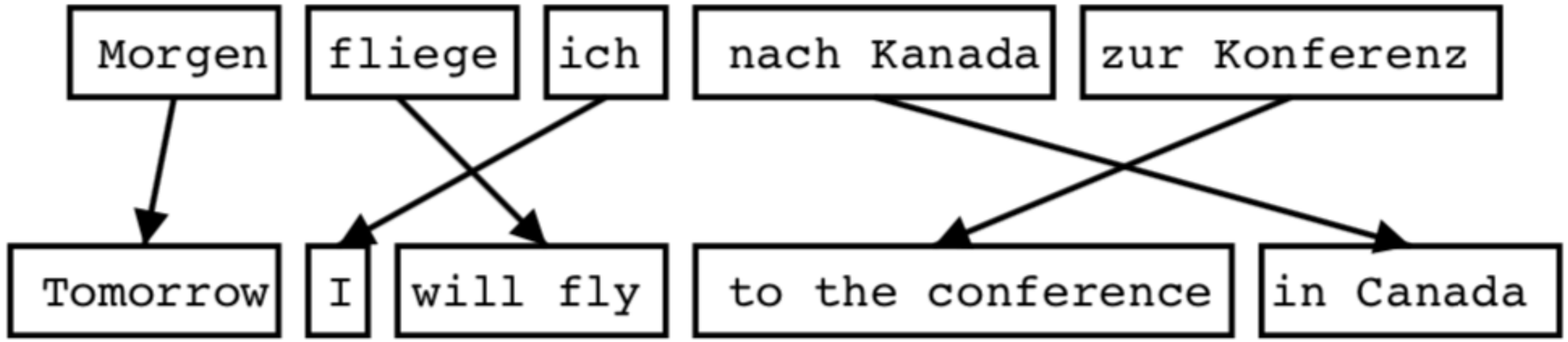
Phrase-Based Translation

- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases
  - Each phrase is translated into another phrase
  - Phrases are reordered



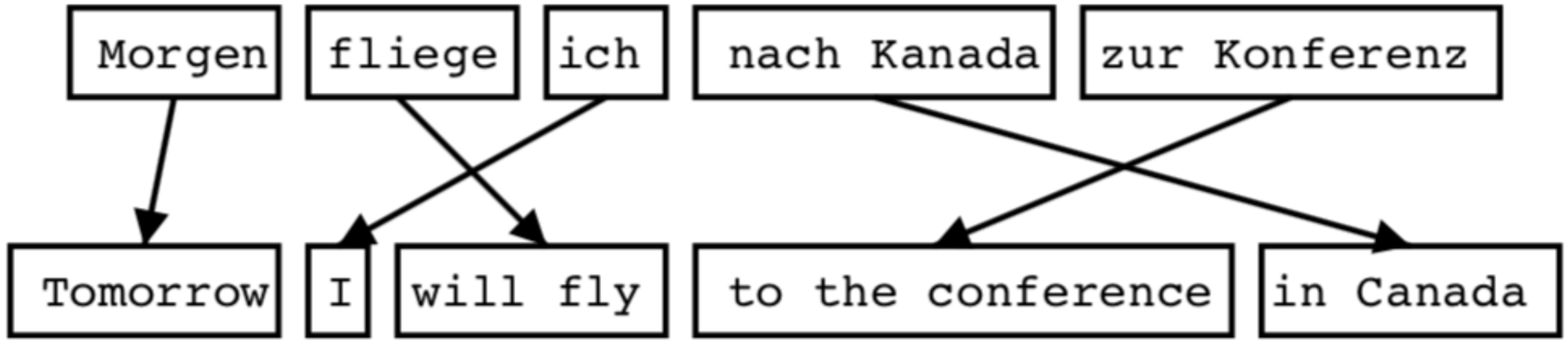
Phrase-Based Translation

- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases
  - Each phrase is translated into another phrase
  - Phrases are reordered
- Many-to-many translations can handle non-compositional cases



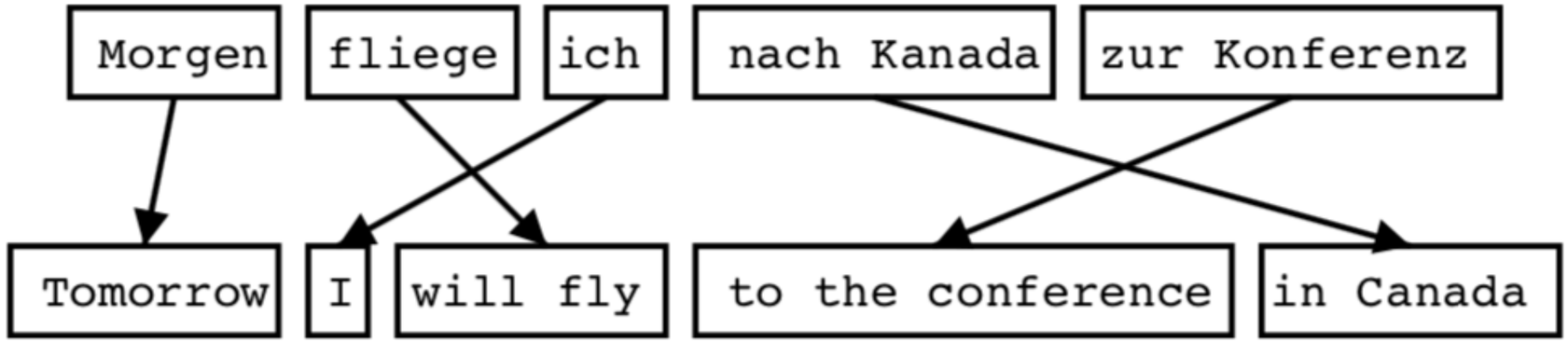
Phrase-Based Translation

- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases
  - Each phrase is translated into another phrase
  - Phrases are reordered
- Many-to-many translations can handle non-compositional cases
- Models local context



Phrase-Based Translation

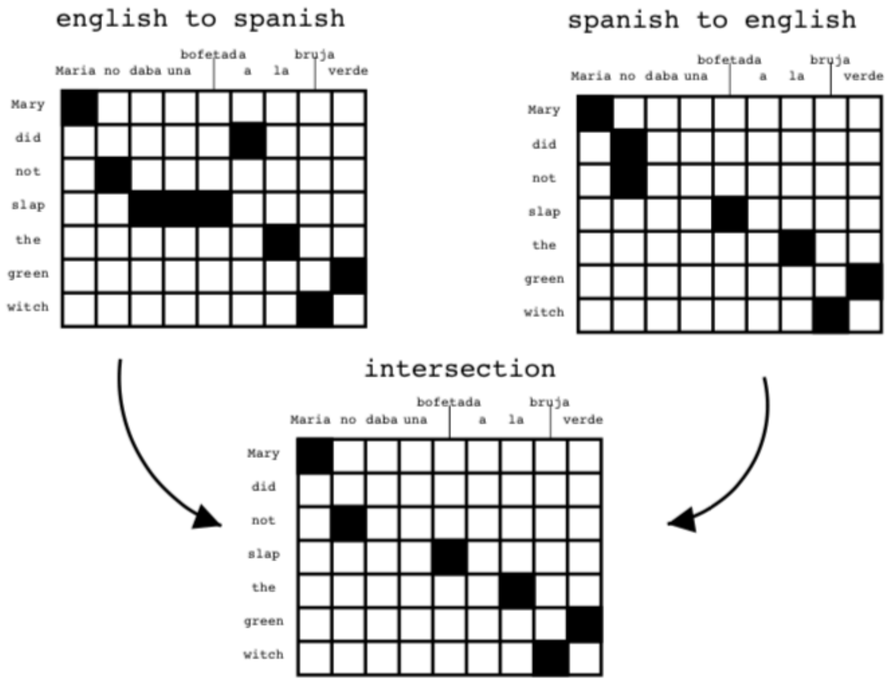
- Koehn, Och & Marcu (2003)
  - Input is segmented into phrases
  - Each phrase is translated into another phrase
  - Phrases are reordered
- Many-to-many translations can handle non-compositional cases
- Models local context
- First Google Translate models were phrase-based (2006)



How do we learn phrase translations?

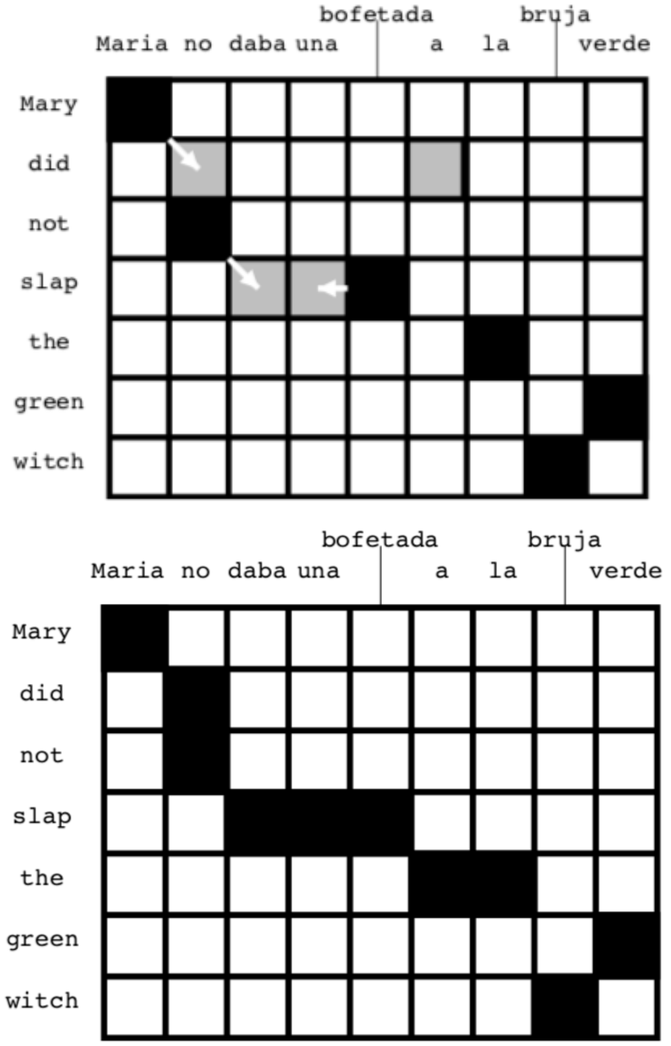
How do we learn phrase translations?

- Start with word alignments



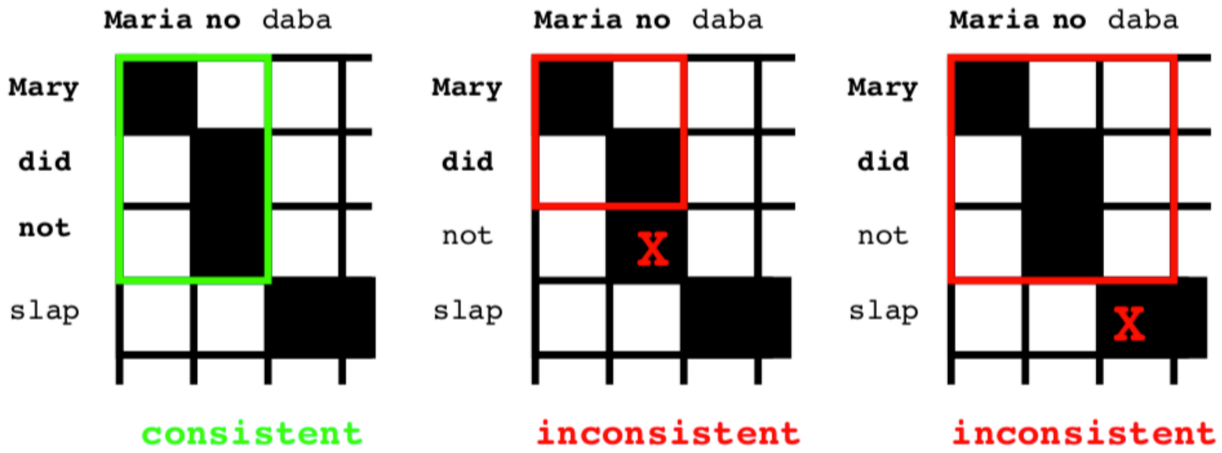
How do we learn phrase translations?

- Start with word alignments
- Preprocess using a heuristic - “Grow-Diag-Final” (Och & Ney, 2003)



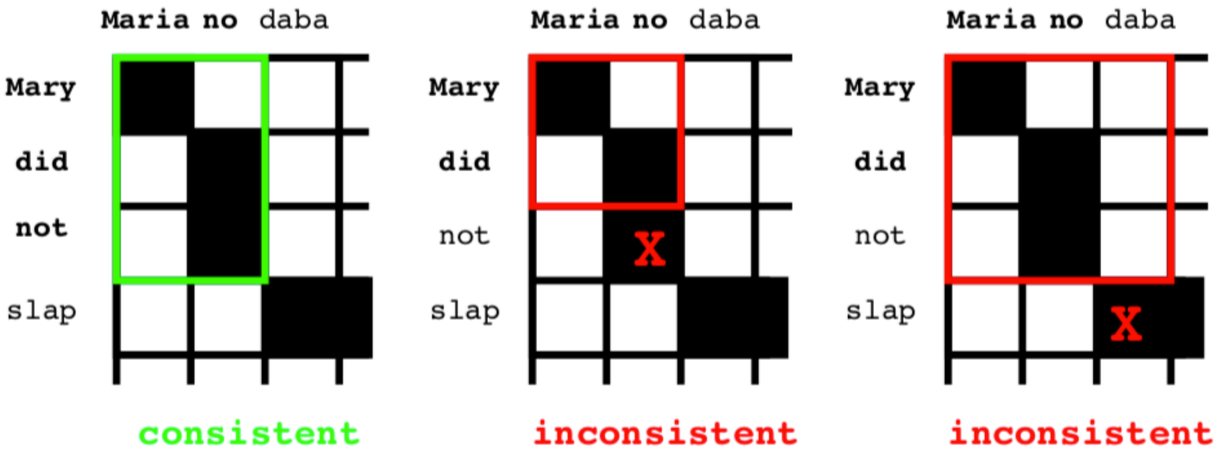
How do we learn phrase translations?

- Start with word alignments
- Preprocess using a heuristic - “Grow-Diag-Final” (Och & Ney, 2003)
- Collect **all** phrase pairs that are **consistent** with the word alignment



How do we learn phrase translations?

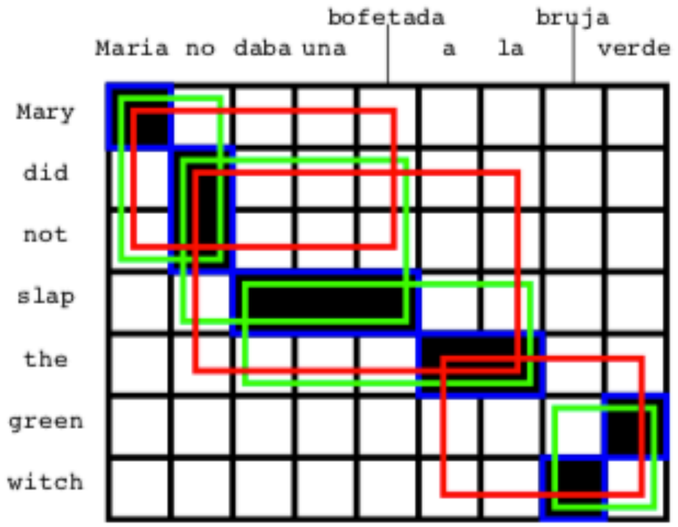
- Start with word alignments
- Preprocess using a heuristic - “Grow-Diag-Final” (Och & Ney, 2003)
- Collect **all** phrase pairs that are **consistent** with the word alignment
  - Phrase alignment has to contain **all** alignment points for **all** covered words



How do we learn phrase translations?

How do we learn phrase translations?

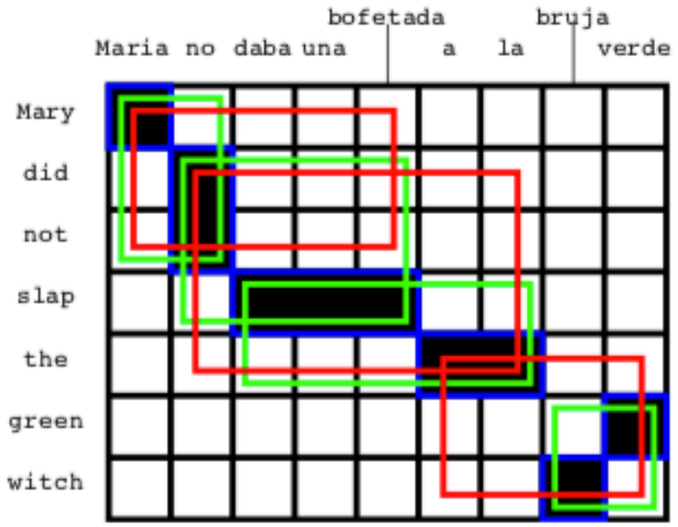
- We can learn phrase pairs with varying lengths:



(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)

How do we learn phrase translations?

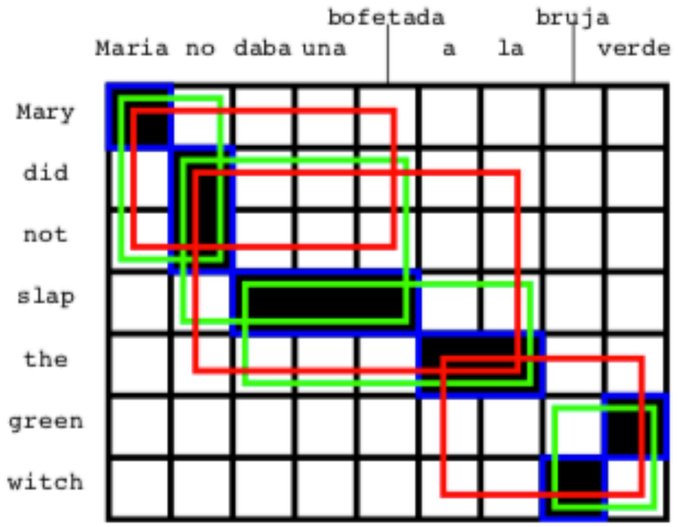
- We can learn phrase pairs with varying lengths:
- Too sparse with little data!



(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)

How do we learn phrase translations?

- We can learn phrase pairs with varying lengths:
- Too sparse with little data!
- Should be tuned



(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)

How do we translate?

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:

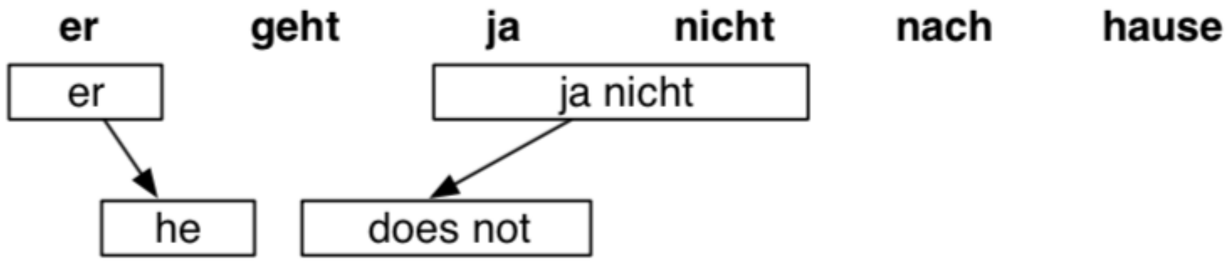
er geht ja nicht nach hause

How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:

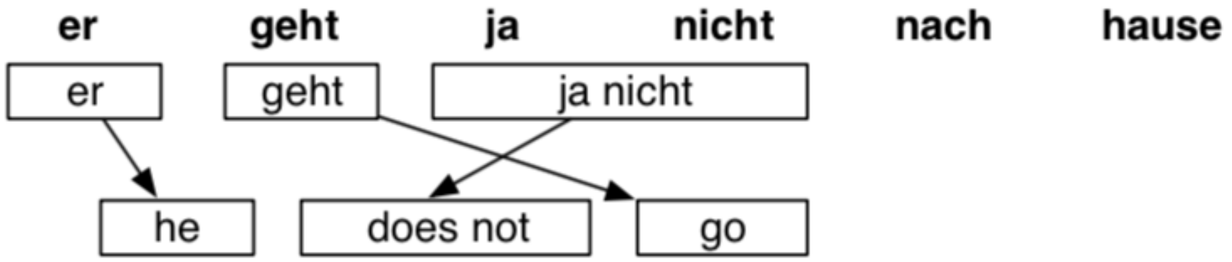
How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:



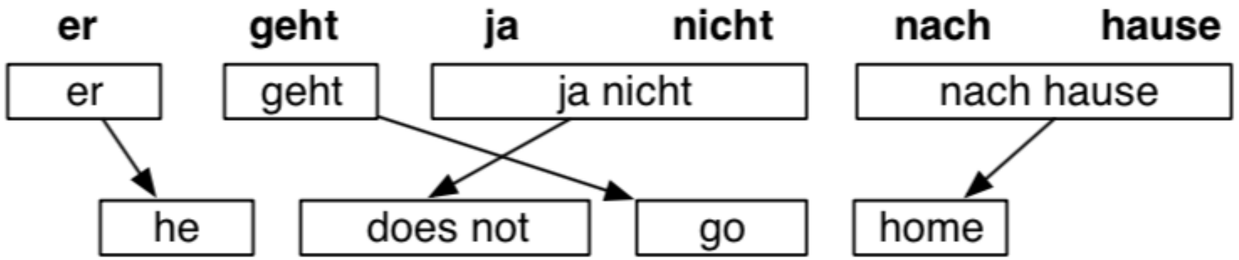
How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:



How do we translate?

- We want to solve the following problem: $\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\, p(\mathbf{e}|\mathbf{f})$
- This task is called **decoding**
- We need to **search** for the most probable translation
- For example:

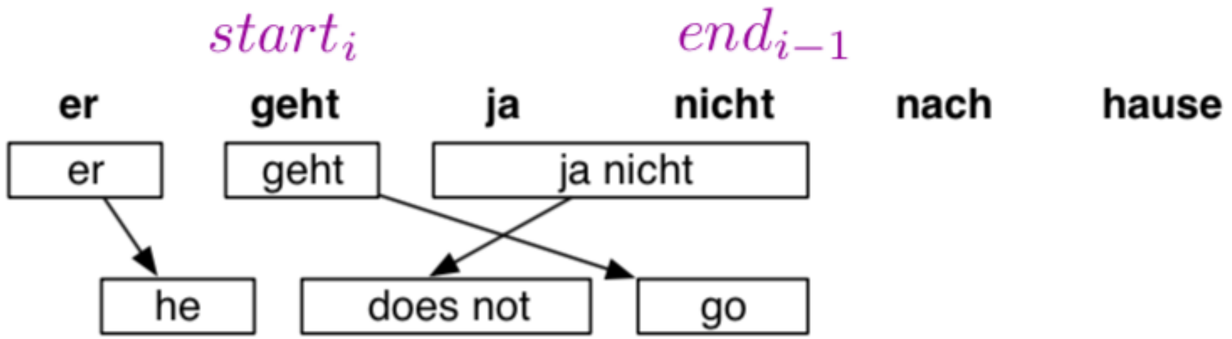


How do we translate?

How do we translate?

- In phrase-based translation, score **partial hypotheses** by**:**

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1)\, p_{\text{LM}}(\mathbf{e})$$

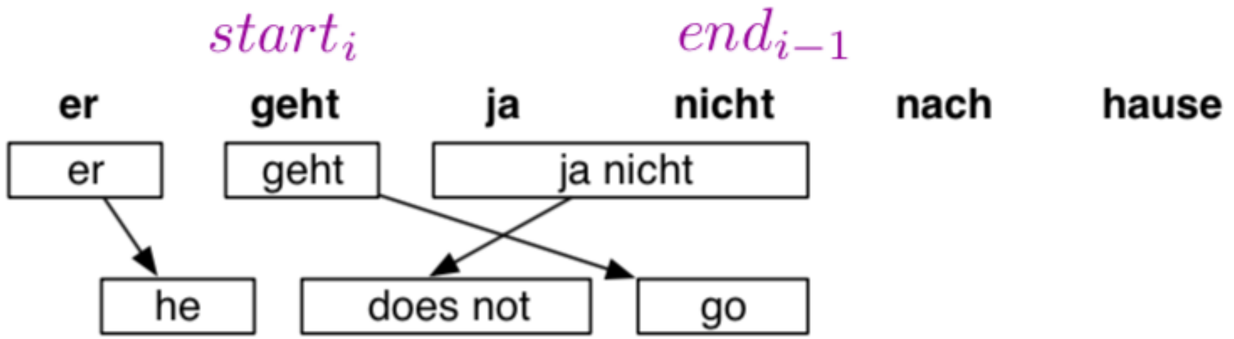


How do we translate?

- In phrase-based translation, score **partial hypotheses** by**:**

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1)\, p_{\text{LM}}(\mathbf{e})$$

- Translation model, reordering model, language model

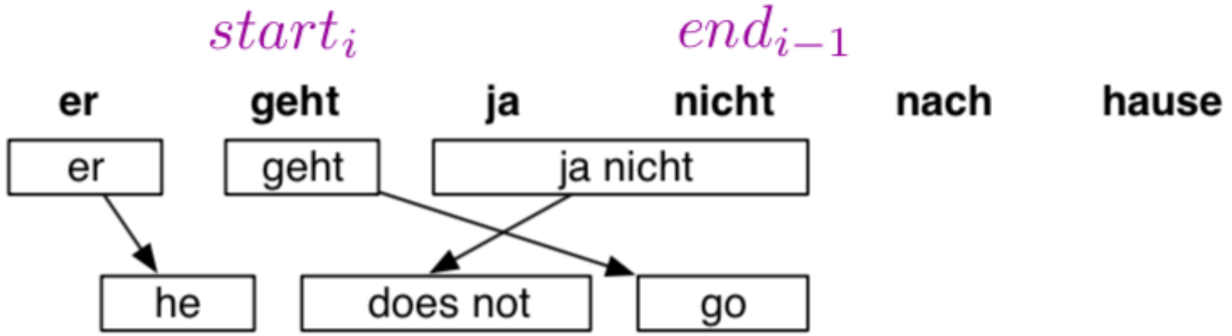


How do we translate?

- In phrase-based translation, score **partial hypotheses** by**:**

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\, d(start_i - end_{i-1} - 1)\, p_{\text{LM}}(\mathbf{e})$$

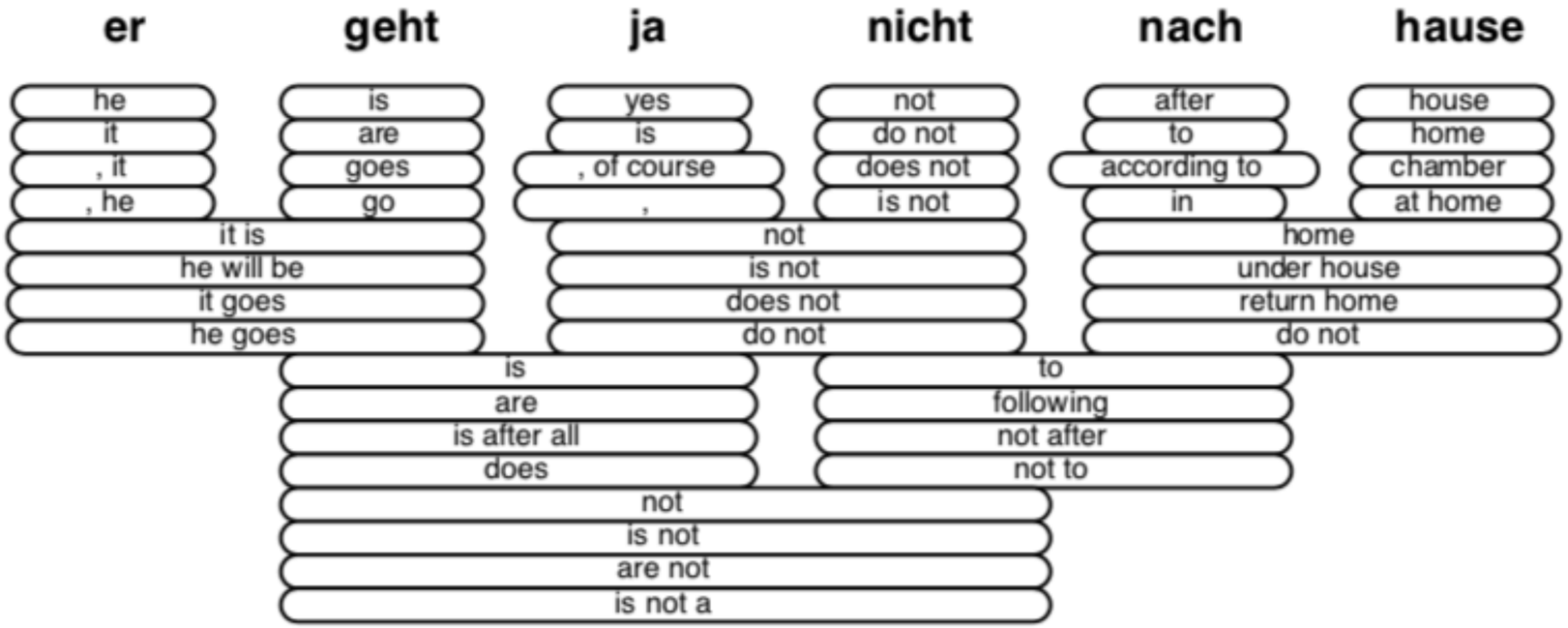
- Translation model, reordering model, language model
- A small problem…



How do we translate?

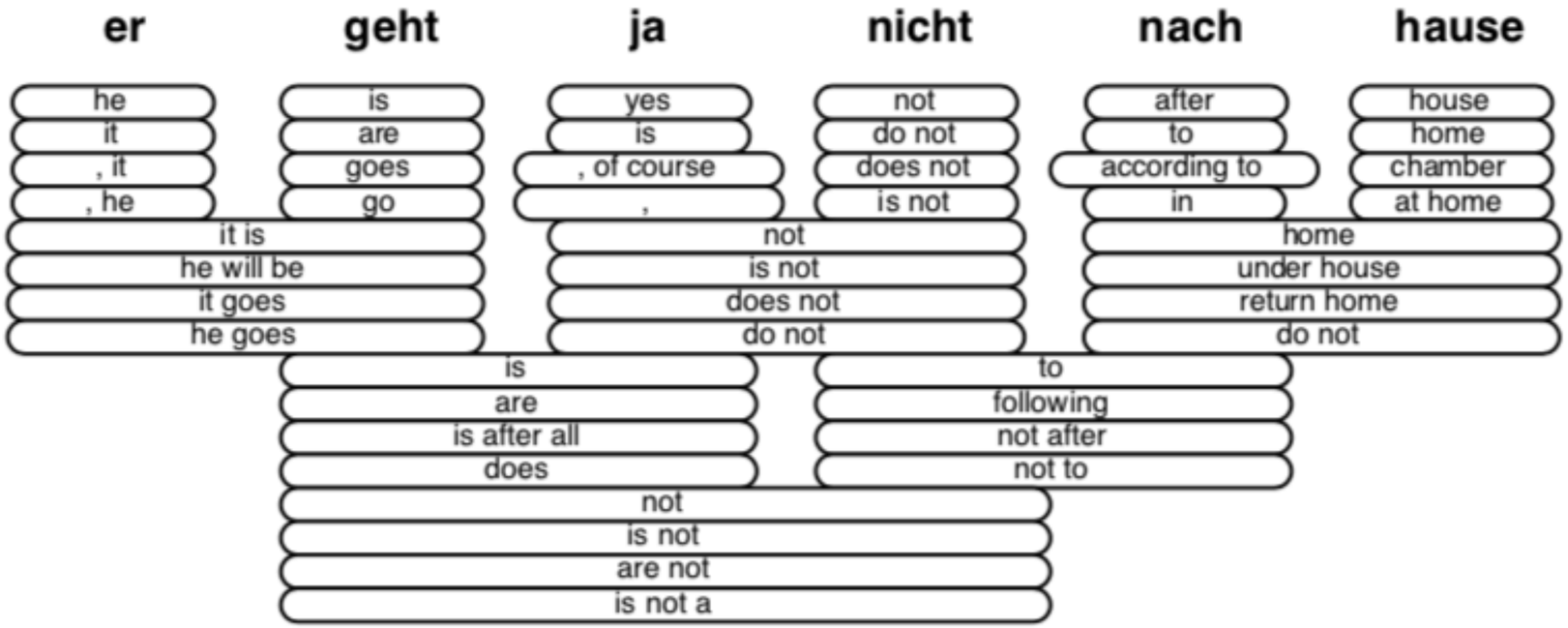
How do we translate?

- Large search space - many options to choose from



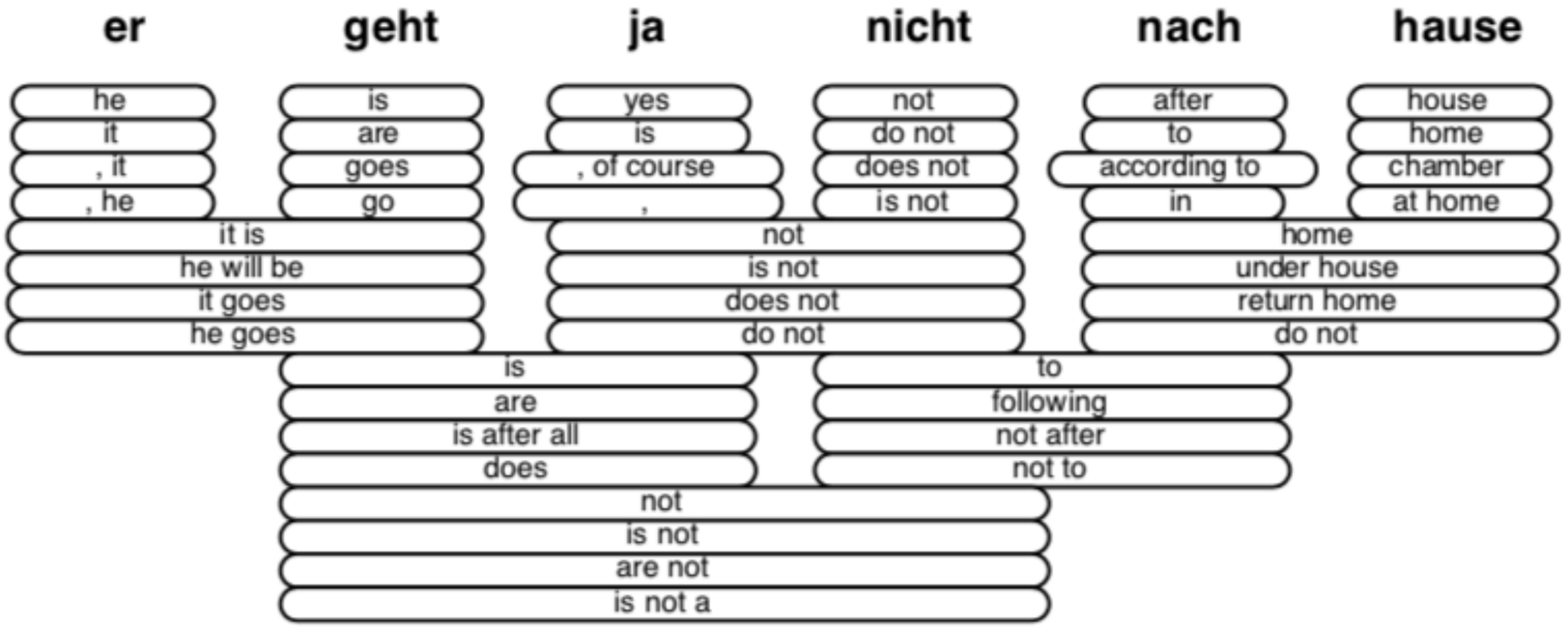
How do we translate?

- Large search space - many options to choose from
- Use words? Phrases? Larger phrases?



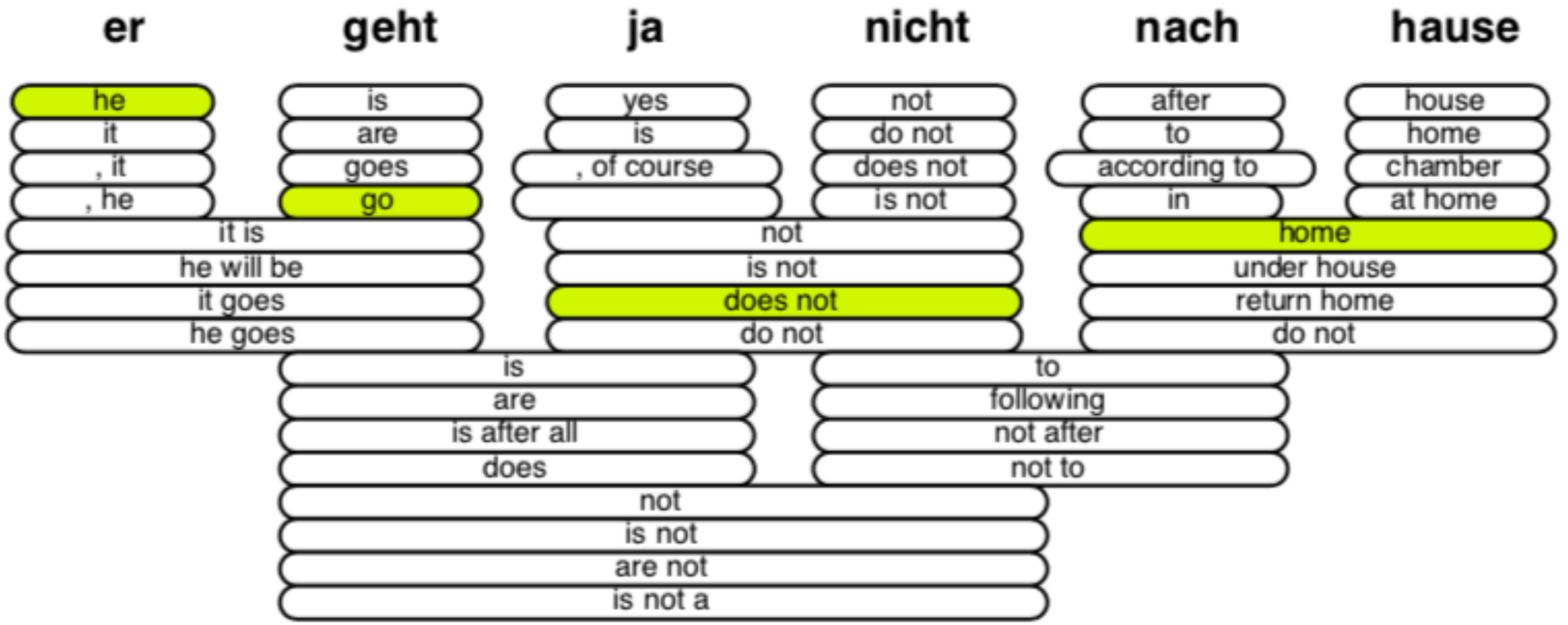
How do we translate?

- Large search space - many options to choose from
- Use words? Phrases? Larger phrases?
- 2727 relevant phrase pairs for this sentence in Europarl alone



How do we translate?

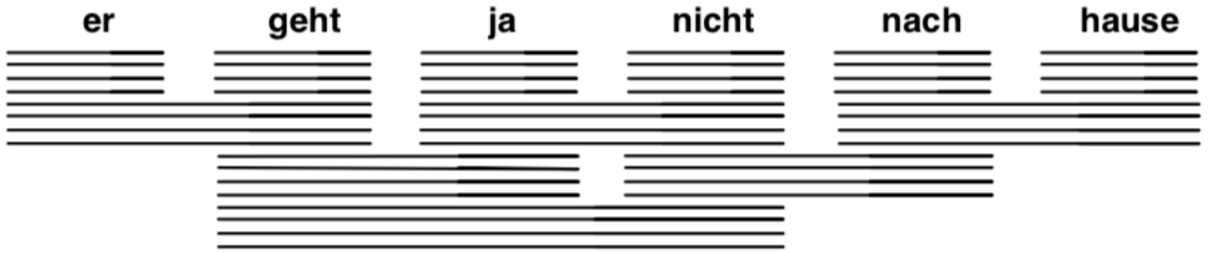
- Large search space - many options to choose from
- Use words? Phrases? Larger phrases?
- 2727 relevant phrase pairs for this sentence in Europarl alone
- Exponential explosion

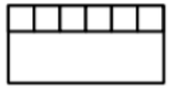


Search Procedure

Search Procedure

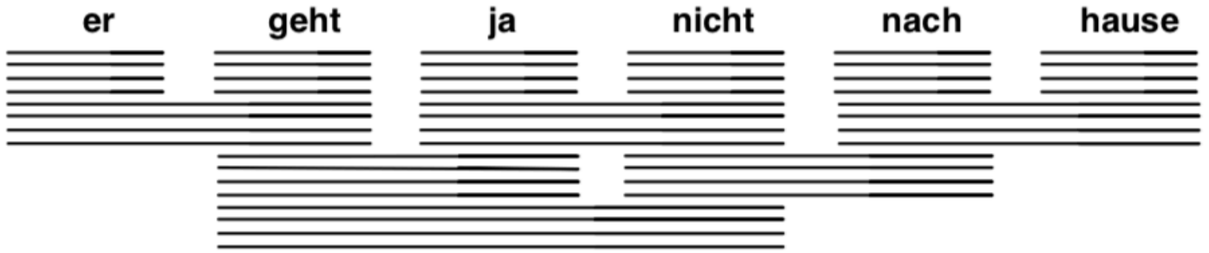
- Start with empty hypothesis

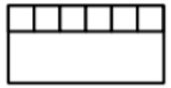




Search Procedure

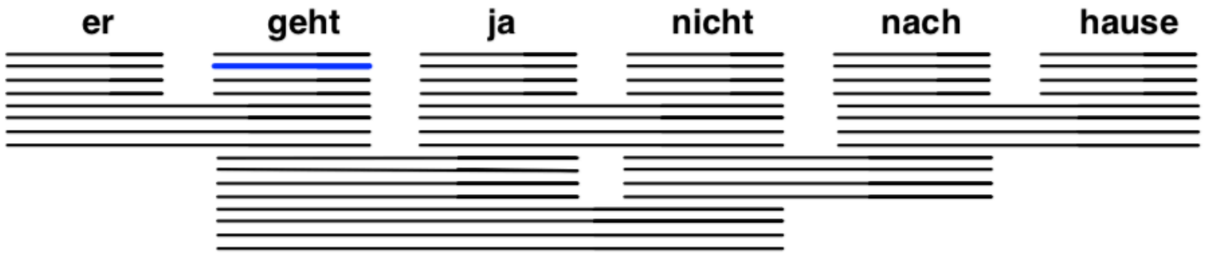
- Start with empty hypothesis
- Fetch all initial hypotheses and score them

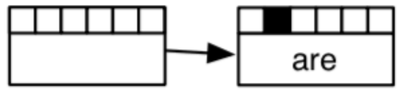




Search Procedure

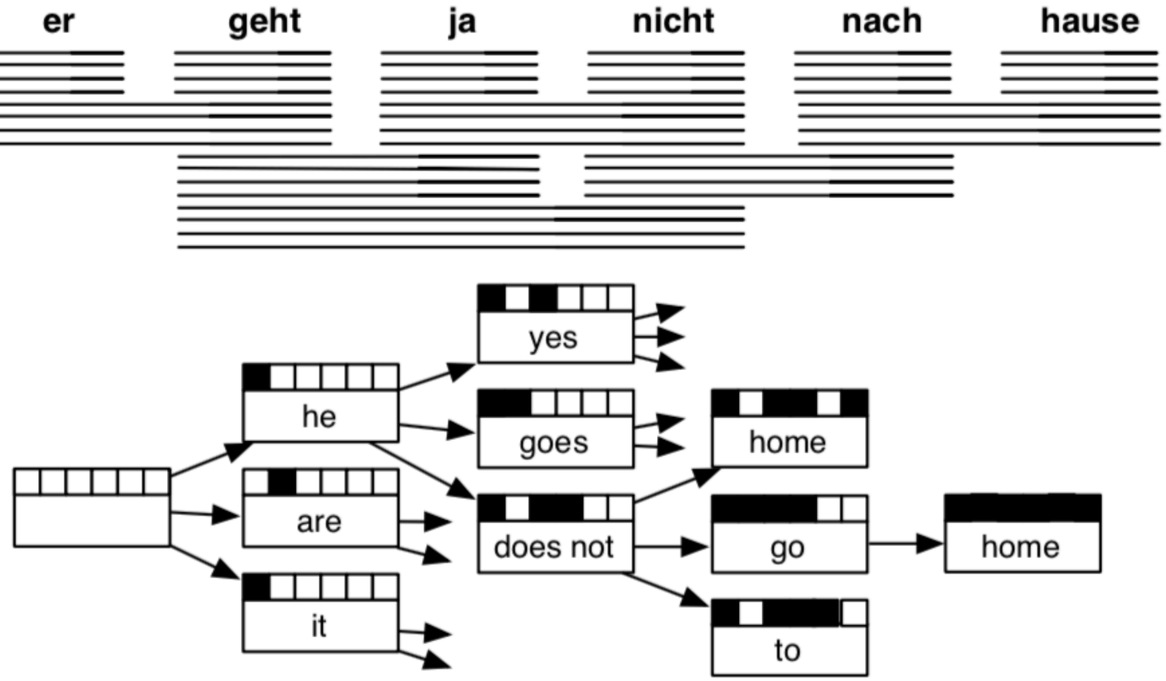
- Start with empty hypothesis
- Fetch all initial hypotheses and score them
- Maintain a coverage vector for each hypothesis





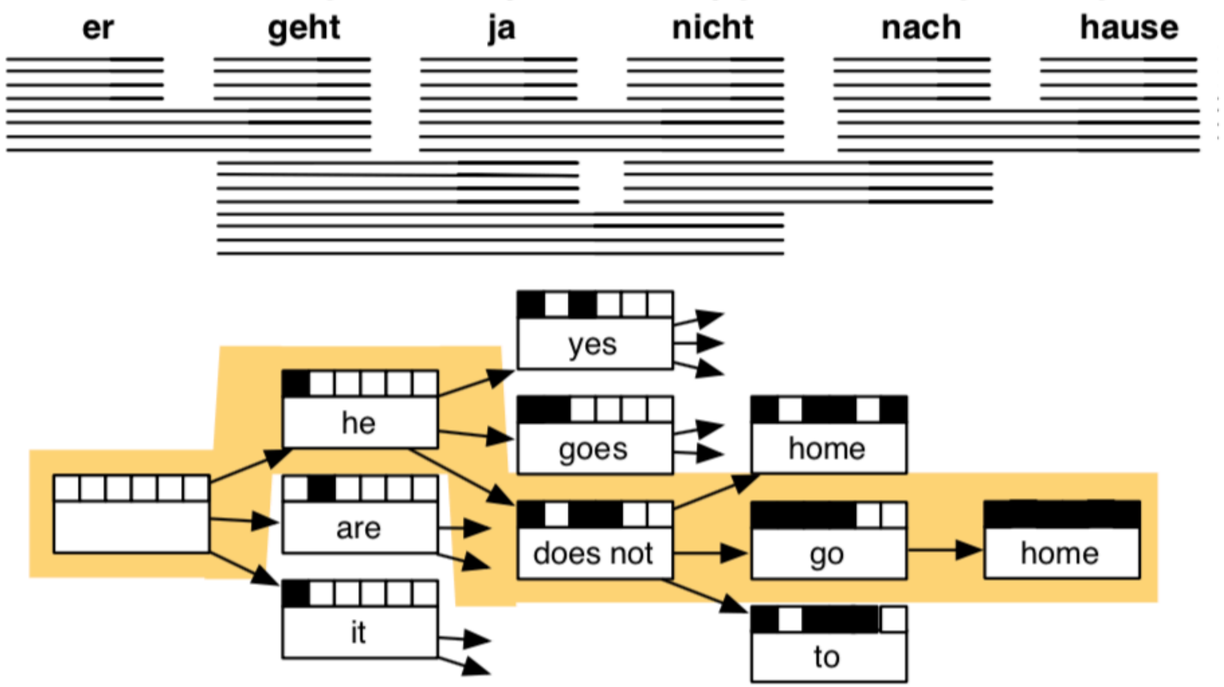
Search Procedure

- Start with empty hypothesis
- Fetch all initial hypotheses and score them
- Maintain a coverage vector for each hypothesis
- Iterate: expand each hypothesis until full cover



Search Procedure

- Start with empty hypothesis
- Fetch all initial hypotheses and score them
- Maintain a coverage vector for each hypothesis
- Iterate: expand each hypothesis until full cover
- Backtrack to get best candidate



Search Procedure

Search Procedure

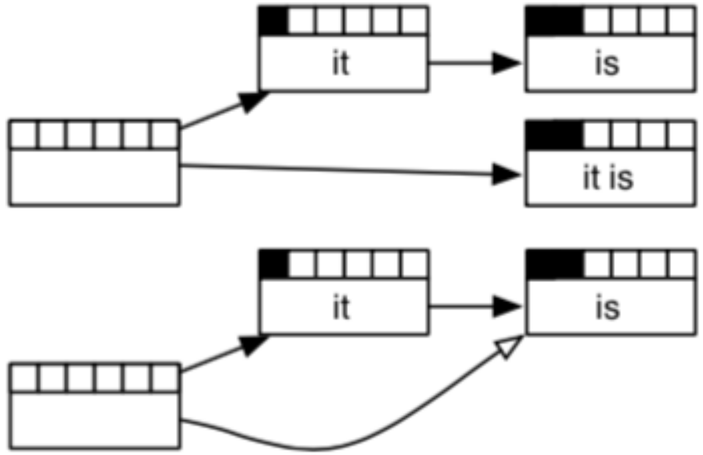
- The proposed approach is still exponential.

Search Procedure

- The proposed approach is still exponential.
- How can we improve?

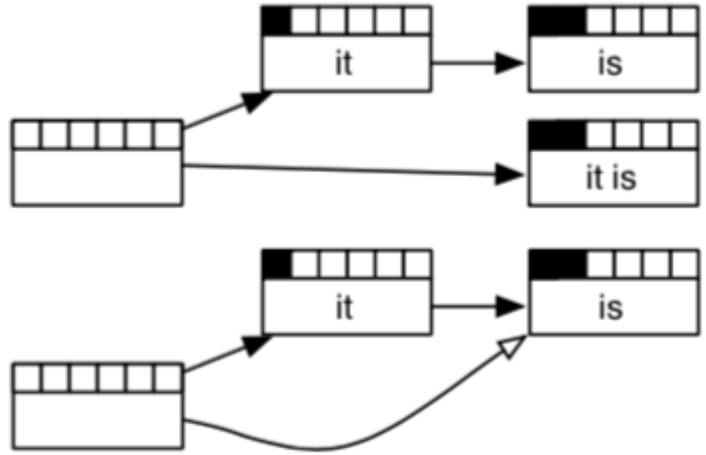
Search Procedure

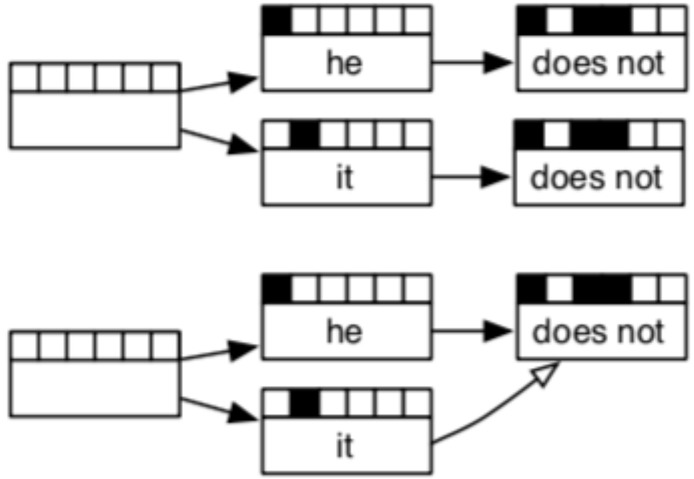
- The proposed approach is still exponential.
- How can we improve?
  - Recombination - merge hypotheses with similar traits



Search Procedure

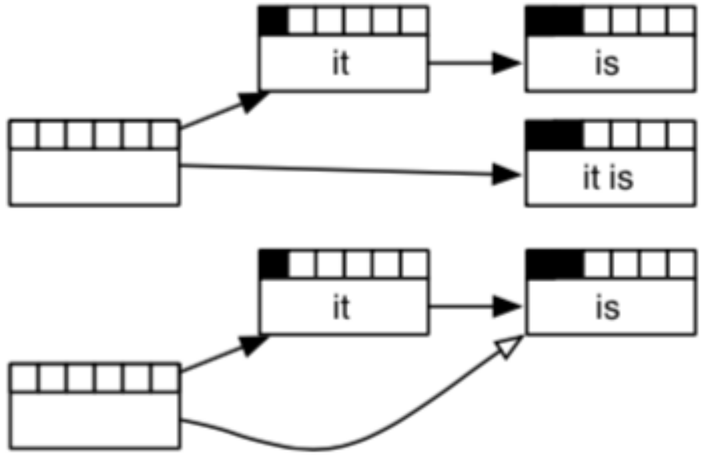
- The proposed approach is still exponential.
- How can we improve?
  - Recombination - merge hypotheses with similar traits
  - Depends on the modelled context

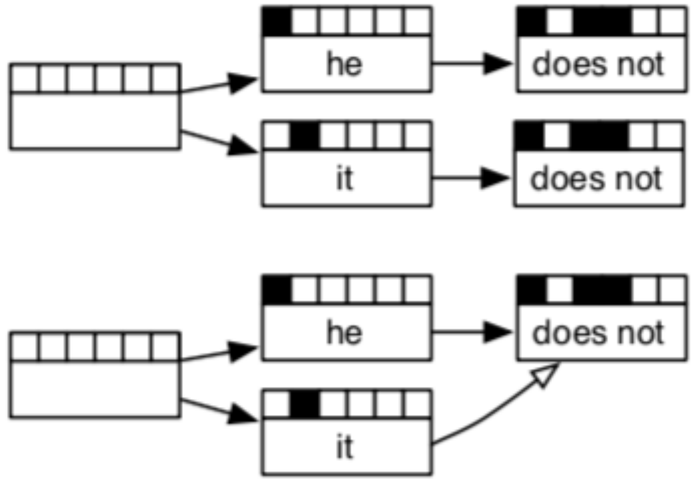




Search Procedure

- The proposed approach is still exponential.
- How can we improve?
  - Recombination - merge hypotheses with similar traits
  - Depends on the modelled context
  - Still not enough, not controllable





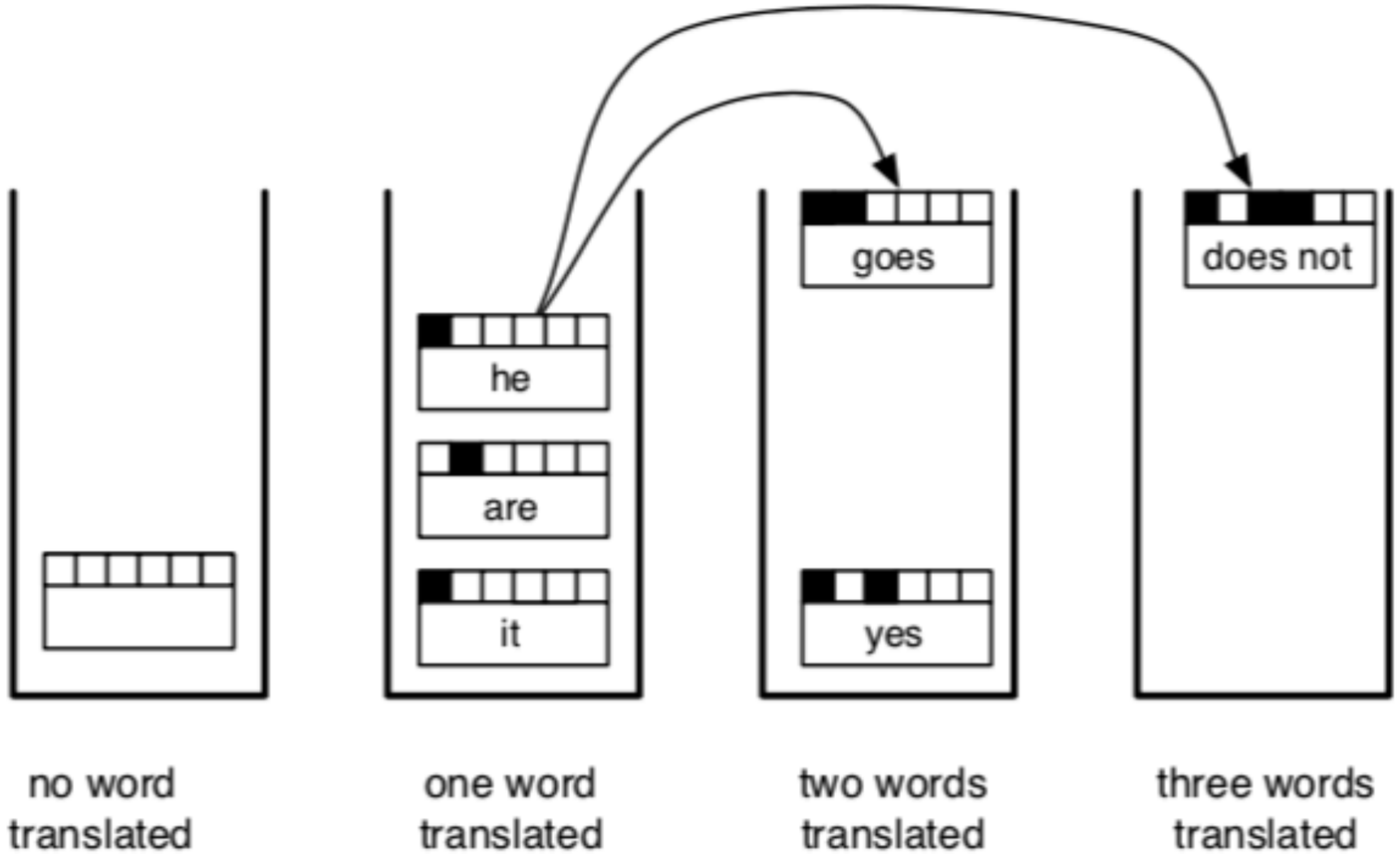
Pruning with Stack Decoding

Pruning with Stack Decoding

- Idea: throw away bad hypothesis early

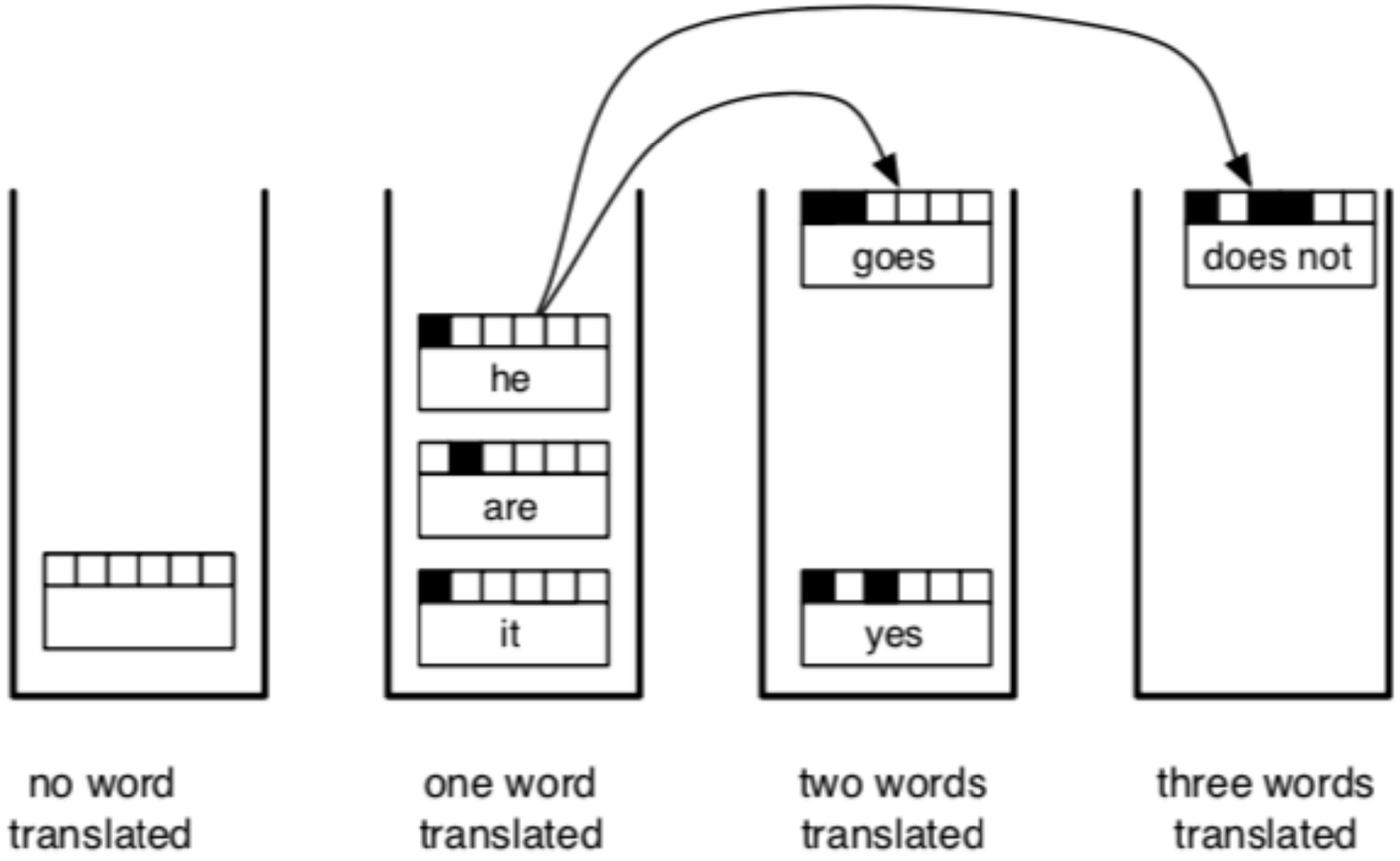
Pruning with Stack Decoding

- Idea: throw away bad hypothesis early
- Put comparable hypotheses into stacks (same coverage)



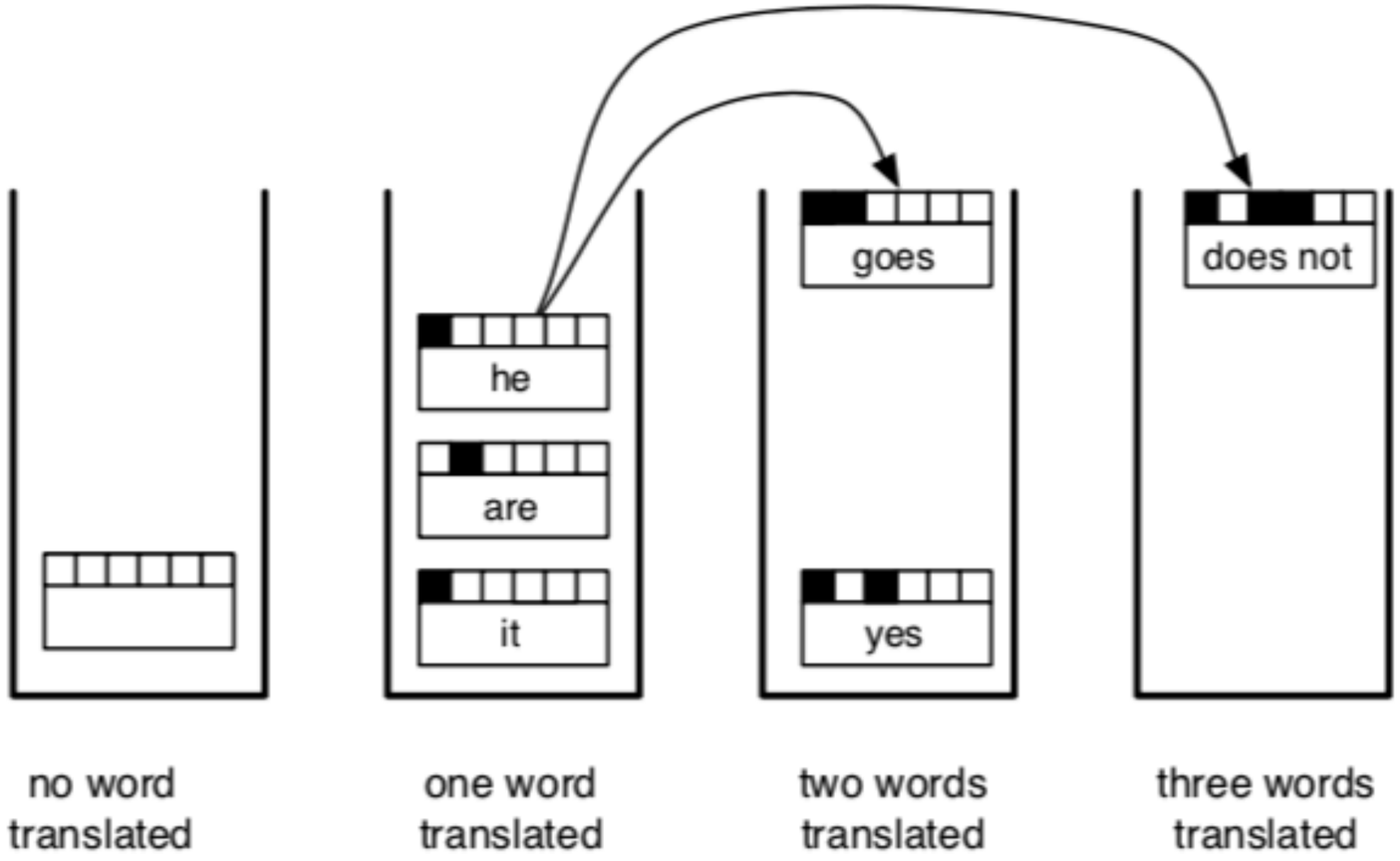
Pruning with Stack Decoding

- Idea: throw away bad hypothesis early
- Put comparable hypotheses into stacks (same coverage)
- Limit the number of hypotheses in each stack - throw away the worst hypothesis when stack is full



Pruning with Stack Decoding

- Idea: throw away bad hypothesis early
- Put comparable hypotheses into stacks (same coverage)
- Limit the number of hypotheses in each stack - throw away the worst hypothesis when stack is full
- Beam search is similar, but without stacks



Pruning with Stack Decoding

Pruning with Stack Decoding

Stack Decoding Algorithm

```
1: place empty hypothesis into stack 0
2: for all stacks 0...n − 1 do
3:    for all hypotheses in stack do
4:       for all translation options do
5:          if applicable then
6:             create new hypothesis
7:             place in stack
8:             recombine with existing hypothesis if possible
9:             prune stack if too big
10:         end if
11:      end for
12:   end for
13: end for
```

Pruning with Stack Decoding

Stack Decoding Algorithm

```
1: place empty hypothesis into stack 0
2: for all stacks 0...n − 1 do
3:    for all hypotheses in stack do
4:       for all translation options do
5:          if applicable then
6:             create new hypothesis
7:             place in stack
8:             recombine with existing hypothesis if possible
9:             prune stack if too big
10:         end if
11:      end for
12:   end for
13: end for
```

$O(\text{max stack size} \times \text{translation options} \times \text{sentence length})$

Summary

Summary

- Additional lexical models

Summary

- Additional lexical models
  - Position based (absolute - Model 2, relative - Model 4)

Summary

- Additional lexical models
  - Position based (absolute - Model 2, relative - Model 4)
  - Fertility (Model 3)

Summary

- Additional lexical models
  - Position based (absolute - Model 2, relative - Model 4)
  - Fertility (Model 3)
- Phrase based models

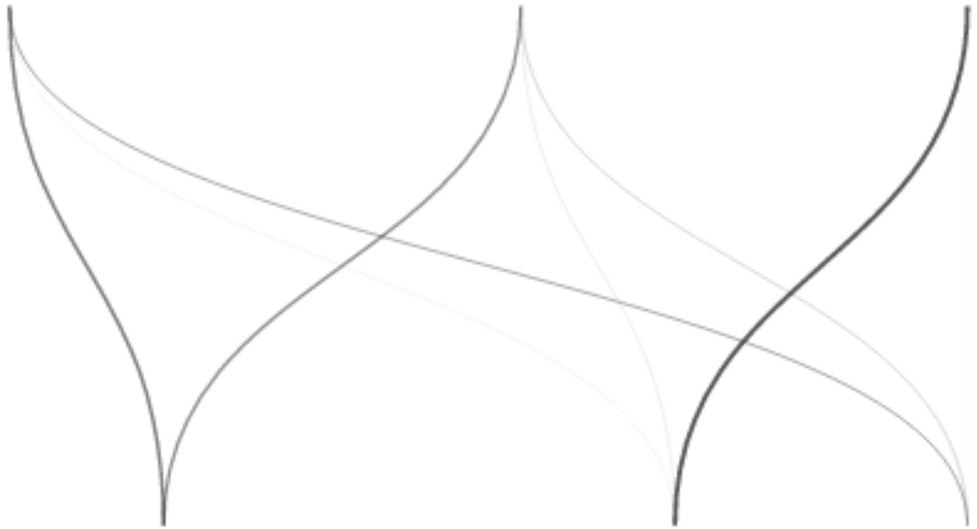
Summary

- Additional lexical models
  - Position based (absolute - Model 2, relative - Model 4)
  - Fertility (Model 3)
- Phrase based models
- Decoding

Summary

- Additional lexical models
  - Position based (absolute - Model 2, relative - Model 4)
  - Fertility (Model 3)
- Phrase based models
- Decoding
  - Recombination, Pruning (Stack decoding/Beam search)

Any Questions ?



Questions diverses ?