Phrase-Based MT

Machine Translation
Lecture 8

Instructor: Chris Callison-Burch
TAs: Mitchell Stern, Justin Chiu

Website: mt-class.org/penn

Translation Process

- Task: translate this sentence from German into English

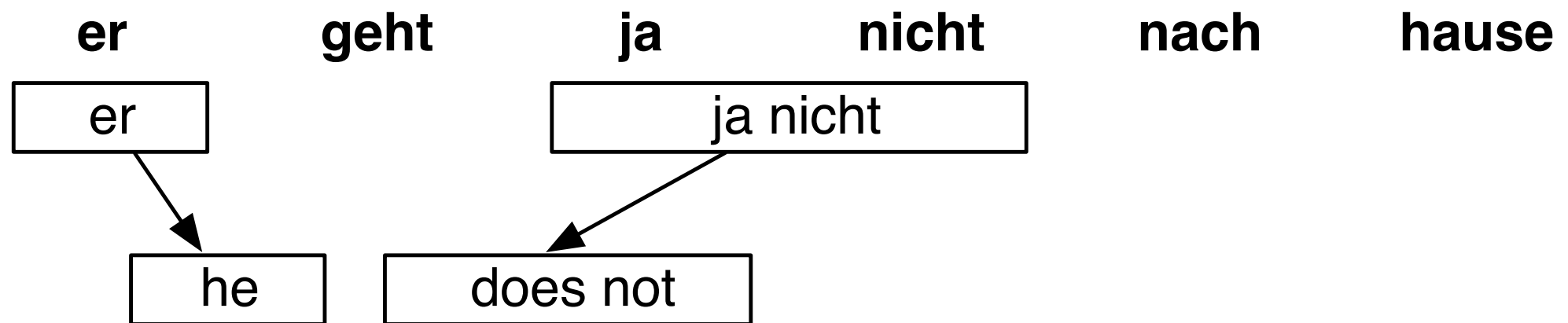**er   geht   ja   nicht   nach   hause**

Translation Process

- Task: translate this sentence from German into English

er   **geht**   **ja**   **nicht**   **nach**   **hause**

er

he

- Pick phrase in input, translate

Translation Process

- Task: translate this sentence from German into English

| er | geht | ja | nicht | nach | hause |

```
┌──────────┐                ┌──────────────────────┐
│    er    │                │      ja nicht        │
└──────────┘                └──────────────────────┘
      ↓                              ↓
   ┌──────┐              ┌──────────────────┐
   │  he  │              │     does not     │
   └──────┘              └──────────────────┘
```

- Pick phrase in input, translate

  - it is allowed to pick words out of sequence reordering
  - phrases may have multiple words: many-to-many translation

Translation Process

- Task: translate this sentence from German into English

er   **geht**   **ja**   **nicht**   **nach**   **hause**

er	geht	ja nicht

he	does not	go

- Pick phrase in input, translate

Translation Process

- Task: translate this sentence from German into English

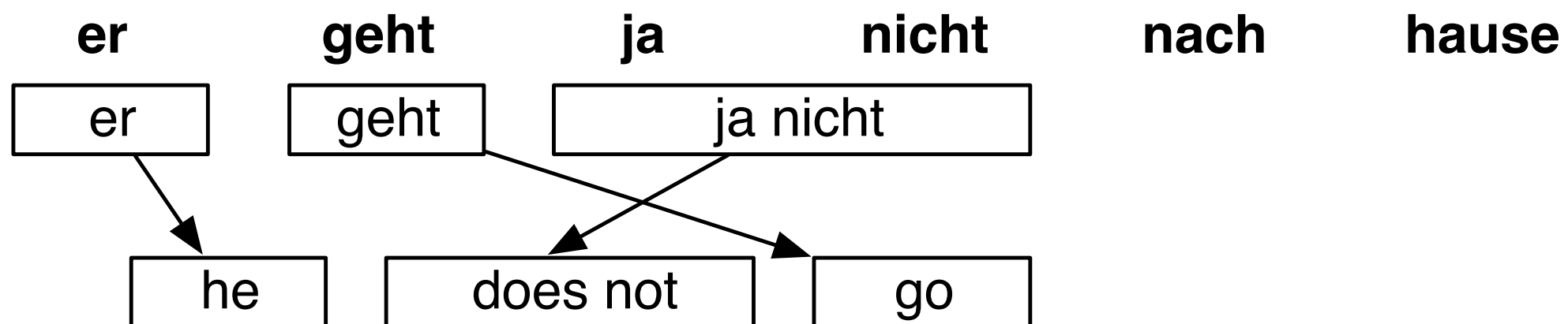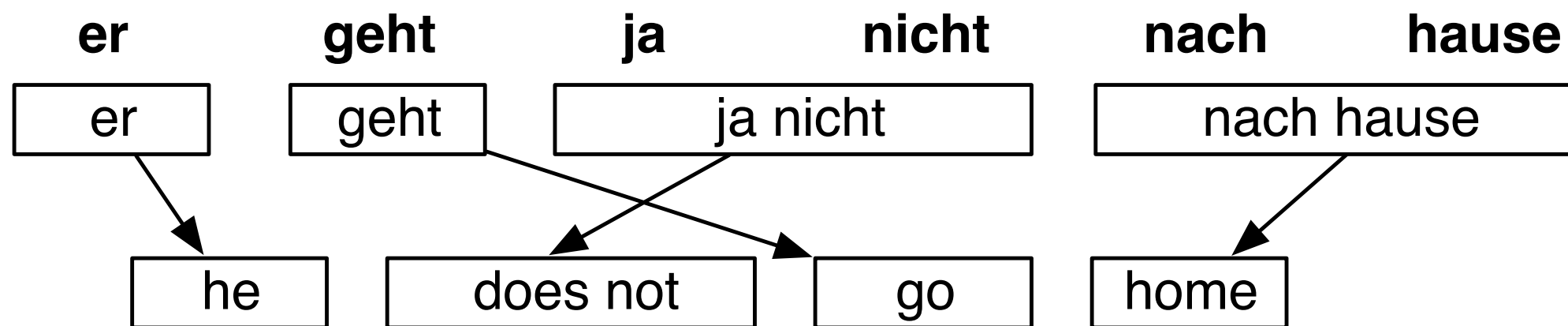- Pick phrase in input, translate

Computing Translation Probability

- Probabilistic model for phrase-based translation:

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

- Score is computed incrementally for each partial hypothesis

- Components

  Phrase translation Picking phrase $\bar{f}_i$ to be translated as a phrase $\bar{e}_i$
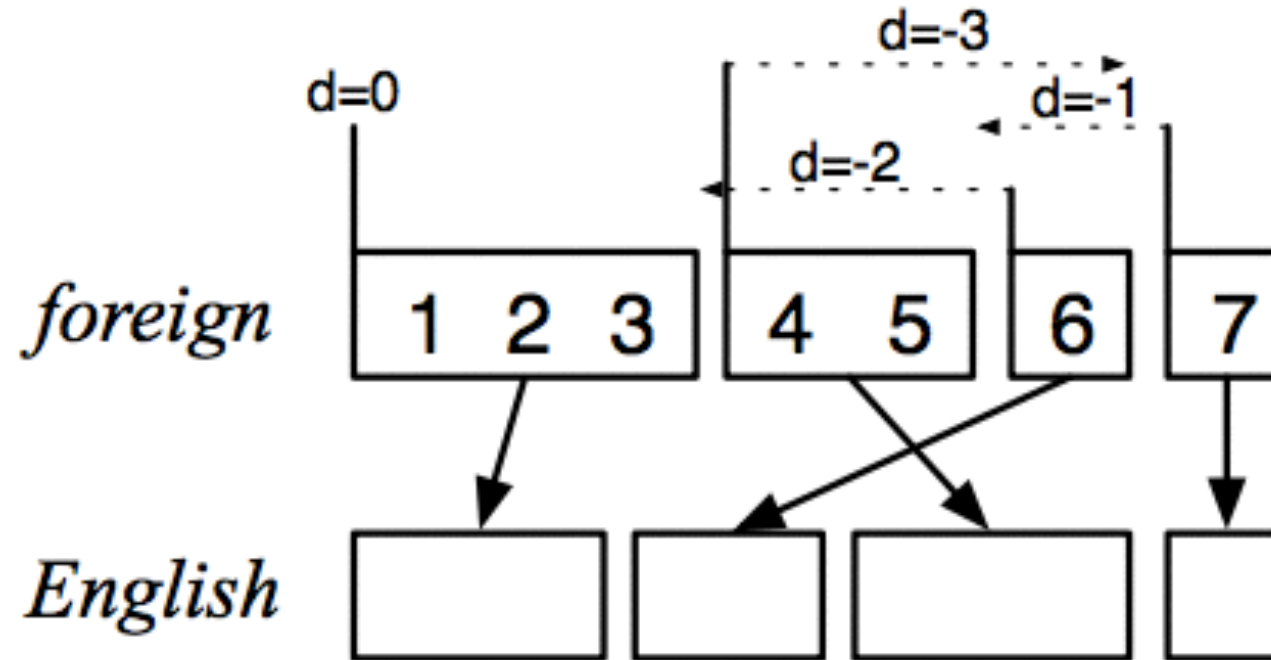 $\rightarrow$ look up score $\phi(\bar{f}_i|\bar{e}_i)$ from phrase translation table
  Reordering Previous phrase ended in end_{i-1}, current phrase starts at $start_i$
    $\rightarrow$ compute $d(start_i - end_{i-1} - 1)$
  Language model For n-gram model, need to keep track of last $n-1$ words
    $\rightarrow$ compute score $p_{\text{LM}}(w_i|w_{i-(n-1)}, ..., w_{i-1})$ for added words w_i

Reordering Model

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

Translation Options

er	**geht**	**ja**	**nicht**	**nach**	**hause**
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home

er geht	ja nicht	nach hause
it is	not	home
he will be	is not	under house
it goes	does not	return home
he goes	do not	do not

geht ja	nicht nach
is	to
are	following
is after all	not after
does	not to

geht ja nicht
not
is not
are not
is not a

- Many translation options to choose from

  – in Europarl phrase table: 2727 matching phrase pairs for this sentence
  – by pruning to the top 20 per phrase, 202 translation options remain

---

Translation Options

er	**geht**	**ja**	**nicht**	**nach**	**hause**
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home

it is		not		home	
he will be		is not		under house	
it goes		does not		return home	
he goes		do not		do not	

is		to
are		following
is after all		not after
does		not to

| not |
|-----|
| is not |
| are not |
| is not a |

- Many translation options to choose from

  – in Europarl phrase table: 2727 matching phrase pairs for this sentence
  – by pruning to the top 20 per phrase, 202 translation options remain

Translation Options

er	geht	ja	nicht	nach	hause
he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	go	,	is not	in	at home

it is	not	home
he will be	is not	under house
it goes	does not	return home
he goes	do not	do not

is	to
are	following
is after all	not after
does	not to

not
is not
are not
is not a

- Many translation options to choose from
  - in Europarl phrase table: 2727 matching phrase pairs for this sentence
  - by pruning to the top 20 per phrase, 202 translation options remain

Translation Options

er	geht	ja	nicht	nach	hause

he	is	yes	not	after	house
it	are	is	do not	to	home
, it	goes	, of course	does not	according to	chamber
, he	**go**		is not	in	at home

it is		not		**home**	
he will be		is not		under house	
it goes		**does not**		return home	
he goes		do not		do not	

is		to	
are		following	
is after all		not after	
does		not to	

| not |
| is not |
| are not |
| is not a |

- The machine translation decoder does not know the right answer
  - picking the right translation options
  - arranging them in the right order

$\rightarrow$ Search problem solved by heuristic beam search

Decoding: Precompute Translation Options
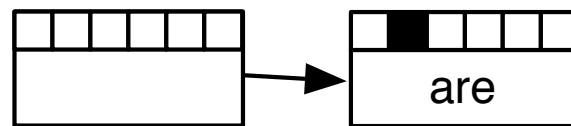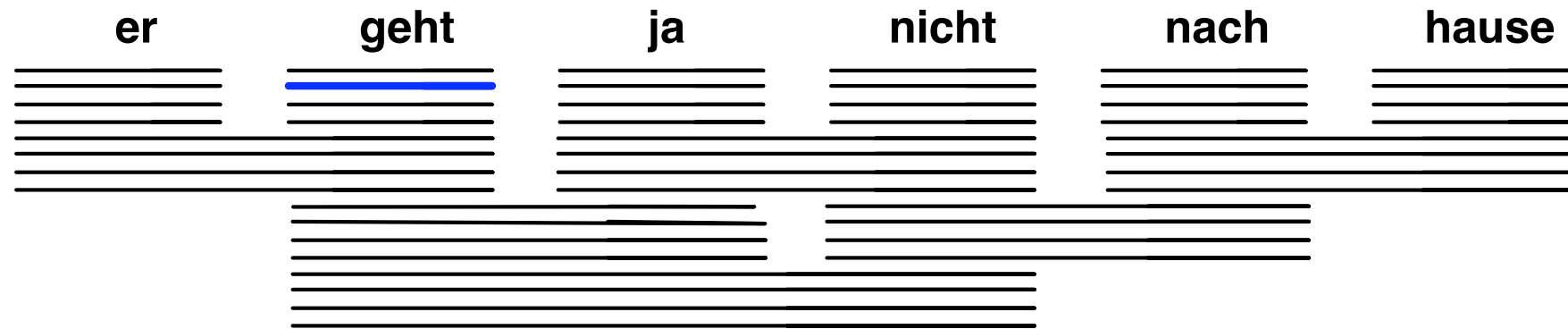

er　　　　geht　　　　ja　　　　nicht　　　　nach　　　　hause
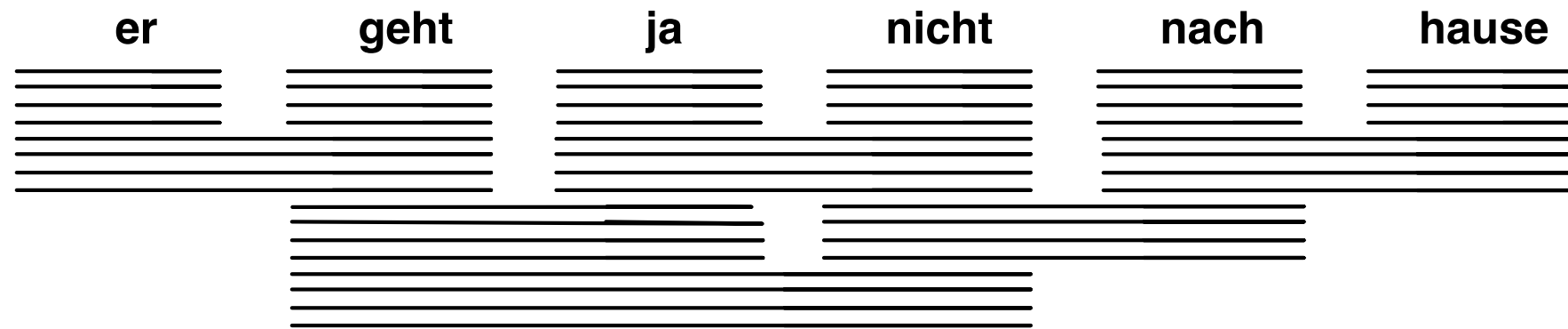
consult phrase translation table for all input phrases

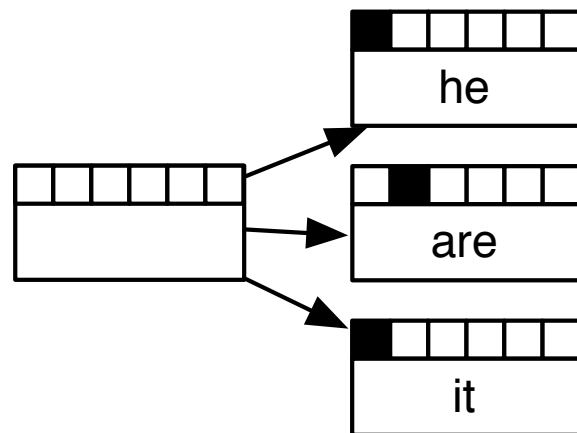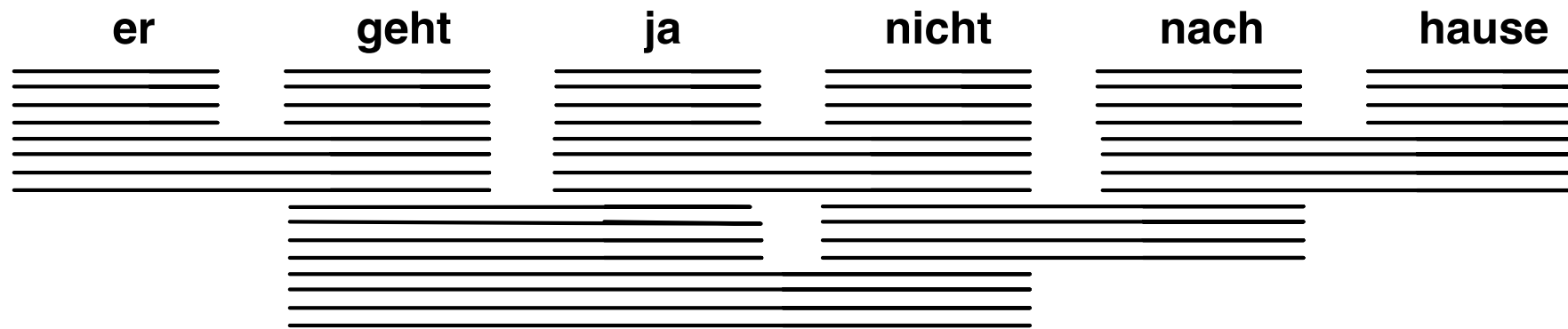# Decoding: Start with Initial Hypothesis
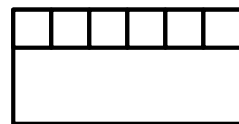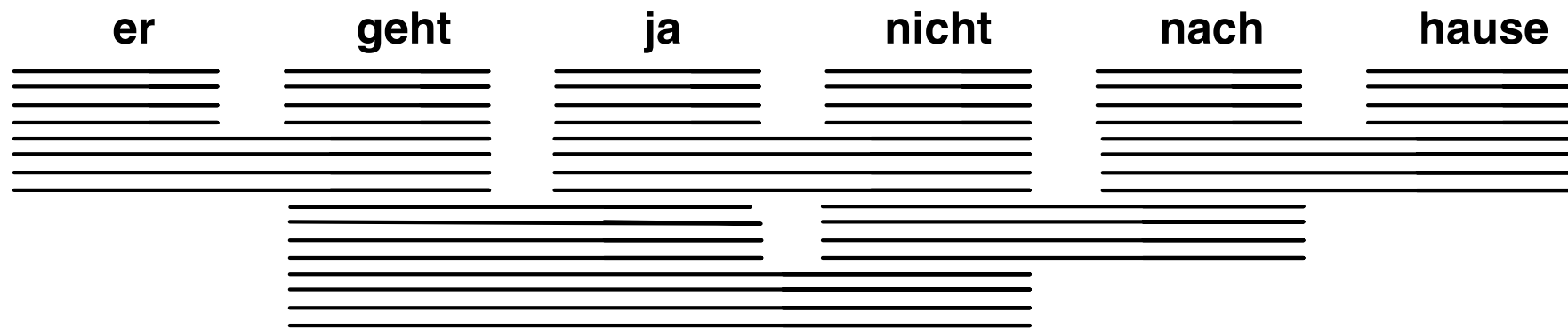
initial hypothesis: no input words covered, no output produced

Decoding: Hypothesis Expansion

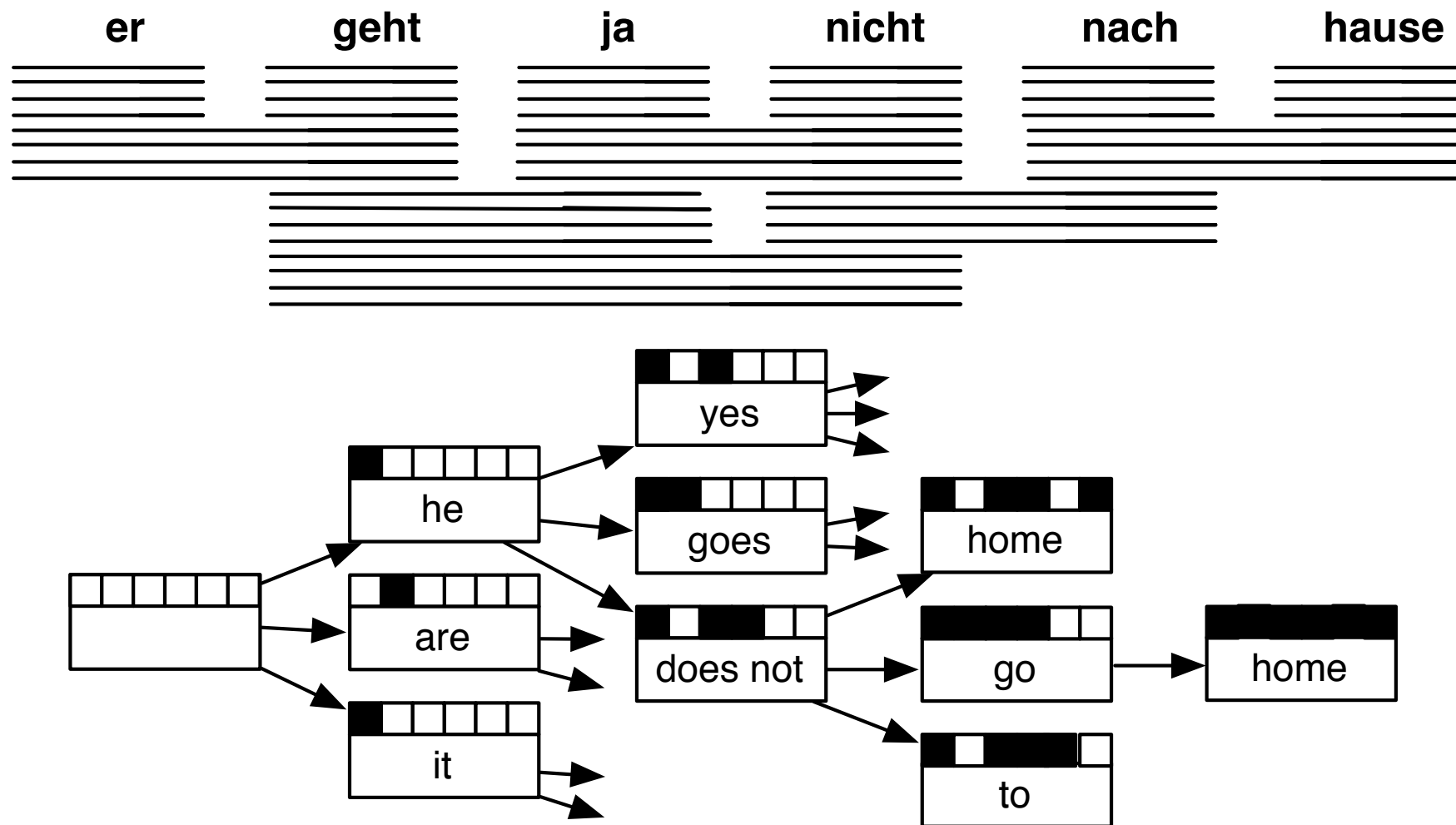pick any translation option, create new hypothesis

Decoding: Hypothesis Expansion

er   **geht**   **ja**   **nicht**   **nach**   **hause**

create hypotheses for all other translation options
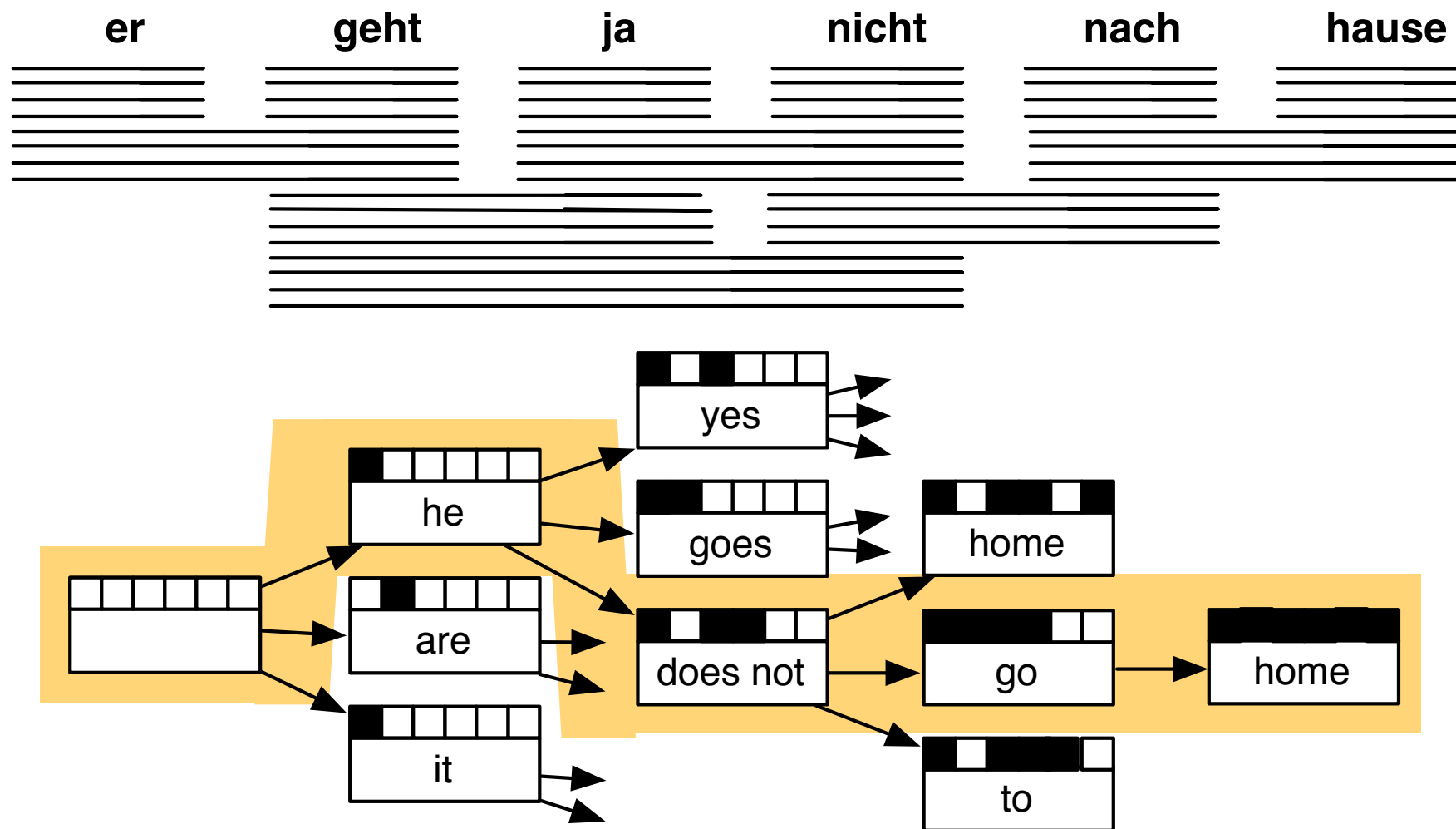
Decoding: Hypothesis Expansion

also create hypotheses from created partial hypothesis

Decoding: Find Best Path

backtrack from highest scoring complete hypothesis

Size of the search space

- Decoding is an NP-complete problem

- The search space is exponential

- How can we reduce the search?

- Dynamic programming (risk free)

- Heuristic search (risky)

Recombination

- Two hypothesis paths lead to two matching hypotheses

  - same number of foreign words translated
  - same English words in the output
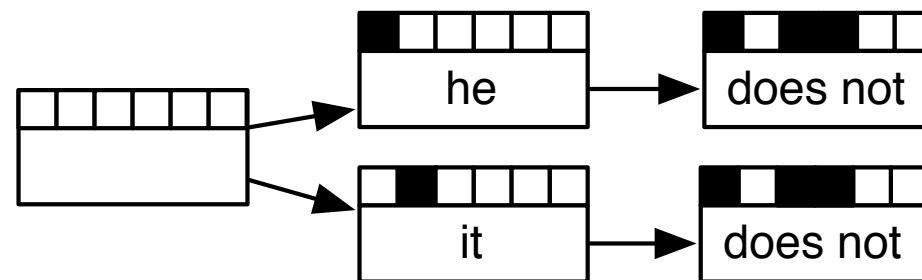 - different scores
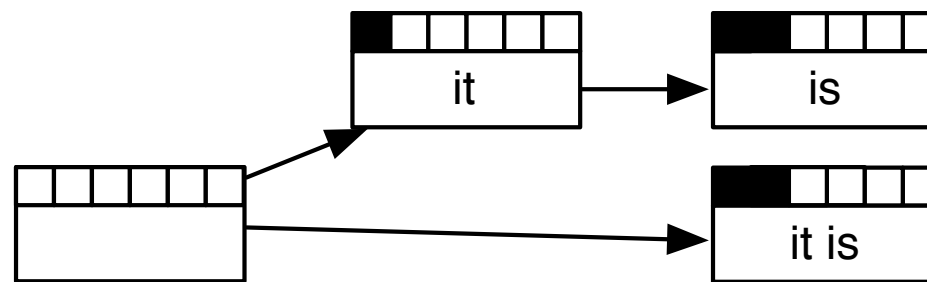
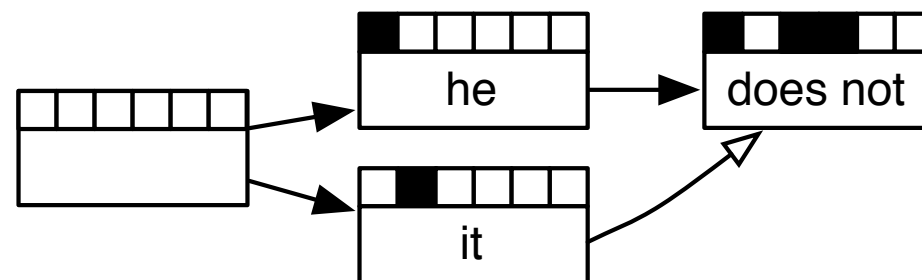- Worse hypothesis is dropped
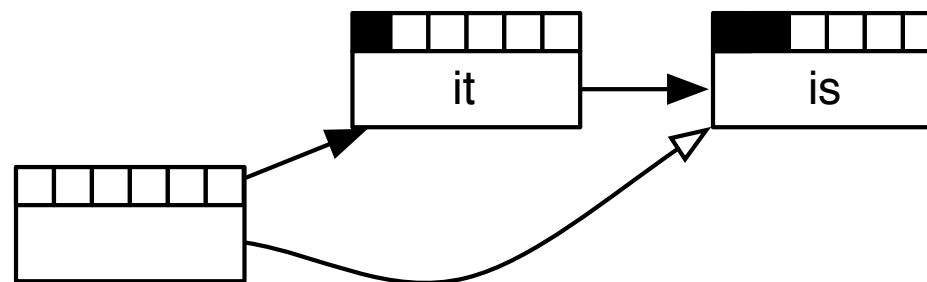
Recombination

- Two hypothesis paths lead to hypotheses indistinguishable in subsequent search

  - same number of foreign words translated
  - same last two English words in output (assuming trigram language model)
  - same last foreign word translated
  - different scores

- Worse hypothesis is dropped

Restrictions on Recombination

- **Translation model:** Phrase translation independent from each other

  $\rightarrow$ no restriction to hypothesis recombination

- **Language model:** Last $n-1$ words used as history in n-gram language model

  $\rightarrow$ recombined hypotheses must match in their last $n-1$ words

- **Reordering model:** Distance-based reordering model based on distance to end position of previous input phrase

  $\rightarrow$ recombined hypotheses must have that same end position

- Other feature function may introduce additional restrictions

Pruning

- Recombination reduces search space, but not enough
  (we still have a NP complete problem on our hands)

- Pruning: remove bad hypotheses early

  - put comparable hypothesis into stacks
    (hypotheses that have translated same number of input words)
  - limit number of hypotheses in each stack

Stacks

| no word
translated | one word
translated | two words
translated | three words
translated |
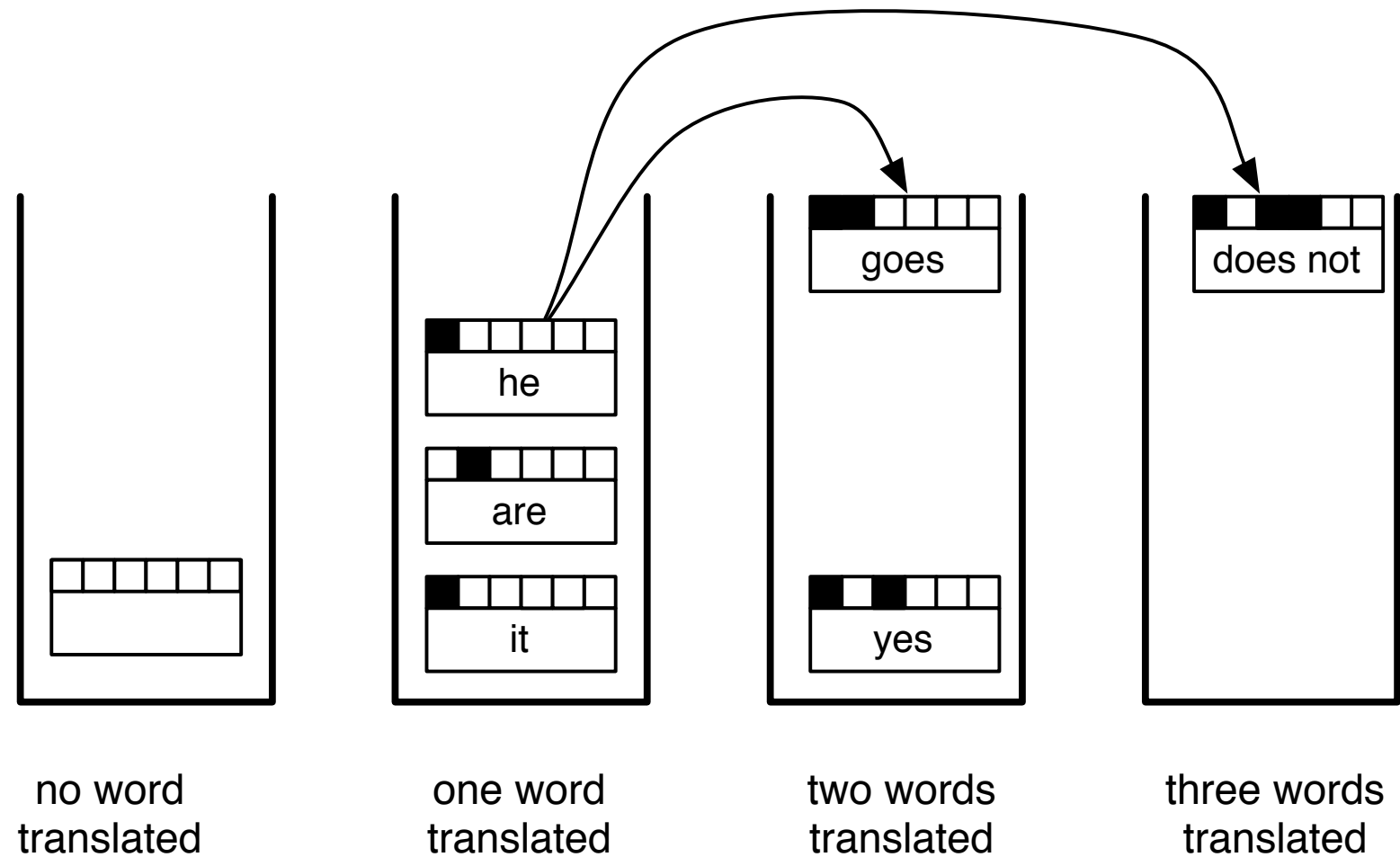
- Hypothesis expansion in a stack decoder
  - translation option is applied to hypothesis
  - new hypothesis is dropped into a stack further down

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**

Decoding complexity

- Finding the best hypothesis is NP-hard

  - Even with no language model, there are an exponential number of states!

  - Solution 1: limit reordering

  - Solution 2: (lossy) pruning

Pruning

- Pruning strategies

  - histogram pruning: keep at most k hypotheses in each stack
  - stack pruning: keep hypothesis with score $\alpha \times$ best score $(\alpha < 1)$

- Computational time complexity of decoding with histogram pruning

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

- Number of translation options is linear with sentence length, hence:

$$O(\text{max stack size} \times \text{sentence length}^2)$$

- Quadratic complexity

Reordering Limits

- Limiting reordering to maximum reordering distance

- Typical reordering distance 5–8 words

  - depending on language pair
  - larger reordering limit hurts translation quality

- Reduces complexity to linear

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

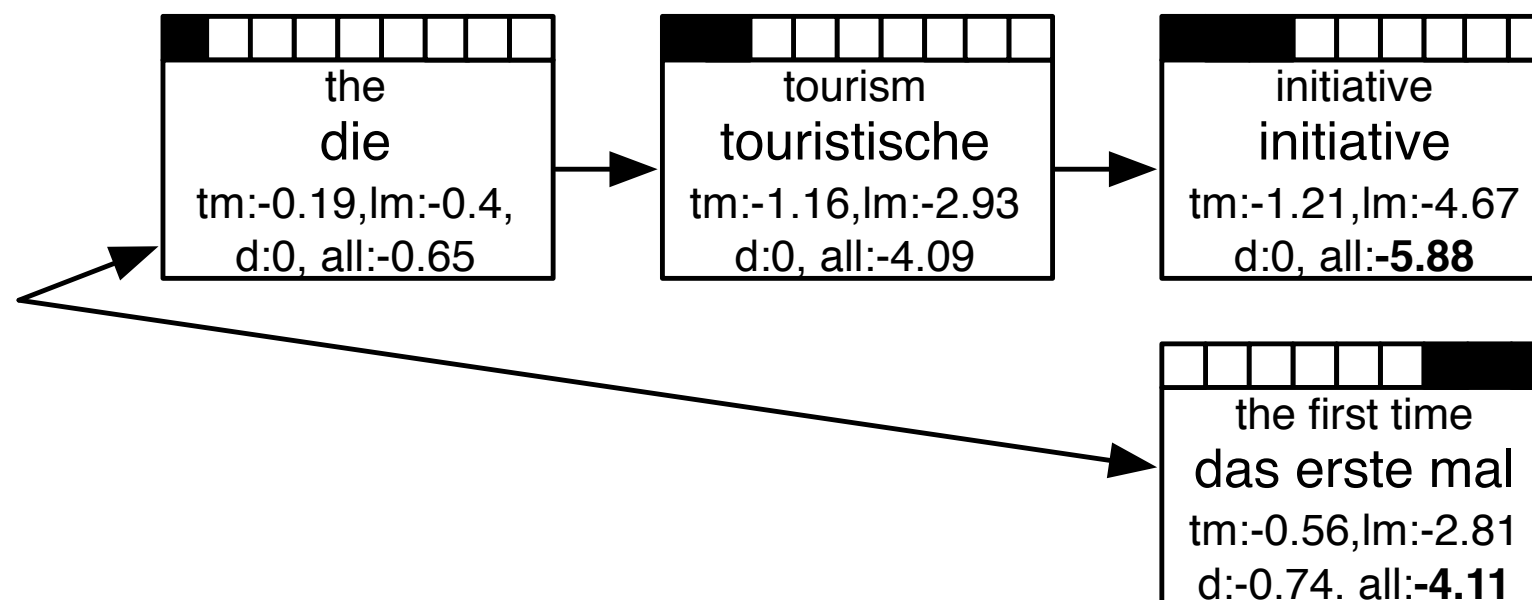- Speed / quality trade-off by setting maximum stack size

Search Errors

- We are using a **heuristic search** to prune the search space

- There are no guarantees of admissibility (like in A* search)

- We may therefore prune out a partial hypothesis that would have lead to the most probable translation, if we hadn't pruned it early on

Model Errors

- Model errors are different than search errors

- If the model scores an incorrect translation higher than higher than the correct translation, then it is the model's fault not the search strategy's fault

- How can these two interact?

Translating the Easy Part First?

the tourism initiative addresses this for the first time

the **die** tm:-0.19,lm:-0.4, d:0, all:-0.65	tourism **touristische** tm:-1.16,lm:-2.93 d:0, all:-4.09	initiative **initiative** tm:-1.21,lm:-4.67 d:0, all:**-5.88**

the first time
das erste mal
tm:-0.56,lm:-2.81
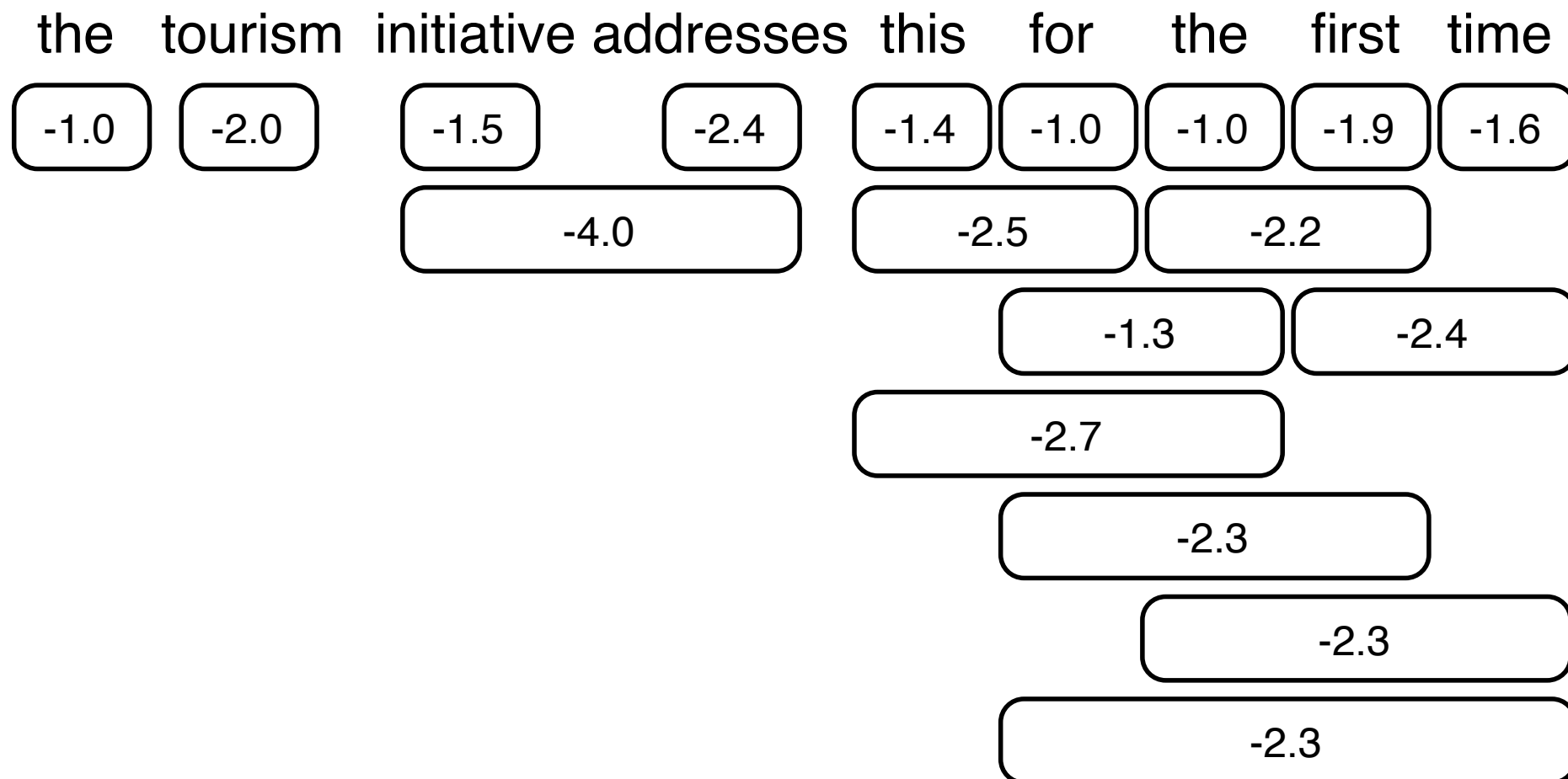d:-0.74. all:**-4.11**

both hypotheses translate 3 words
worse hypothesis has better score

Estimating Future Cost

- Future cost estimate: how expensive is translation of rest of sentence?

- Optimistic: choose cheapest translation options

- Cost for each translation option

  - **translation model**: cost known

  - **language model:** output words known, but not context
    $\rightarrow$ estimate without context

  - **reordering model:** unknown, ignored for future cost estimation

Cost Estimates from Translation Options

the   tourism   initiative   addresses   this   for   the   first   time

| -1.0 | -2.0 | | -1.5 | | -2.4 | -1.4 | -1.0 | -1.0 | -1.9 | -1.6 |

| -4.0 | -2.5 | -2.2 |

| -1.3 | -2.4 |

| -2.7 |

| -2.3 |

| -2.3 |

| -2.3 |

cost of cheapest translation options for each input span (log-probabilities)


Cost Estimates for all Spans

- Compute cost estimate for all contiguous spans by combining cheapest options

first word	future cost estimate for n words (from first)								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
the	-1.0	-3.0	-4.5	-6.9	-8.3	-9.3	-9.6	-10.6	-10.6
tourism	-2.0	-3.5	-5.9	-7.3	-8.3	-8.6	-9.6	-9.6	
initiative	-1.5	-3.9	-5.3	-6.3	-6.6	-7.6	-7.6		
addresses	-2.4	-3.8	-4.8	-5.1	-6.1	-6.1			
this	-1.4	-2.4	-2.7	-3.7	-3.7				
for	-1.0	-1.3	-2.3	-2.3					
the	-1.0	-2.2	-2.3						
first	-1.9	-2.4							
time	-1.6								

- Function words cheaper (the: -1.0) than content words (tourism -2.0)
- Common phrases cheaper (for the first time: -2.3)
  than unusual ones (tourism initiative addresses: -5.9)

Combining Score and Future Cost
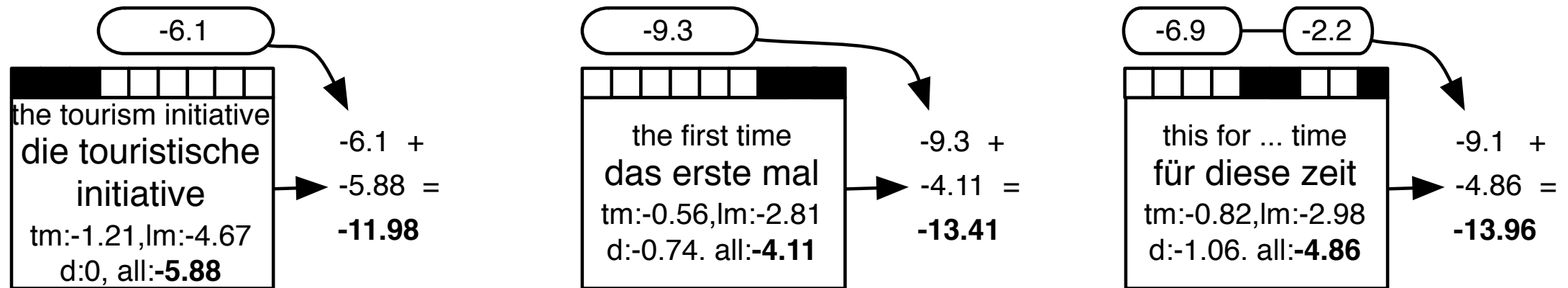
- Hypothesis score and future cost estimate are combined for pruning

  - left hypothesis starts with hard part: the tourism initiative
    score: -5.88, future cost: -6.1 → total cost -11.98

  - middle hypothesis starts with easiest part: the first time
    score: -4.11, future cost: -9.3 → total cost -13.41

  - right hypothesis picks easy parts: this for ... time
   score: -4.86, future cost: -9.1 → total cost -13.96

Alternate to Future Cost Estimation

- In the setup that we described so far, we had N stacks, N=number of source words

- Hypotheses were added to the stack based on **how many** source words they covered

- We could group hypotheses into stacked based on **which source words** they cover

- This eliminates the needs for FCE, but requires more stacks

Other Decoding Algorithms

- A* search

- Greedy hill-climbing

- Using finite state transducers (standard toolkits)

A* Search
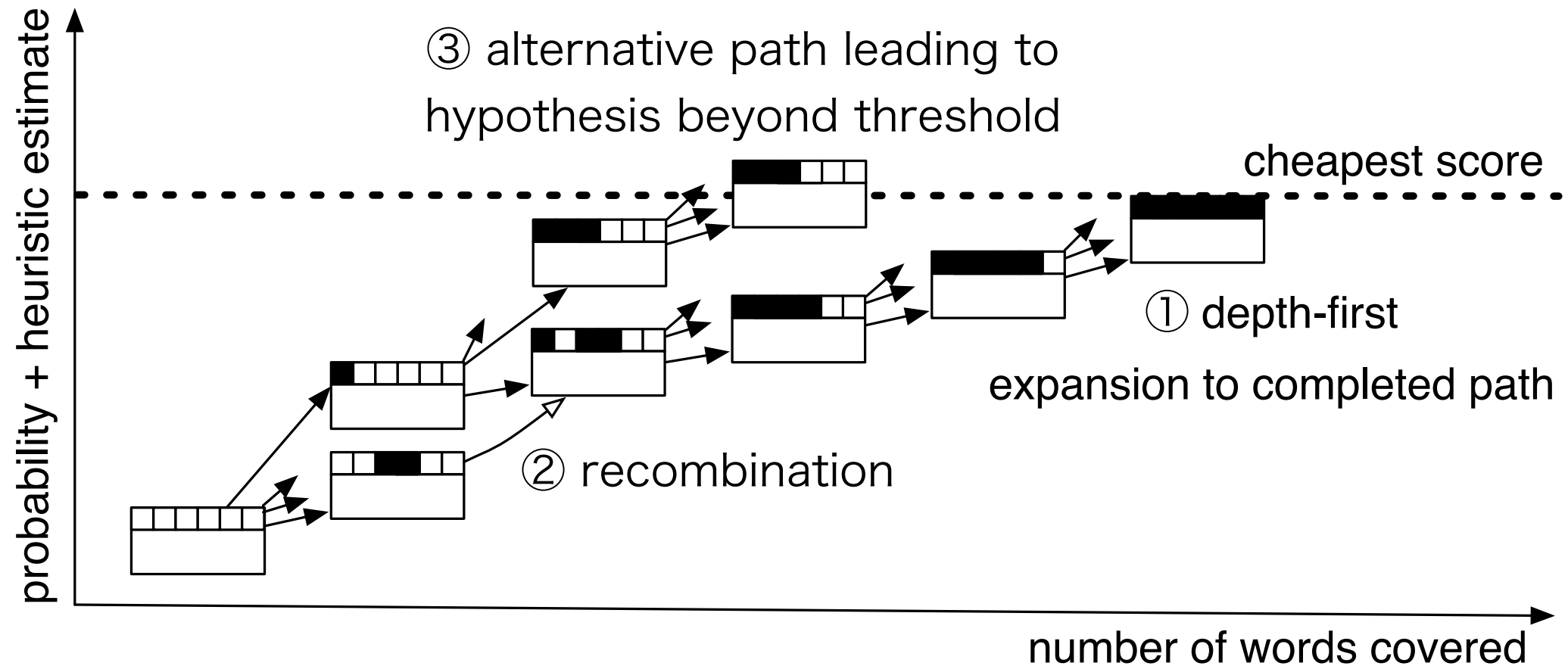
- Uses *admissible* future cost heuristic: never overestimates cost
- Translation agenda: create hypothesis with lowest score $+$ heuristic cost
- Done, when complete hypothesis created

Greedy Hill-Climbing

- Create one complete hypothesis with depth-first search (or other means)

- Search for better hypotheses by applying change operators

  - change the translation of a word or phrase
  - combine the translation of two words into a phrase
  - split up the translation of a phrase into two smaller phrase translations
  - move parts of the output into a different position
  - swap parts of the output with the output at a different part of the sentence

- Terminates if no operator application produces a better translation

Reading

- Read 6 from the textbook

Statistical Machine Translation

Philipp Koehn

CAMBRIDGE
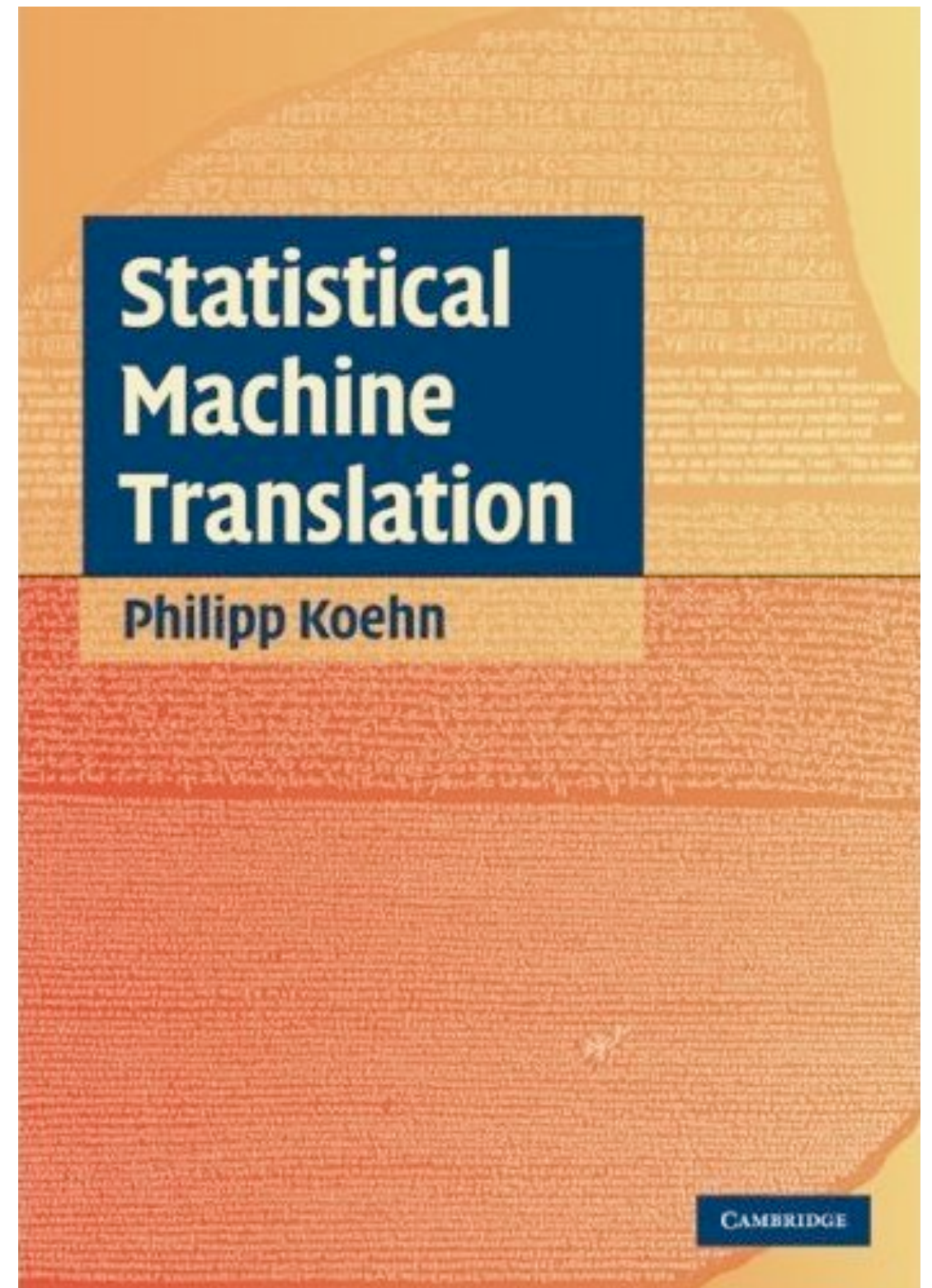
Announcements

- HW2 was released on Friday

- HW2 due **Thursday Feb 19th** at 11:59pm

- **Start early!!**

- This assignment takes longer to run than the last HW, and a good strategy for doing well involves running the search with wider beam widths (taking more CPU cycles).