Phrase-based Machine Translation

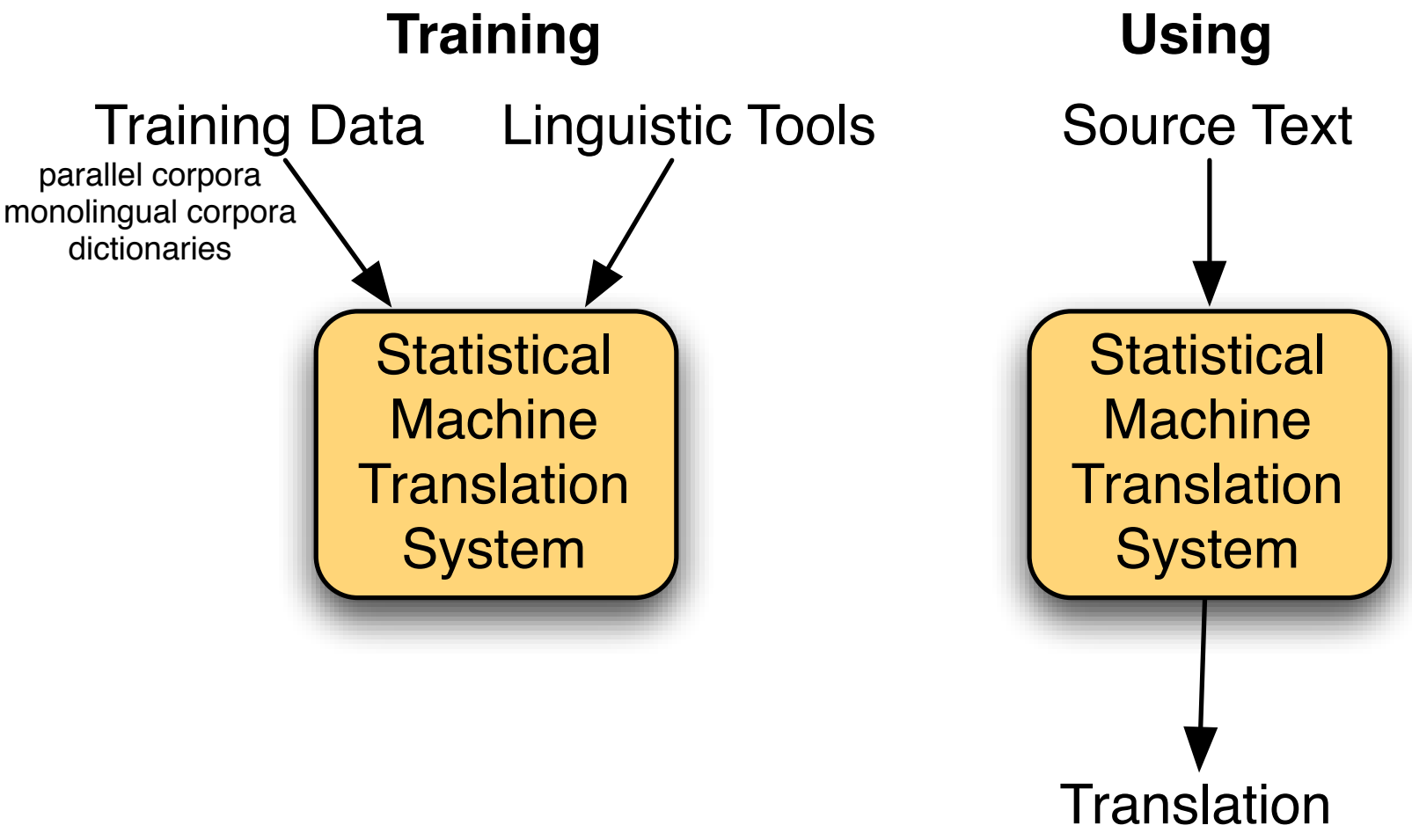
Matthias Huck

(partially adapted from slides originally by Philipp Koehn)

Center for Information and Language Processing
LMU Munich

10 January 2019

Statistical Machine Translation (SMT): Basic Idea



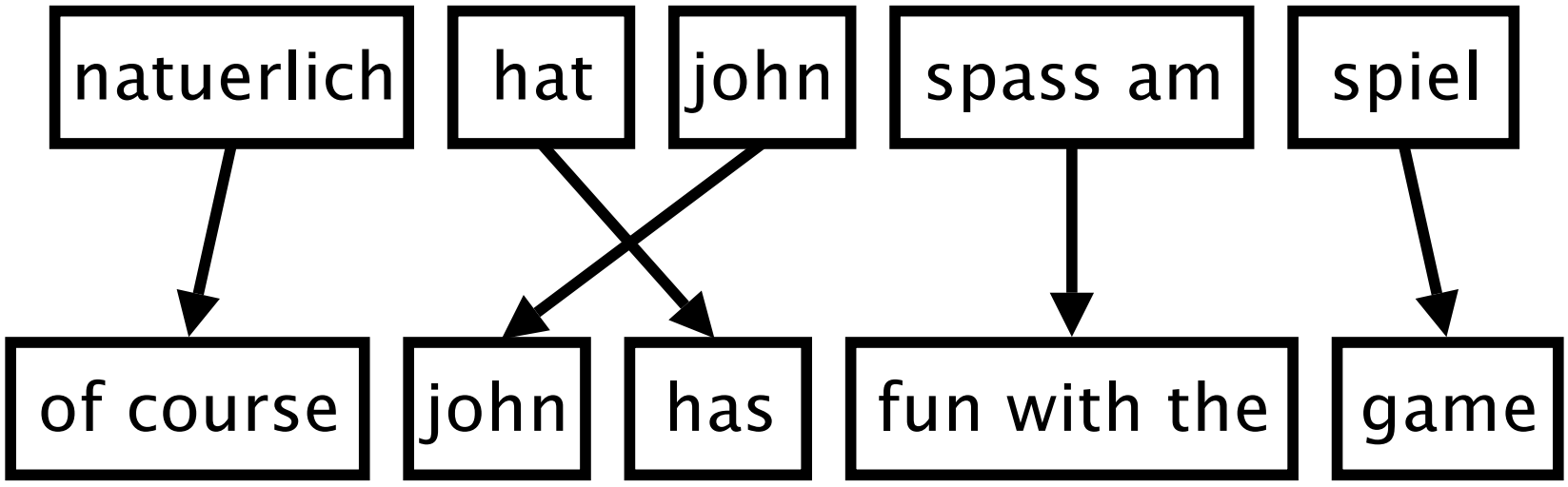
Different Paradigms (1)

- **Word-based translation**
  - Word-based models translate *words* as atomic units
- **Phrase-based translation**
  - Phrase-based models translate *phrases* as atomic units
  - Advantages:
    - larger atomic units capture more context in translation
    - local word reorderings are handled within phrases
    - the more training data, the longer phrases can be learned reliably
  - Dominant approach for many years

Different Paradigms (2)

- **Hierarchical translation**
  - Allows for *gaps* in the phrases
  - Formalization as a synchronous context-free grammar (SCFG)
  - Parsing-based decoding (CYK+)
- **Syntax-based translation**
  - Comparable to hierarchical, but with *linguistic well-formedness constraints*
  - Source syntax, target syntax, or both
  - SCFG or synchronous tree substitution grammar (STSG)
- **Neural machine translation**
  - Sequence-to-sequence classification using artificial neural networks

Phrase-based Approach (1)



- Foreign input is segmented into phrases
- Each phrase is translated to English
- Phrases can be reordered

Phrase-based Approach (2)

- Main knowledge source: a **phrase table**, containing bilingual word sequences and their translation probabilities

Foreign $\bar{f}$	**English** $\bar{e}$	**Probability** $\phi(\bar{e}\|\bar{f})$
natürlich	of course	0.5
natürlich	naturally	0.3
natürlich	of course ,	0.15
natürlich	, of course ,	0.05

Phrase-based Approach (3)

- To translate a foreign sentence **f**, we have to solve

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f})$$

- In reality what we do is (Viterbi approximation)

$$(a, \mathbf{e})_{\text{best}} = \text{argmax}_{(a,\mathbf{e})}\ p(a, \mathbf{e}|\mathbf{f})$$

- This is known as **decoding**
  - it's a search problem
  - the search space is huge

Phrase-based Approach (4)

- **Bayes' rule**:

$$\begin{aligned} \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}}\ \frac{p_{\text{TM}}(\mathbf{f}|\mathbf{e})\ p_{\text{LM}}(\mathbf{e})}{p(\mathbf{f})} \\ &= \text{argmax}_{\mathbf{e}}\ p_{\text{TM}}(\mathbf{f}|\mathbf{e})\ p_{\text{LM}}(\mathbf{e}) \end{aligned}$$

  - translation model $p_{\text{TM}}(\mathbf{f}|\mathbf{e})$
  - language model $p_{\text{LM}}(\mathbf{e})$

- **Decomposition** of the translation model:

$$p_{\text{TM}}(\mathbf{f}|\mathbf{e}) = p_{\text{TM}}(\bar{f}_1^K|\bar{e}_1^K) = \prod_{k=1}^{K} \phi(\bar{f}_k|\bar{e}_k)$$

Phrase-based Approach: Illustration (by Kevin Knight)

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				

Phrase-based Approach: Illustration (by Kevin Knight)

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 but	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			
		including 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				

Phrase-based Approach: Illustration (by Kevin Knight)

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				

Phrase-based Approach: Illustration (by Kevin Knight)

SMT Pipeline (1)

1. **Data collection**
   - Parallel corpora for training translation model (TM)
   - Target-side monolingual corpora for training language model (LM)
   - Separate development and test sets for tuning
2. **Data preparation, preprocessing**
   - Cleaning: unrealiable parts should be removed
   - Sentence segmentation, sentence alignment of parallel corpora
   - Preprocessing steps (word tokenization, segmentation, casing, ...)

SMT Pipeline (2)

3. **Training**
   - Word alignment
   - Phrase extraction
   - Language model estimation
   - ...
4. **Tuning**
   - In practice, the decoder combines a set of features
   - Each feature has a weight which determines its contribution to the overall model score
   - We can optimize the values of the feature weights such that translation quality is maximized on a development set

SMT Pipeline (3)

5. **Decoding**
   - Translate unseen source text
   - Decoding = searching the space of possible translations for the best hypothesis according to the model score
   - Dynamic programming, beam search
6. **Postprocessing**
   - Detokenization, recasing, ...
7. **Evaluation**
   - Automatic metrics to assess translation quality by comparing the hypothesis with a reference translation: BLEU, TER, METEOR, ...
   - Human judgement

Data Collection

ALBERTO ALESINA
Alberto Alesina is Professor of Economics at *Harvard University*. READ MORE

FRANCESCO GIAVAZZI
Francesco Giavazzi is Professor of Economics at *Bocconi University*, Milan. READ MORE

JUN 24, 2003 0

Berlusconi at Bay

Silvio Berlusconi, who becomes President of the European Union on July 1 st , is a man of vision who once loved risk--and whose business bets paid off big. In the 1960's, he was the first to see that Milan, then a traditional Italian city where people walked to work, would become a modern metropolis, surrounded by American-style suburbs. So his fortune began in real estate development.

Fifteen years later, Signor Berlusoni understood that the Italian state's monopoly of television would not survive and jump-started what became Italy's main privately owned media group. But you don't win in TV and the real estate business without the right political connections. On both occasions, Berlusconi outwitted his competitors by siding with the Socialists, at the time the rising stars of Italian political life. His long association with Bettino

Data Collection

ALBERTO ALESINA
Alberto Alesina is Professor of Economics at *Harvard University*. READ MORE

FRANCESCO GIAVAZZI
Francesco Giavazzi is Professor of Economics at *Bocconi University*, Milan. READ MORE

JUN 24, 2003 0

Berlusconi in Bedrängnis

Silvio Berlusconi, der am 1. Juli die Ratspräsidentschaft in der Europäischen Union übernehmen wird, ist ein Mann mit Visionen, der einst das Risiko liebte - und dessen Geschäftssinn sich in großem Stil bezahlt machte. In den sechziger Jahren des vorigen Jahrhunderts war er der erste, der die damalige Provinzhauptstadt Mailand, wo die Menschen zu Fuß zur Arbeit gingen, als moderne Metropole mit Vororten in amerikanischem Stil sah. So begann er mit Immobilien sein Vermögen aufzubauen.

Fünfzehn Jahre später erkannte Signor Berlusconi, dass die Tage des staatlichen Fernsehmonopols in Italien gezählt sein würden und so gründete er ein Unternehmen, das später Italiens wichtigste private Mediengruppe werden sollte. Aber ohne die richtigen Verbindungen in die Politik lässt sich im Immobiliengeschäft und im TV nichts ausrichten. In

Sentence Segmentation

Roxy Ann Peak is a 3,576-foot-tall (1,090 m) mountain in the Western Cascade Range in the U.S. state of Oregon. Composed of several geologic layers, the majority of the peak is of volcanic origin and dates to the early Oligocene. It is primarily covered by oak savanna and open grassland on its lower slopes, and mixed coniferous forest on its upper slopes and summit. Despite the peak's relatively small topographic prominence of 753 feet (230 m), it rises 2,200 feet (670 m) above Medford, and it is the city's most important viewshed, open space reserve, and recreational resource.

Sentence Segmentation

Roxy Ann Peak is a 3,576-foot-tall (1,090 m) mountain in the Western Cascade Range in the U.S. state of Oregon.

Composed of several geologic layers, the majority of the peak is of volcanic origin and dates to the early Oligocene.

It is primarily covered by oak savanna and open grassland on its lower slopes, and mixed coniferous forest on its upper slopes and summit.

Despite the peak's relatively small topographic prominence of 753 feet (230 m), it rises 2,200 feet (670 m) above Medford, and it is the city's most important viewshed, open space reserve, and recreational resource.

Parallel Sentence Alignment

Je vous invite à vous lever pour cette minute de silence.

(Le Parlement, debout, observe une minute de silence)

Madame la Présidente, c'est une motion de procédure.

Please rise, then, for this minute's silence.

(The House rose and observed a minute's silence)

Madam President, on a point of order.

Astronomes Introduction Vidéo d'introduction

Qu'est-ce que l'astronomie?

Souvent considéré comme la plus ancienne des sciences, elle découle de notre étonnement et de nos questionnements envers le ciel L'astronomie est la science qui étudie l'Univers au-delà de l'atmosphère terrestre.

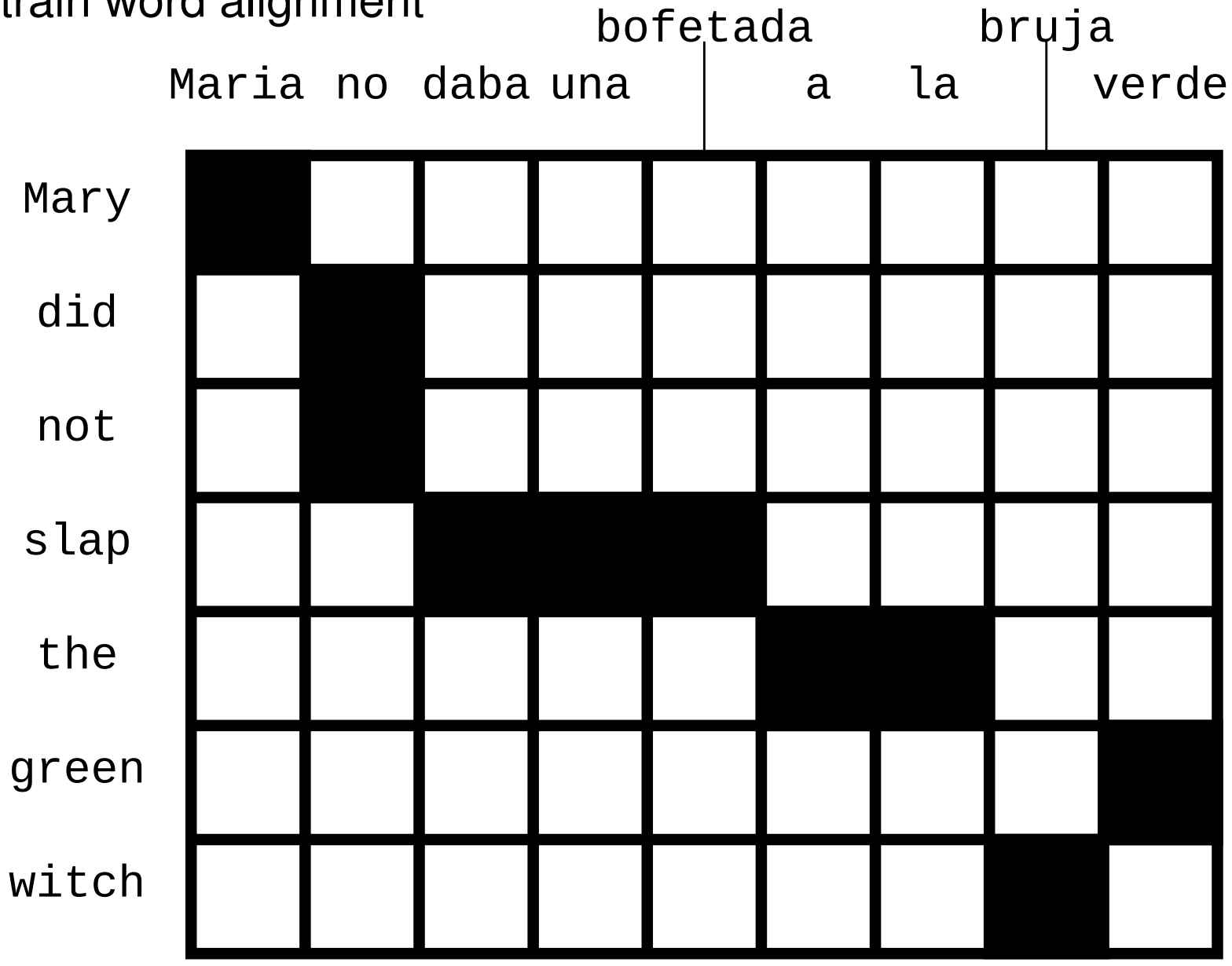
Astronomers Introduction Introduction video

What is Astronomy?

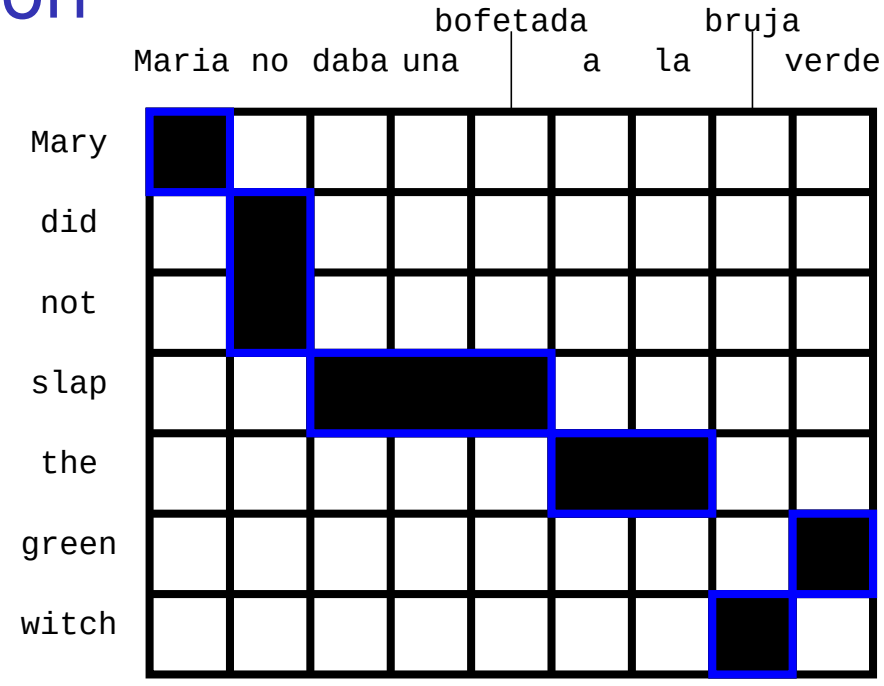
Often considered the oldest science, it was born of our amazement at the sky and our need to question Astronomy is the science of space beyond Earth's atmosphere.

Word Alignment

- After further preprocessing (tokenization, word segmentation), train word alignment

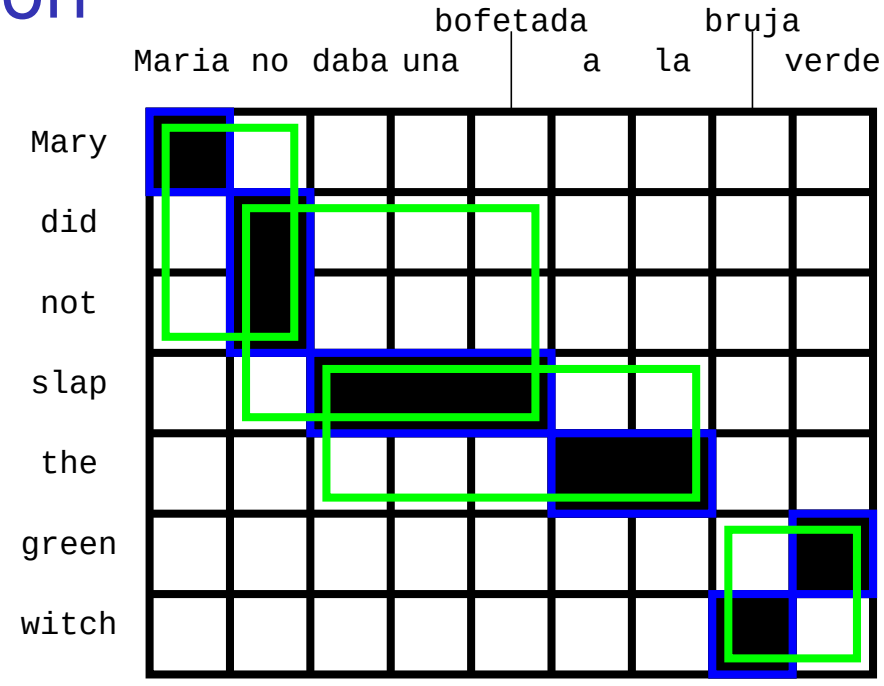


Phrase Extraction



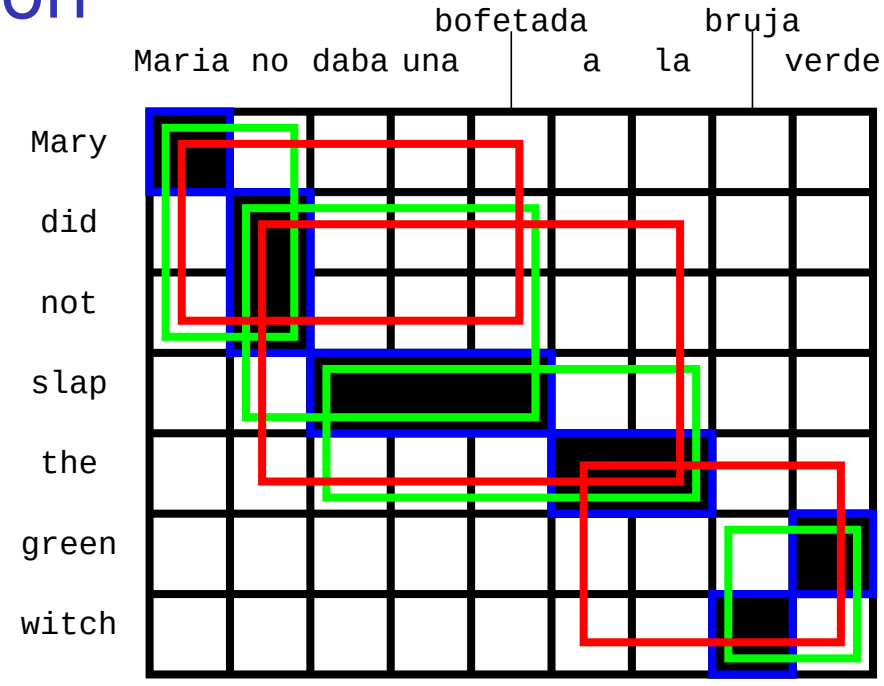
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Phrase Extraction



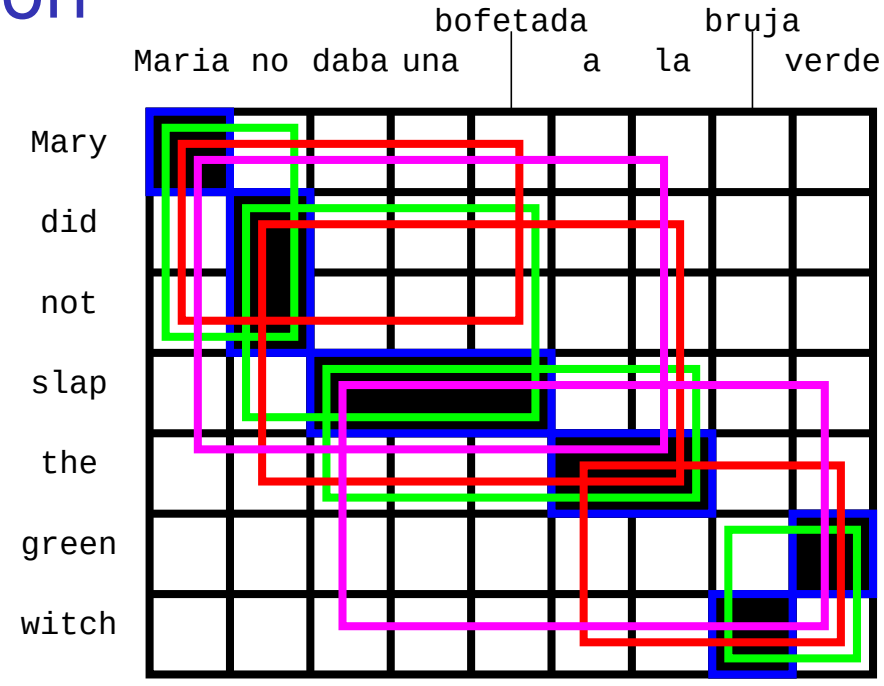
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Phrase Extraction



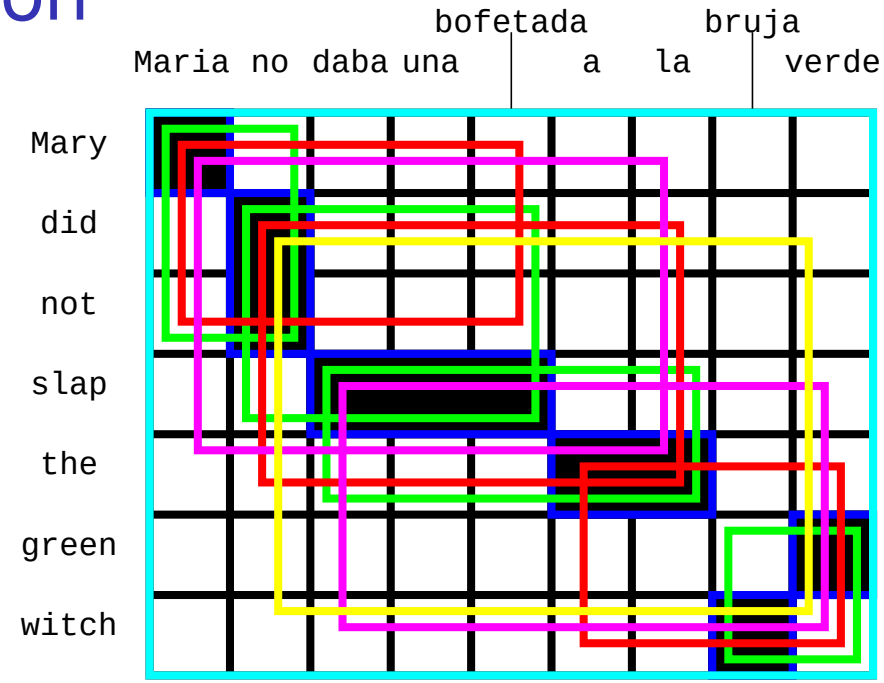
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap), (no daba una bofetada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

Phrase Extraction



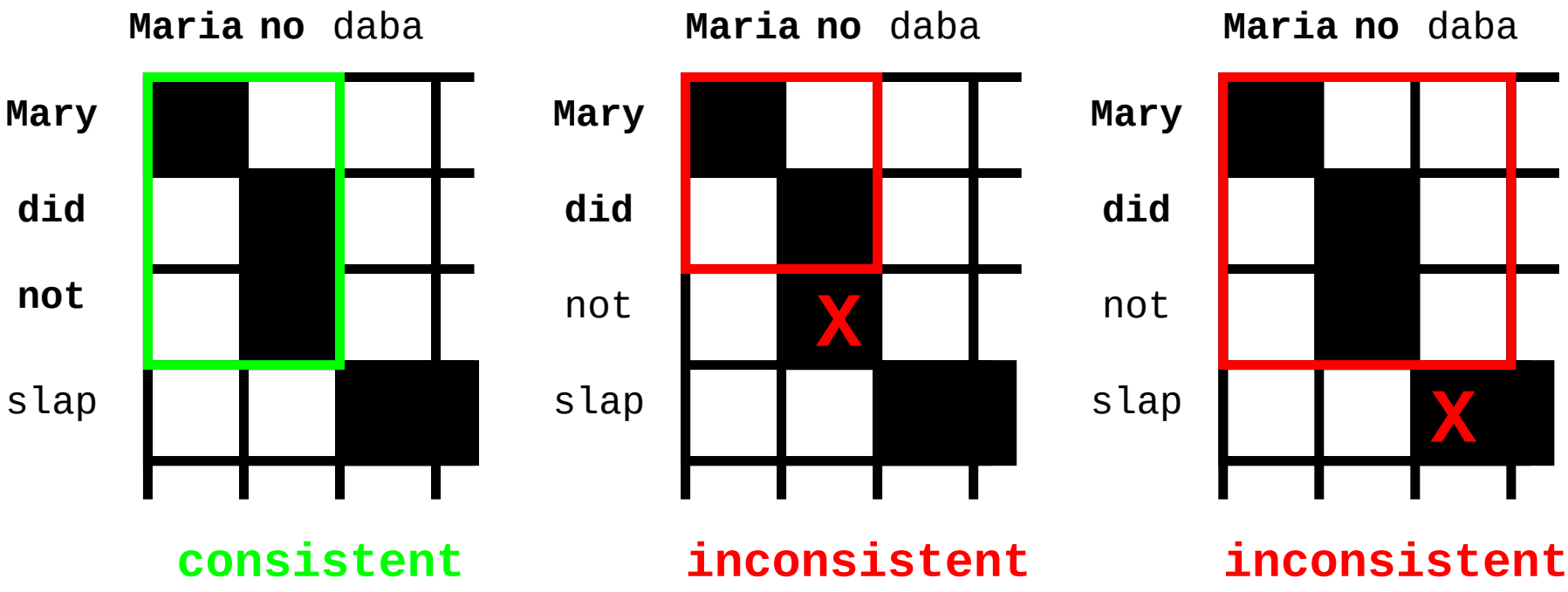
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch),
(Maria no daba una bofetada, Mary did not slap), (no daba una bofetada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

Phrase Extraction



(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)

Phrase Extraction Criteria



- A valid phrase has to contain all alignment points for all covered words
- A valid phrase has to contain at least one alignment point

Phrase Translation Probabilities

- **Phrase table**: stores all phrases $(\bar{e}, \bar{f})$ extracted from the data
- We need to assign **probabilities** to extracted phrases
- Calculate relative frequencies:

$$\phi(\bar{e}|\bar{f}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{e}'} \text{count}(\bar{e}', \bar{f})}$$

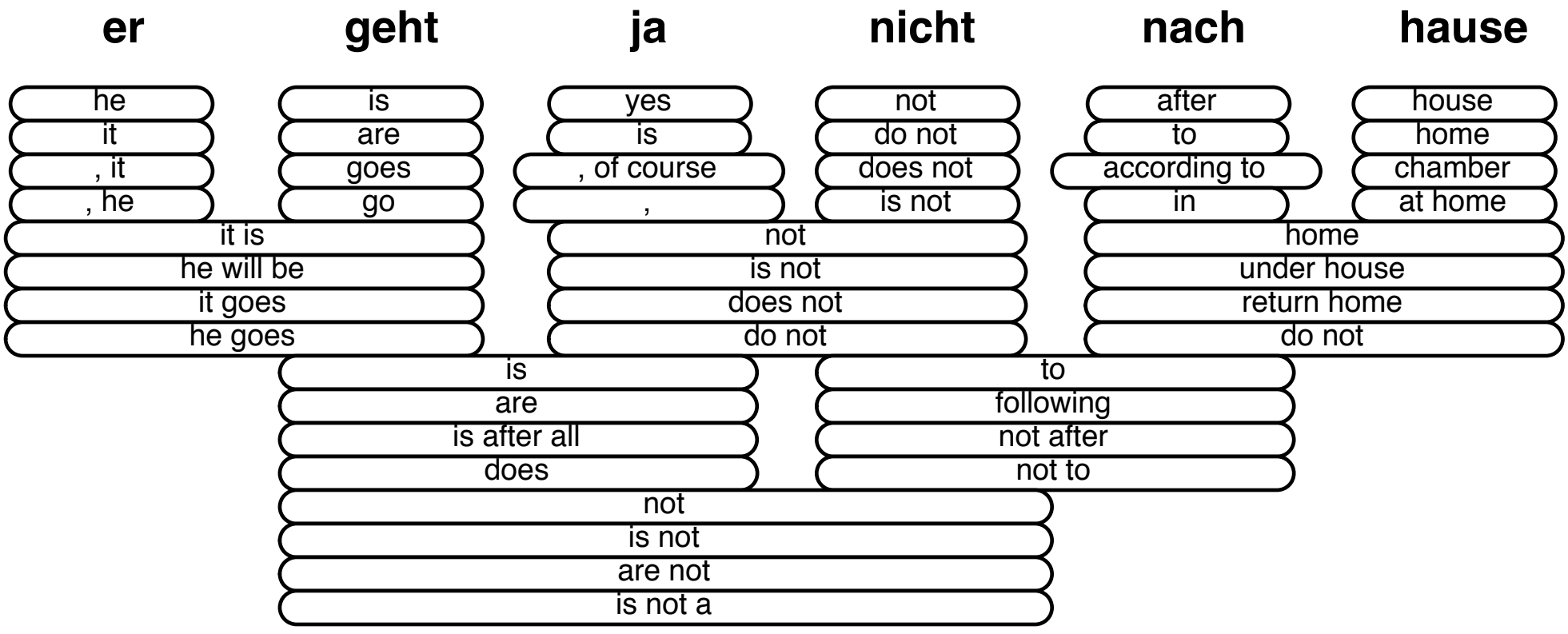
$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}'} \text{count}(\bar{e}, \bar{f}')}$$

Real Example

- Phrase translations for den Vorschlag from German into English:

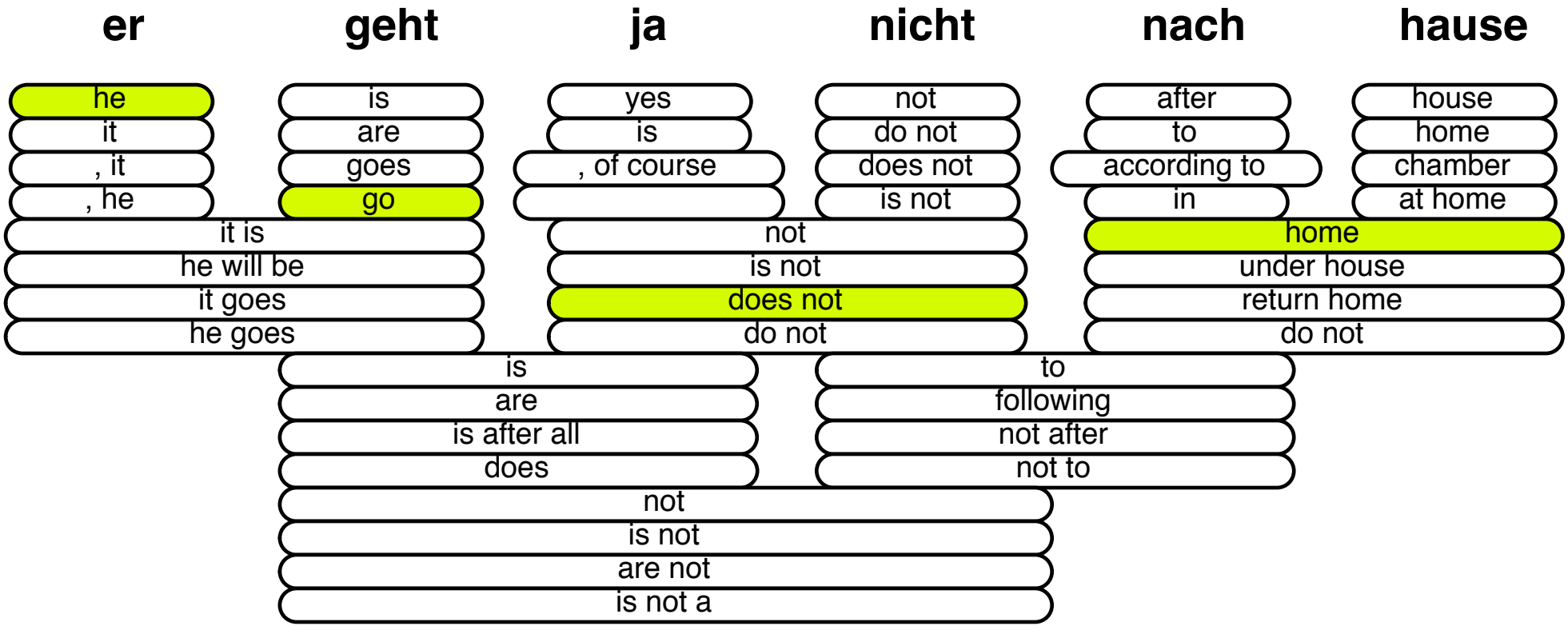
English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Phrase-based Decoding



- Many translation options to choose from

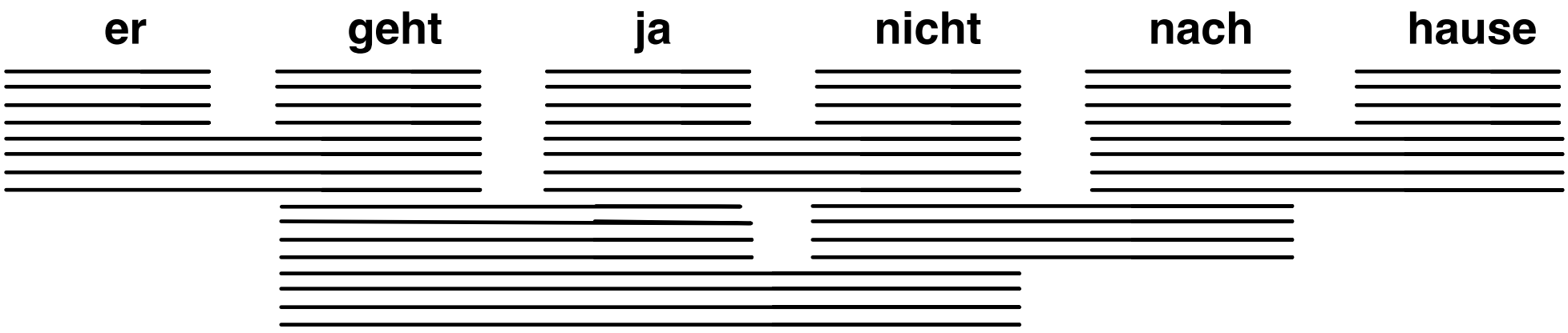
Phrase-based Decoding



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

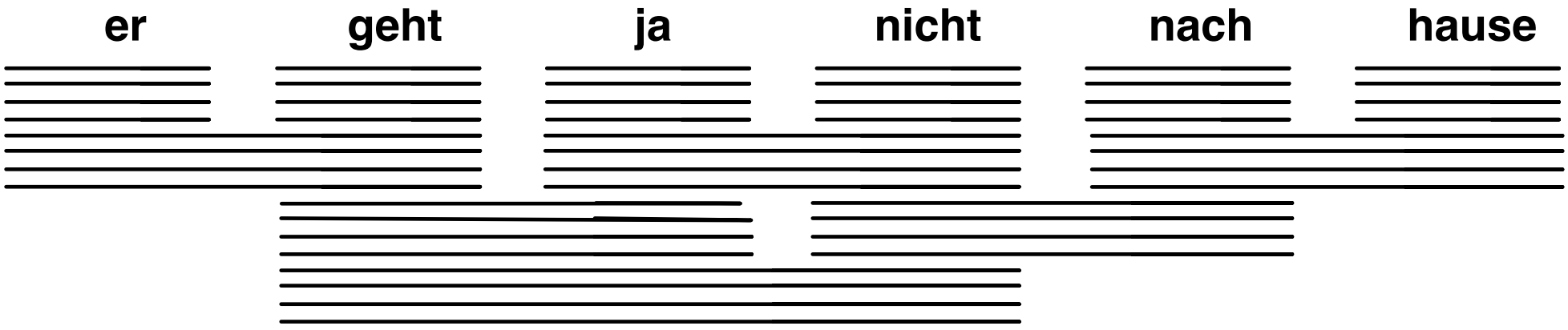
→ Search problem solved by beam search

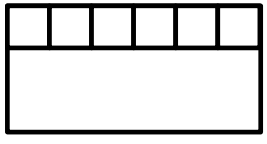
Decoding: Precompute Translation Options



consult phrase translation table for all input phrases

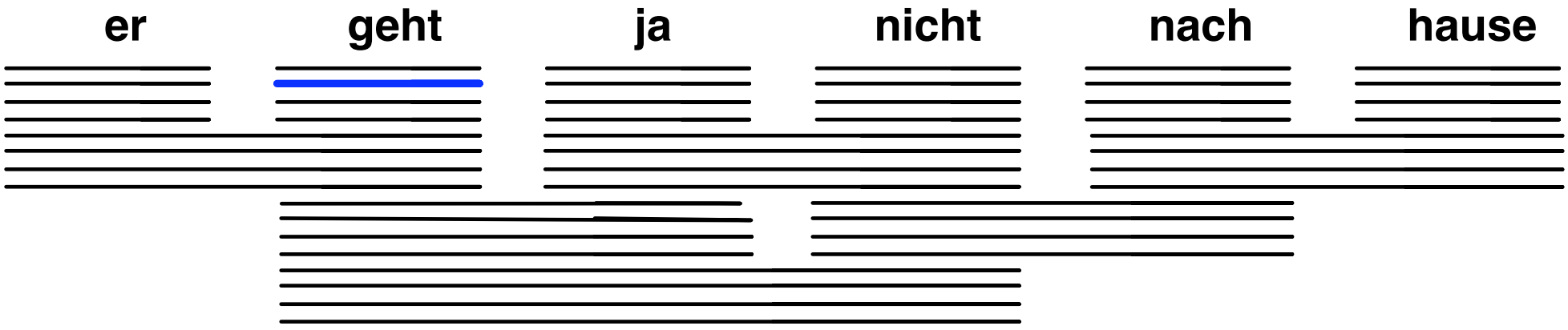
Decoding: Start with Initial Hypothesis

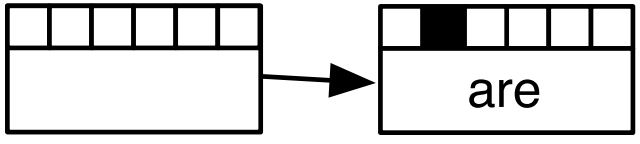




initial hypothesis: no input words covered, no output produced

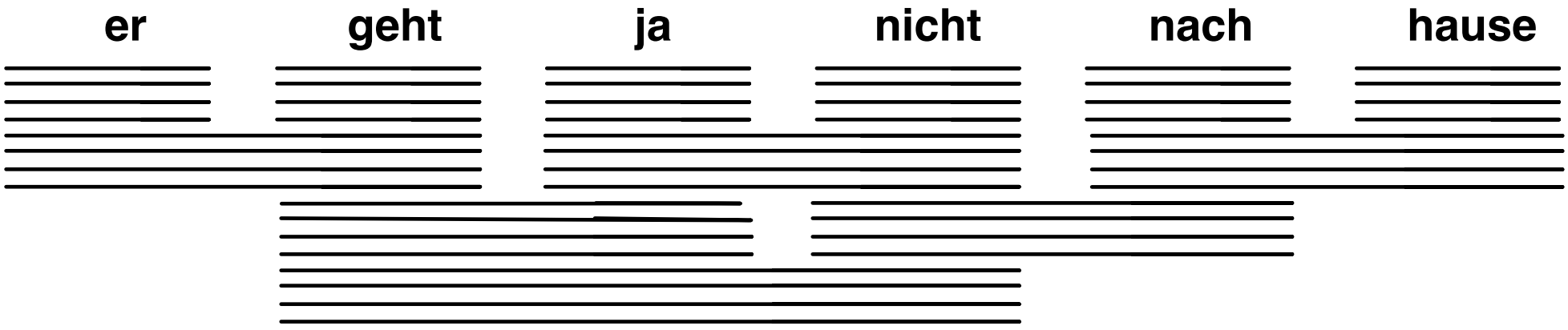
Decoding: Hypothesis Expansion

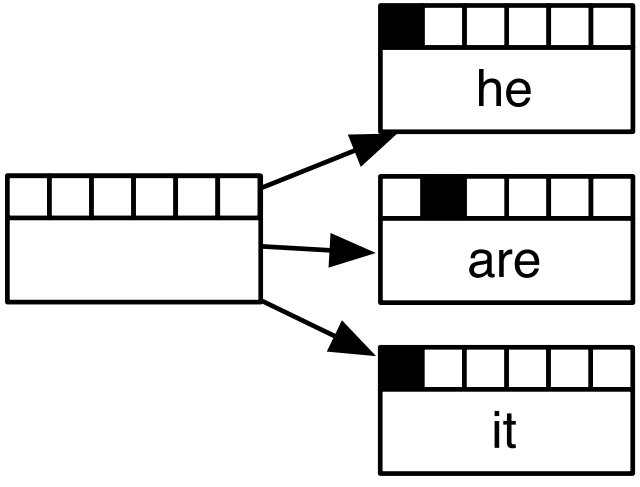




pick any translation option, create new hypothesis

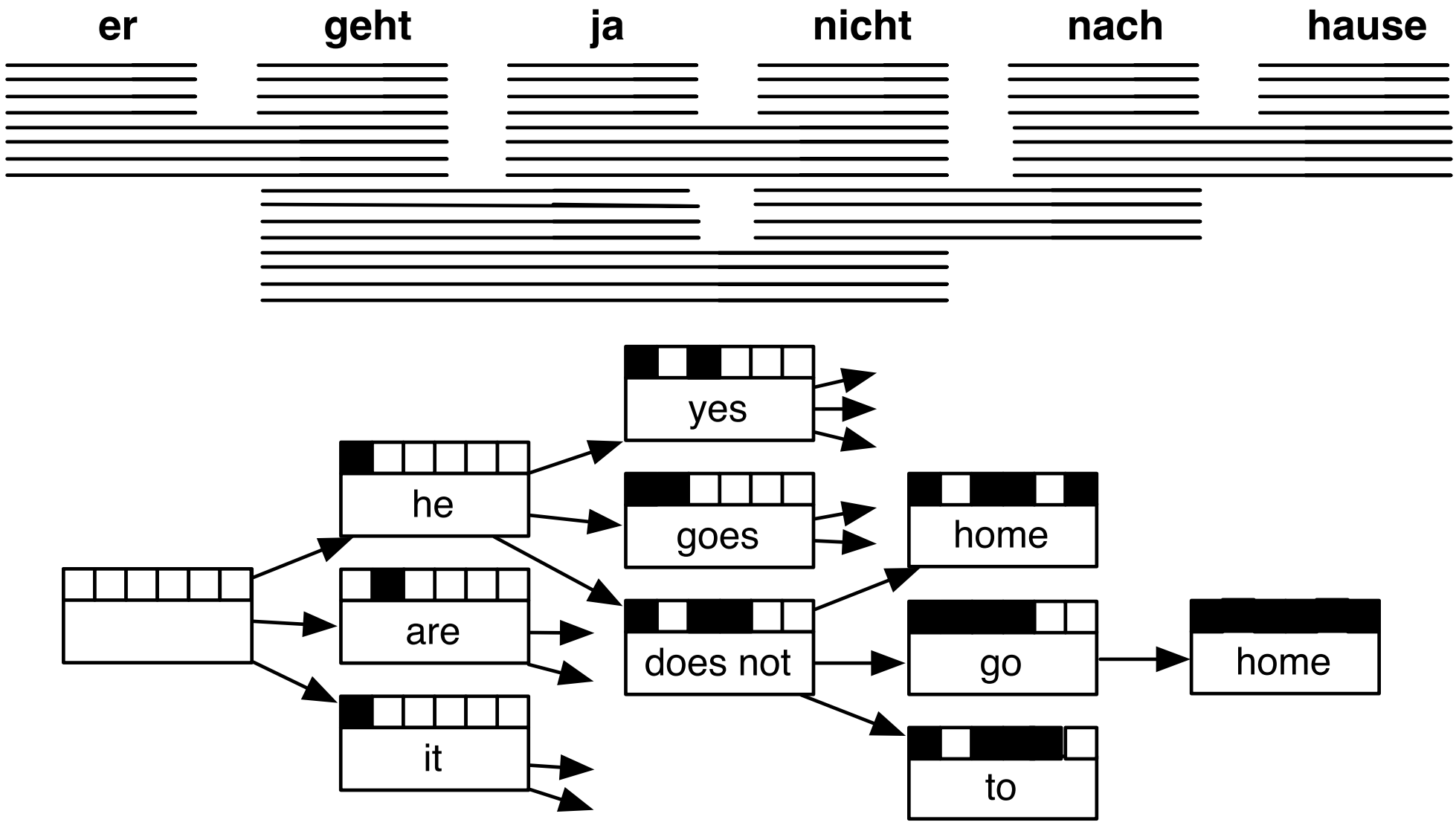
Decoding: Hypothesis Expansion





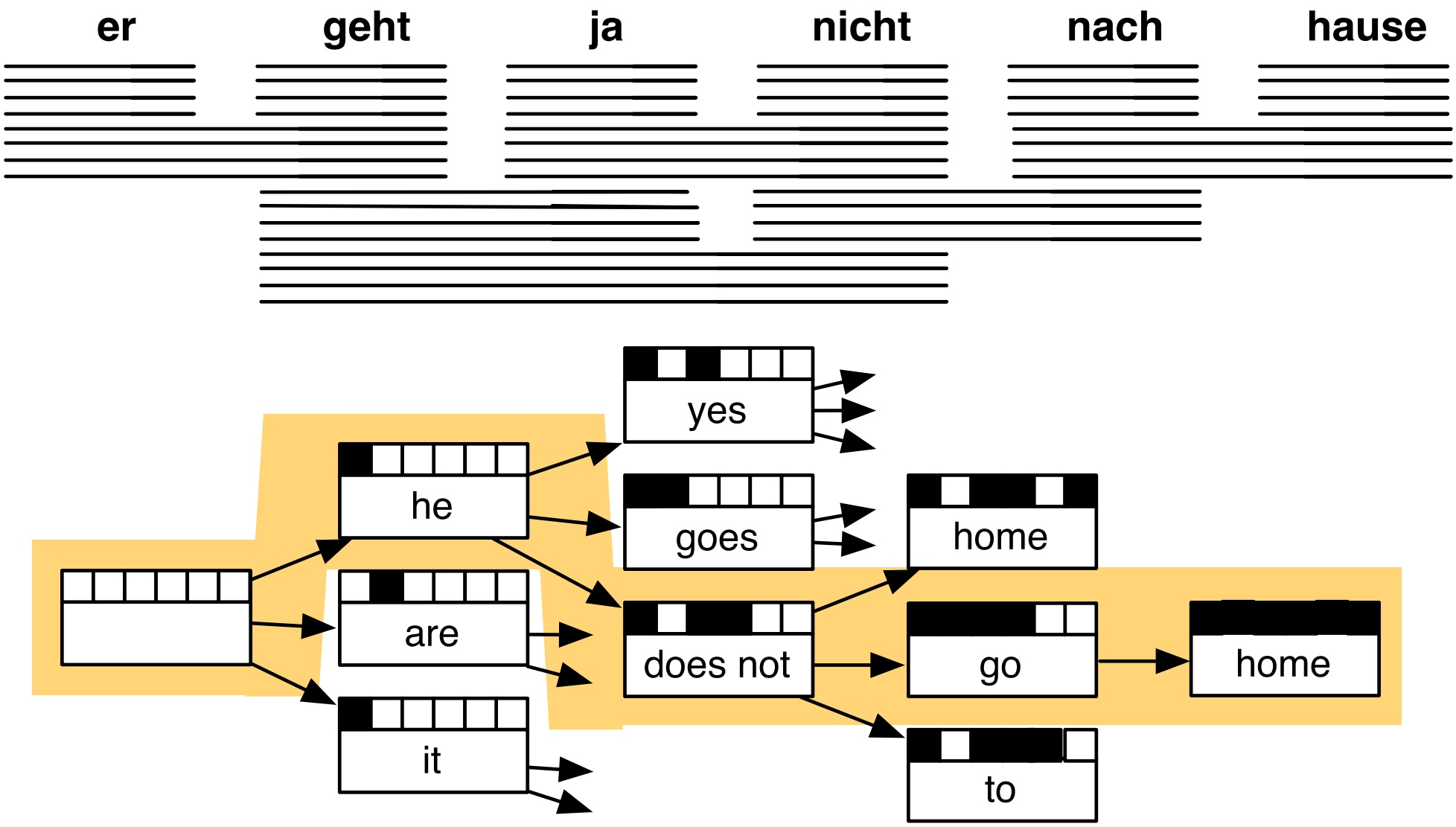
create hypotheses for all other translation options

Decoding: Hypothesis Expansion



also create hypotheses from created partial hypotheses

Decoding: Find Best Path

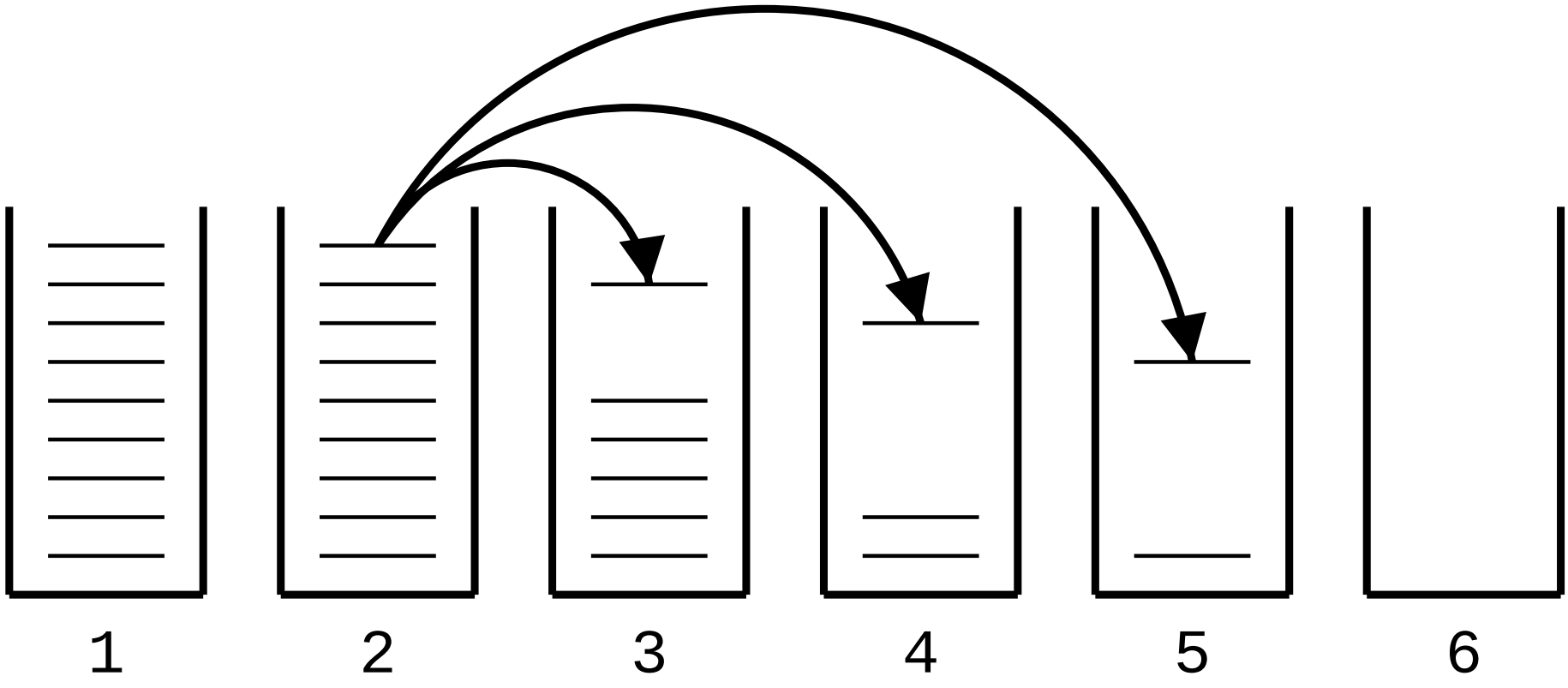


backtrack from highest scoring complete hypothesis

Pruning

- Heuristically discard weak hypotheses early: **beam search**
- Organize hypotheses in **stacks** (actually priority queues), e.g. by
  - same source words covered
  - same number of source words covered
- Compare hypotheses in stacks, discard bad ones
  - **histogram pruning**: keep top *k* hypotheses in each stack (e.g., *k*=100)
  - **threshold pruning**: keep hypotheses that are at least α times the score of the best hypothesis in the stack (e.g., α = 0.001)

Hypothesis Stacks



- Organization of hypotheses into stacks
  - here: based on **number of source words** translated
  - during translation all hypotheses from one stack are expanded
  - expanded hypotheses are placed into next stacks

From Bayes to a Combination of Many Features

- We work in logarithmic space:

$$\begin{aligned} \mathbf{e}_{\text{best}} &= \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}}\ p_{\text{TM}}(\mathbf{f}|\mathbf{e})\ p_{\text{LM}}(\mathbf{e}) \\ &= \text{argmax}_{\mathbf{e}}\ \log p_{\text{TM}}(\mathbf{f}|\mathbf{e}) + \log p_{\text{LM}}(\mathbf{e}) \end{aligned}$$

- As the TM and the LM might not be equally important, we could weight them differently:

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\ \lambda_{\text{TM}} \log p_{\text{TM}}(\mathbf{f}|\mathbf{e}) + \lambda_{\text{LM}} \log p_{\text{LM}}(\mathbf{e})$$

- More feature functions can be introduced by linearly combining all of them:

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f})$$

(Log-)linear Model

- Decoding: Given the model, find the best translation

$$\mathbf{e}_{\text{best}} = \text{argmax}_{\mathbf{e}}\ p(\mathbf{e}|\mathbf{f})$$

- Our model is a **weighted combination** of many components

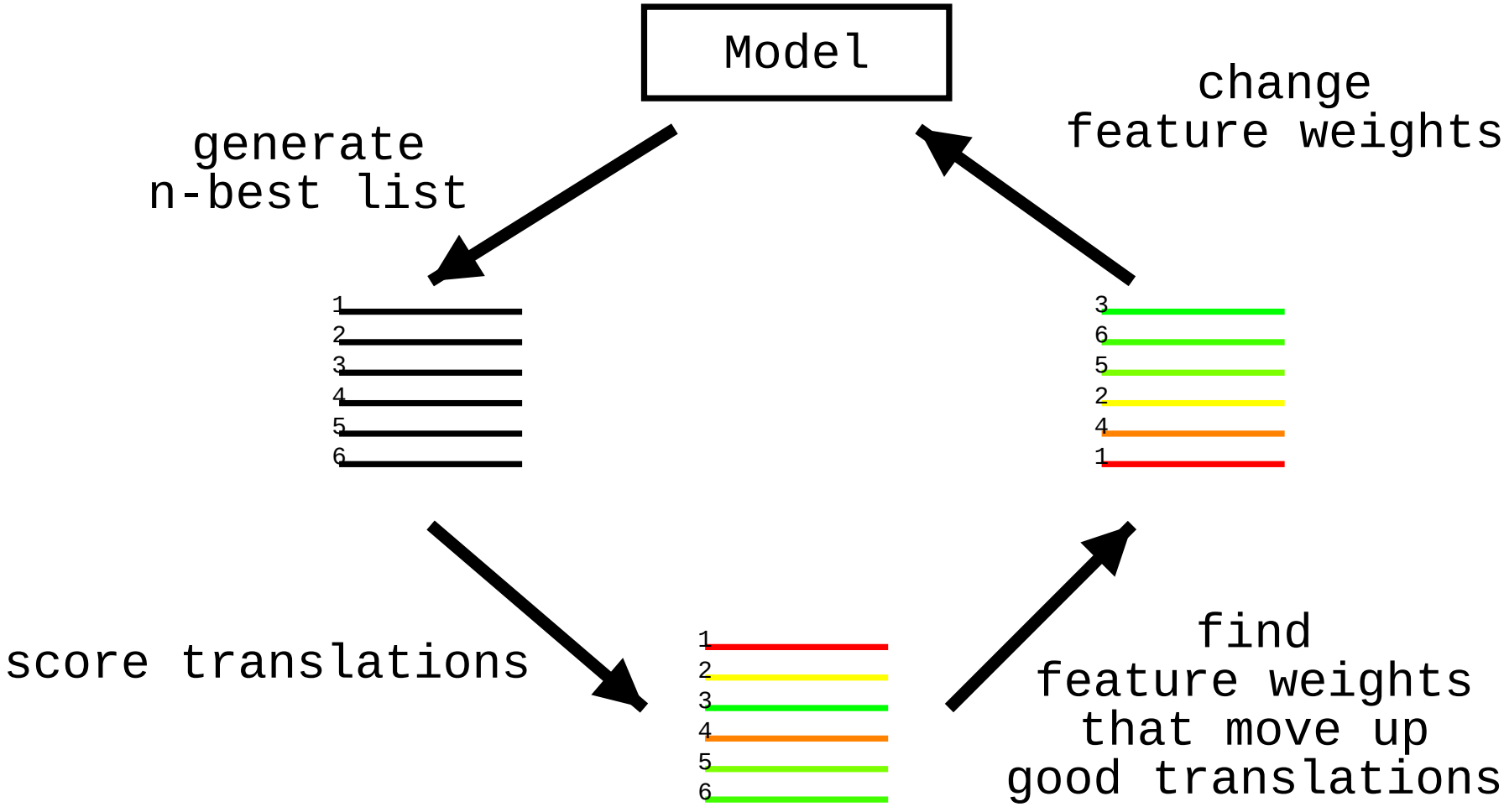
$$p(\mathbf{e}|\mathbf{f}) \propto \exp \sum_{m=1}^{M} \lambda_m \cdot h_m(\mathbf{e}, \mathbf{f})$$

where $h_m(\mathbf{e}, \mathbf{f})$ are **feature functions** and λ_m are **feature weights**

Features

- The decoder will employ the phrase translation probabilities along with other features to assign an overall score to each possible hypothesis
- **Standard features**:
  - Phrase translation log-probabilities
  - Lexical translation log-probabilities (from single-word based models)
  - Language model log-probability (from a target-side *n*-gram LM)
  - Phrase count
  - Word count
  - Distortion cost based on jump distances
  - Lexicalized reordering score
- Why might these be useful?
- And where do the feature weights come from?

Tuning: Optimizing the Feature Weights



- Minimum error rate training (MERT)

Evaluation

- **Automatic evaluation metrics**
  - Measure similarity between machine translation output and human-generated reference translation
  - Good metrics correlate well with human judgement
- Decode a held-out test set and – after postprocessing (detokenization, truecasing) – score with an automatic metric to **determine translation quality**
- BLEU
  - Most commonly used metric
  - n-gram precision ($n = 1 \ldots 4$), brevity penalty

The End! Questions?

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de