

30 Questions in Neural Machine Translation

Jia Xu

Graduate Center & Hunter College, CUNY

MT Tutorial @ JSALT'19

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

Odyssey, Homer, 700 BC

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

Odyssey, Homer, 700 BC



ancient Greek

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

Odyssey, Homer, 700 BC



ancient Greek

Man啊，给我唱一个英雄的故事吧，
宙斯啊，多才多艺的英雄！

Odyssey, Homer, 700 BC

我将歌唱那远航归来的伟业，
歌唱那位英雄，他被命运之神
驱逐，流落天涯，历尽千辛万苦，
在海上漂流了多年，吃尽了苦头。

我将歌唱那位英雄，他被命运之神
驱逐，流落天涯，历尽千辛万苦，
在海上漂流了多年，吃尽了苦头。

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

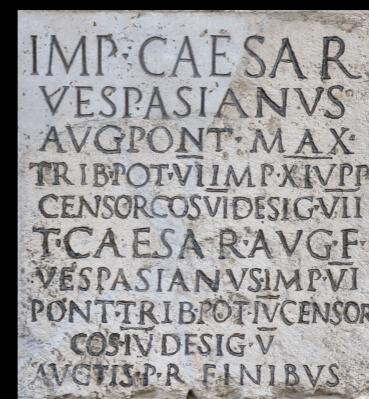
Odyssey, Homer, 700 BC



Virum mihi , Camena, insece versutum...



ancient Greek



Latin

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

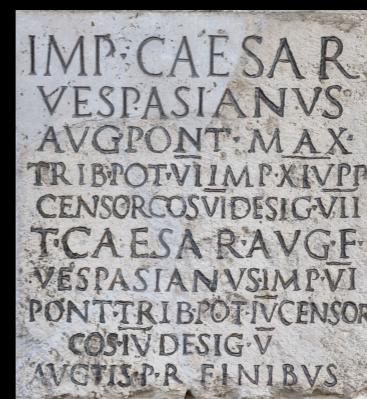
Odyssey, Homer, 700 BC



Virum mihi , Camena, insece versutum...



ancient Greek



Latin

2000 year old translation of
Sing in me, Muse, and through me tell the story of the man of twists and turns...

human language translation: old problem

Ἄνδρα μοι ἔννεπε, Μοῦσα, πολύτροπον...

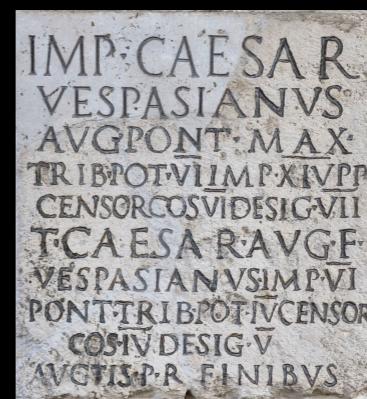
Odyssey, Homer, 700 BC



Virum mihi , Camena, insece versutum...



ancient Greek



Latin

2000 year old translation of
Sing in me, Muse, and through me tell the story of the man of twists and turns...

translating between human languages is an old problem

human language translation: nowadays

human language translation: nowadays



human language translation: nowadays



human language translation: nowadays



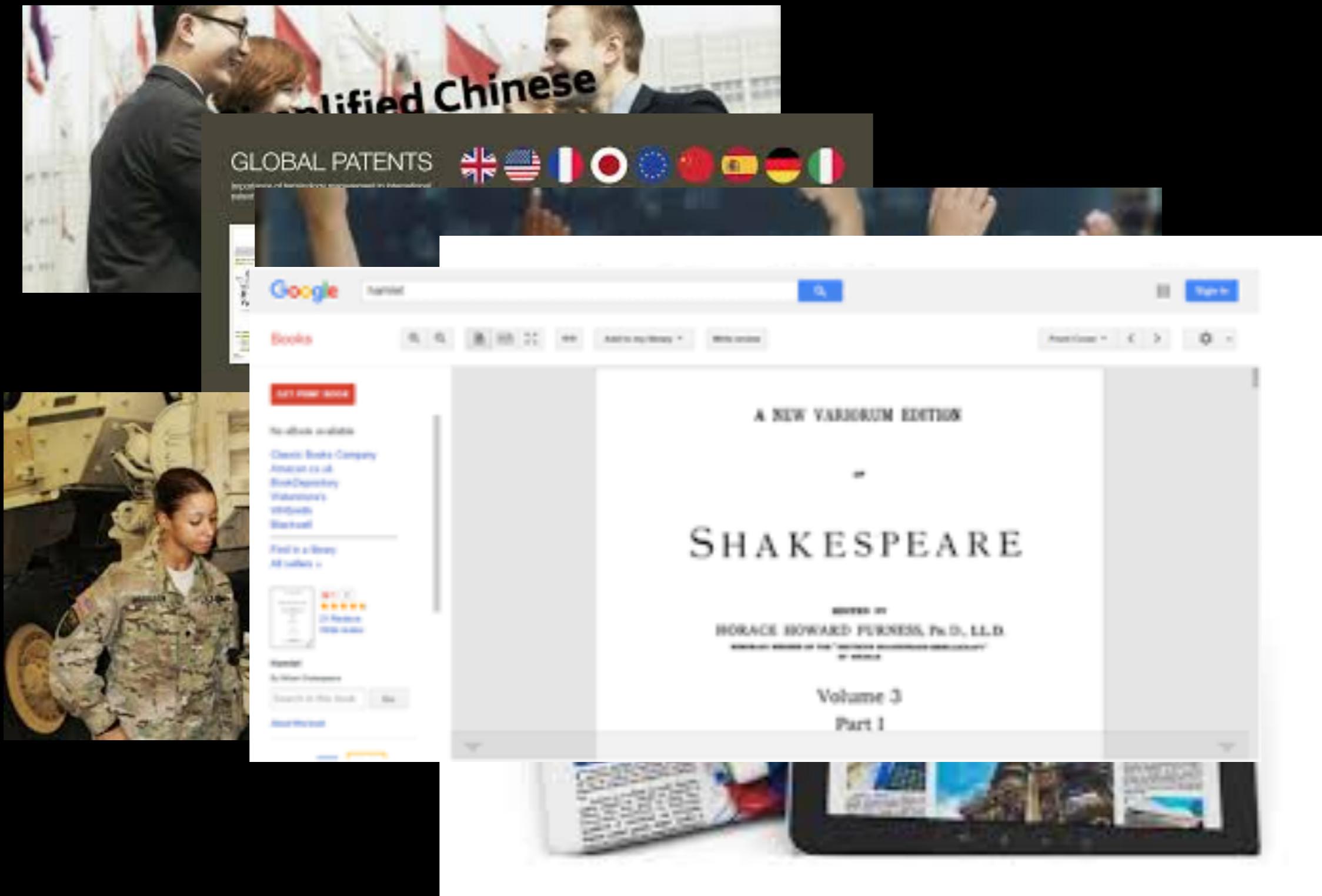
human language translation: nowadays



human language translation: nowadays



human language translation: nowadays



human language translation: nowadays



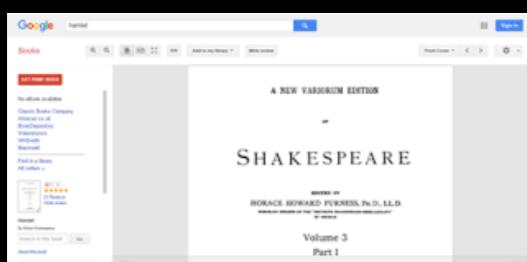
human language translation: nowadays



human language translation: nowadays



nowadays machine translation
is applied in society, science,
arts, commerce and finance,
literature, military, ...



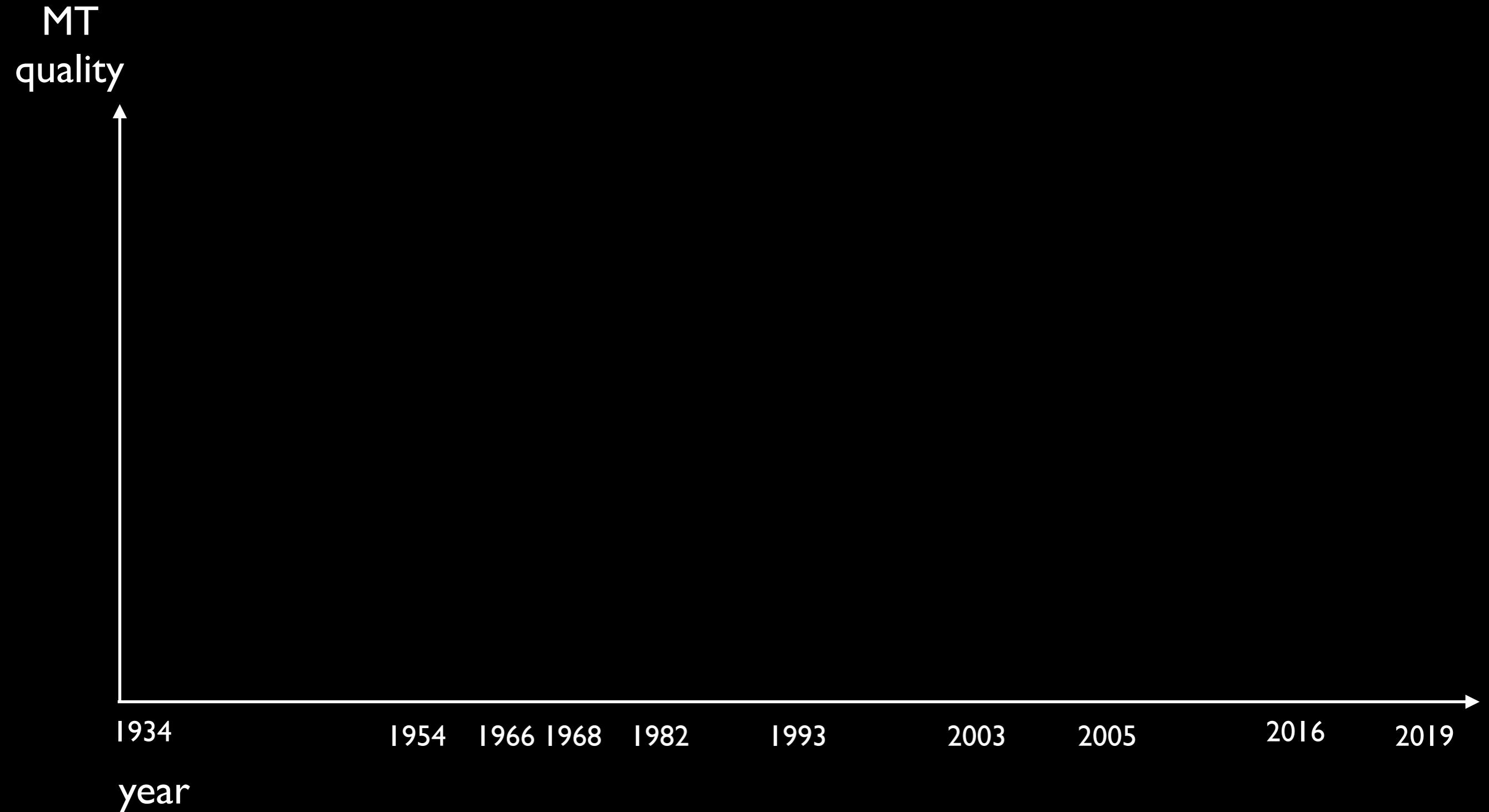
machine translation:

automatically translate from
one human language to another

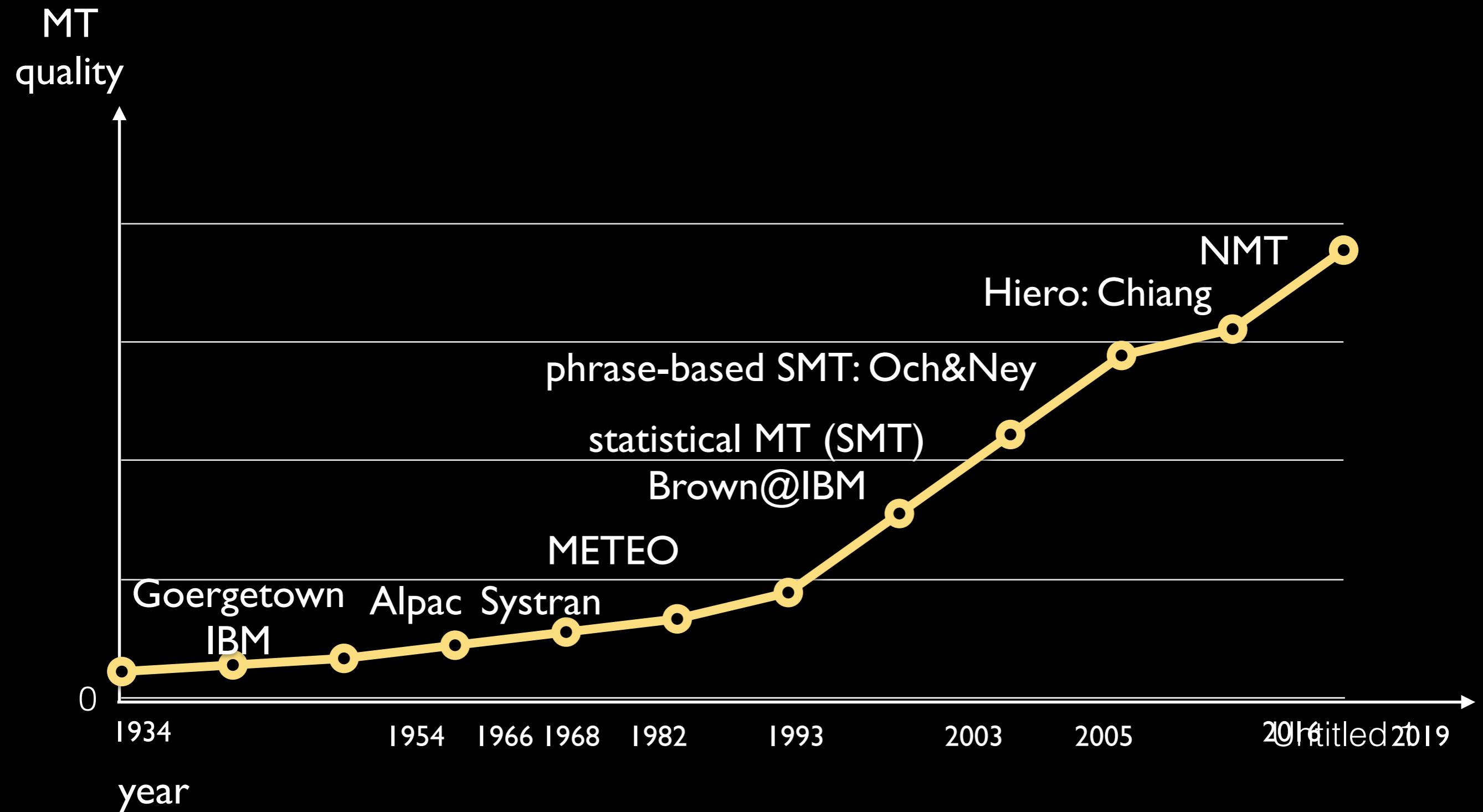
machine translation evaluations & resources

- competitions with data resources and baseline platforms, e.g.
 - NIST: 2002- [<https://www.nist.gov/itl/iad/mig/openmt15-evaluation>]
 - WMT: 2006- [<http://www.statmt.org>]
 - IWSLT: 2004- [<https://workshop2019.iwslt.org>]
- projects, e.g.
 - GALE, TC-Star, EuroMatrix, BOLT, and more and more
- datasets, e.g.
 - LDC: [<https://www.ldc.upenn.edu>]

machine translation development



machine translation development



machine translation history & challenges

1934

1954

1966 1968

1982

1993

2003

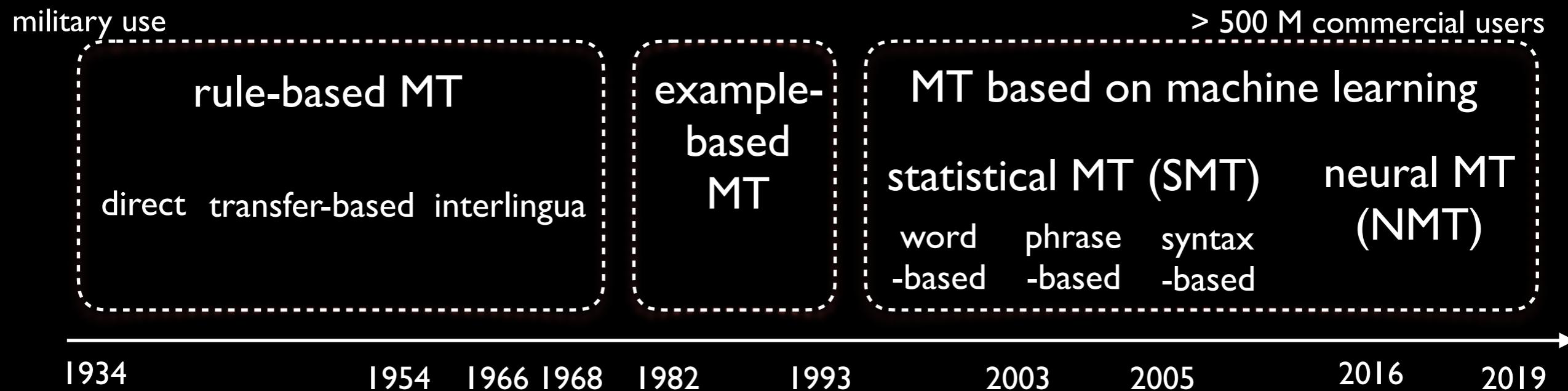
2005

2016

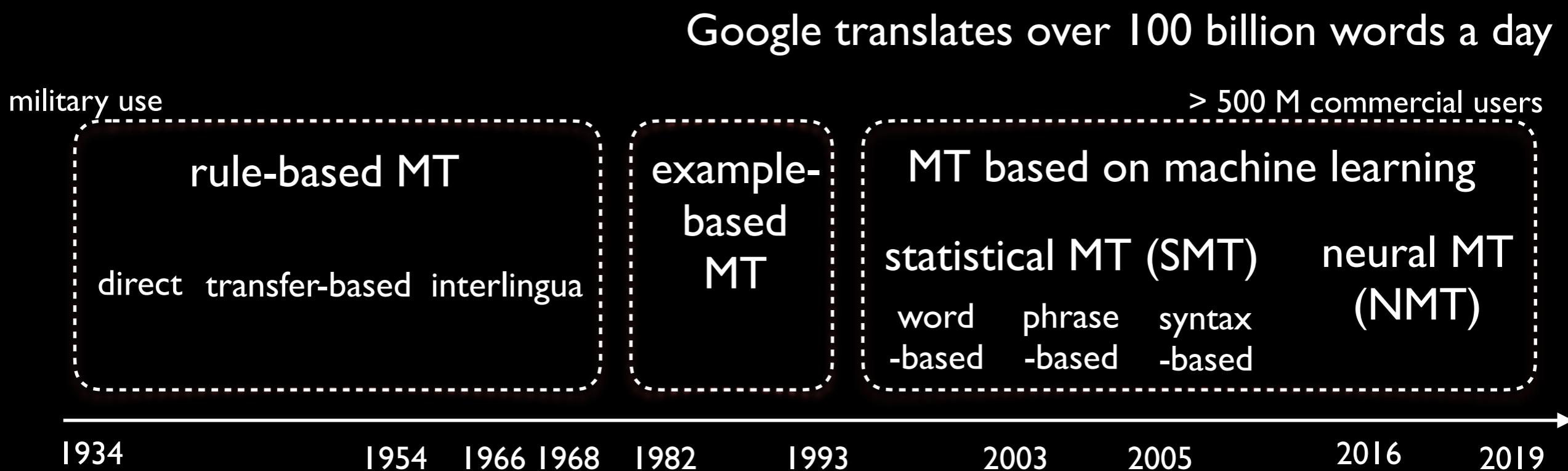
2019



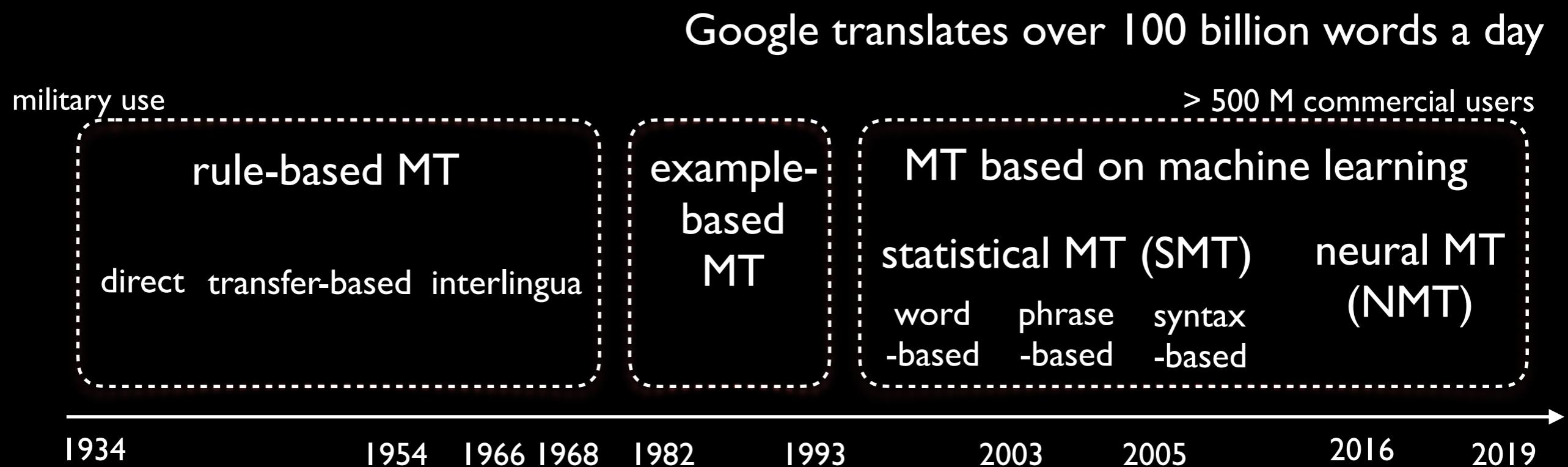
machine translation history & challenges



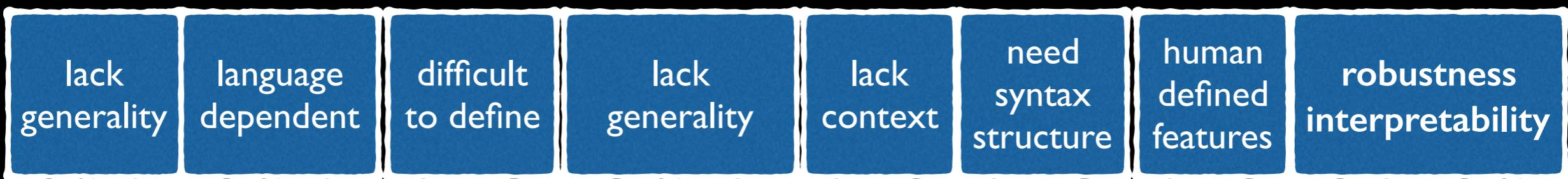
machine translation history & challenges



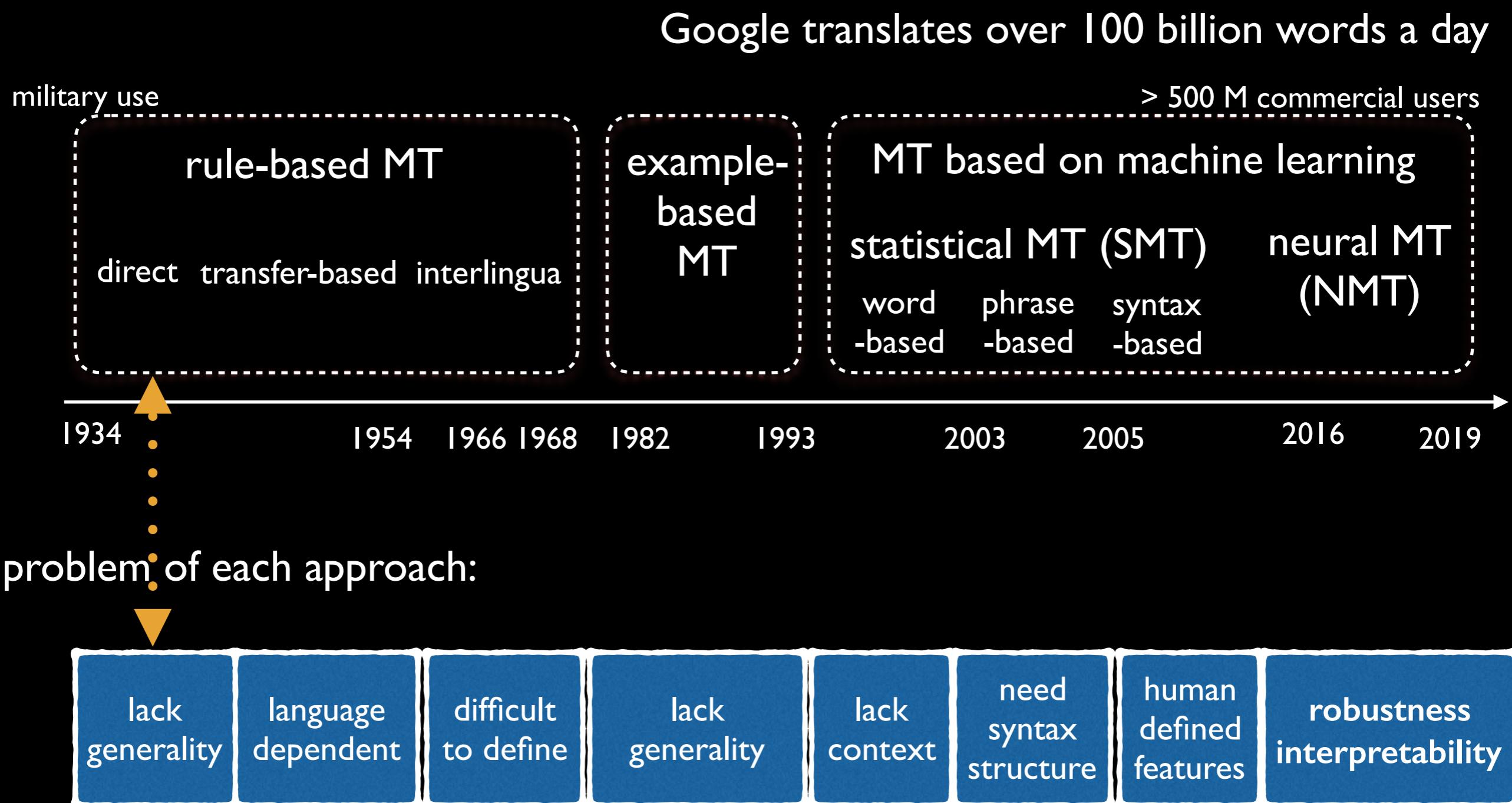
machine translation history & challenges



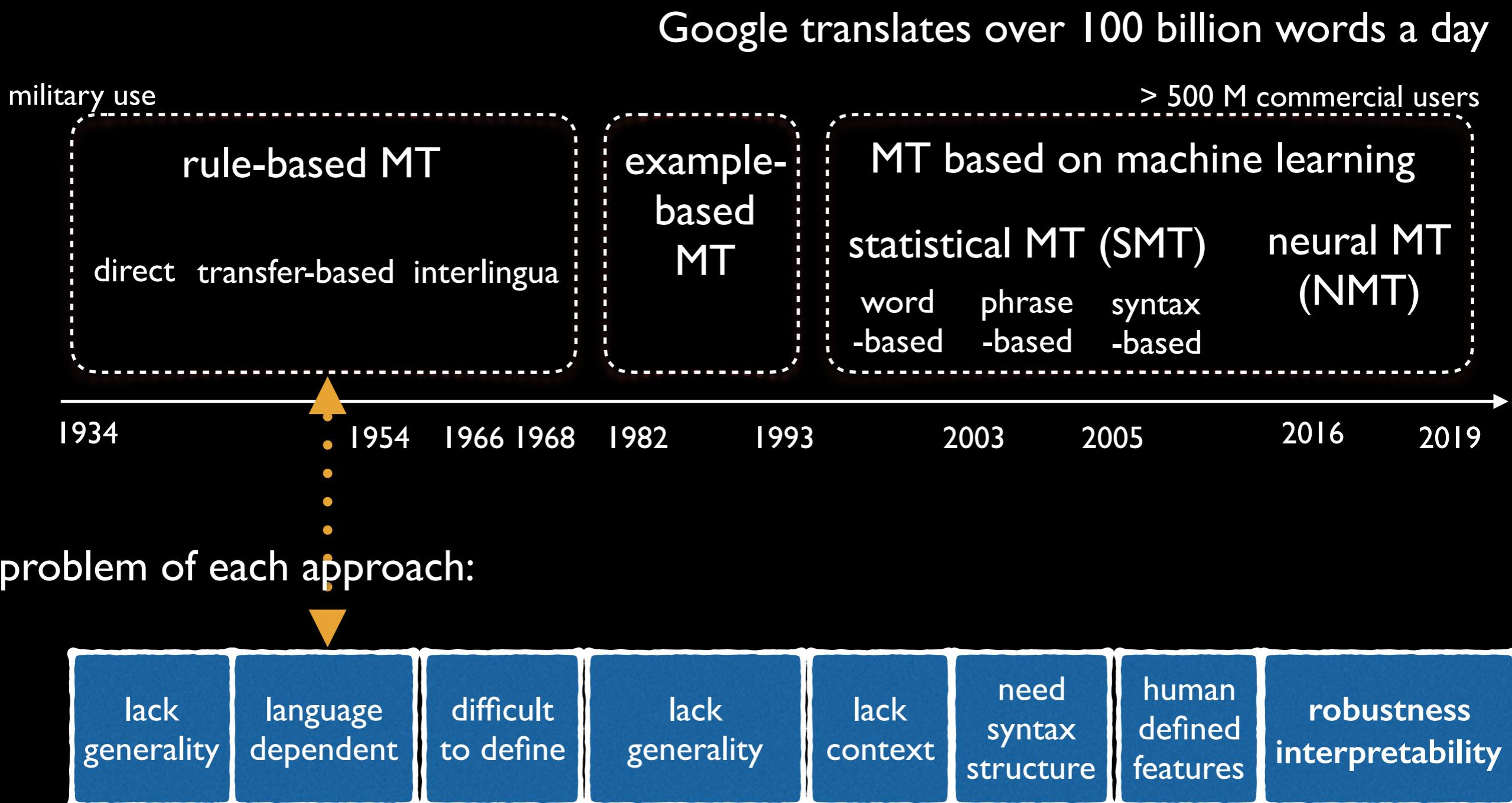
problem of each approach:



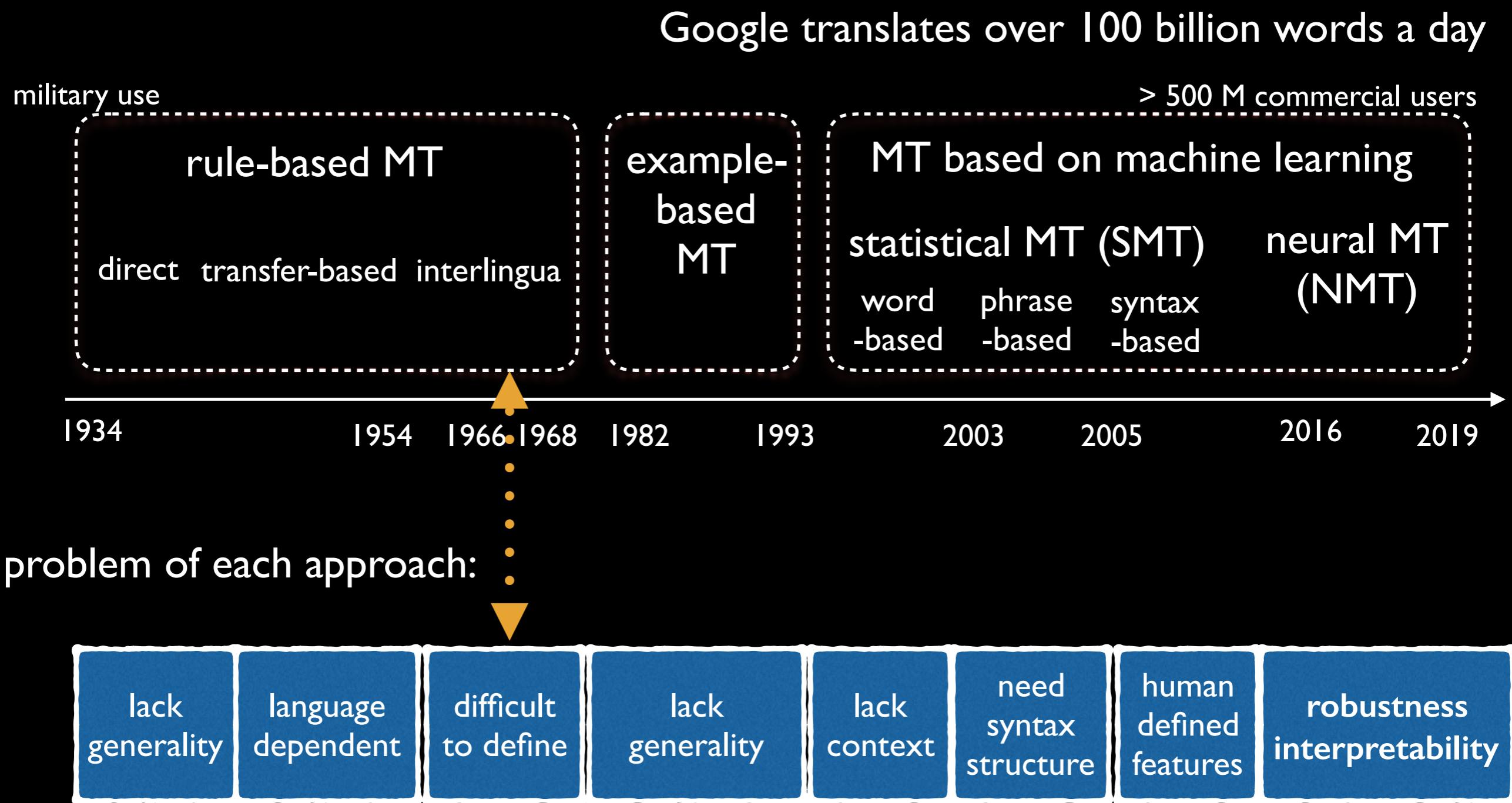
machine translation history & challenges



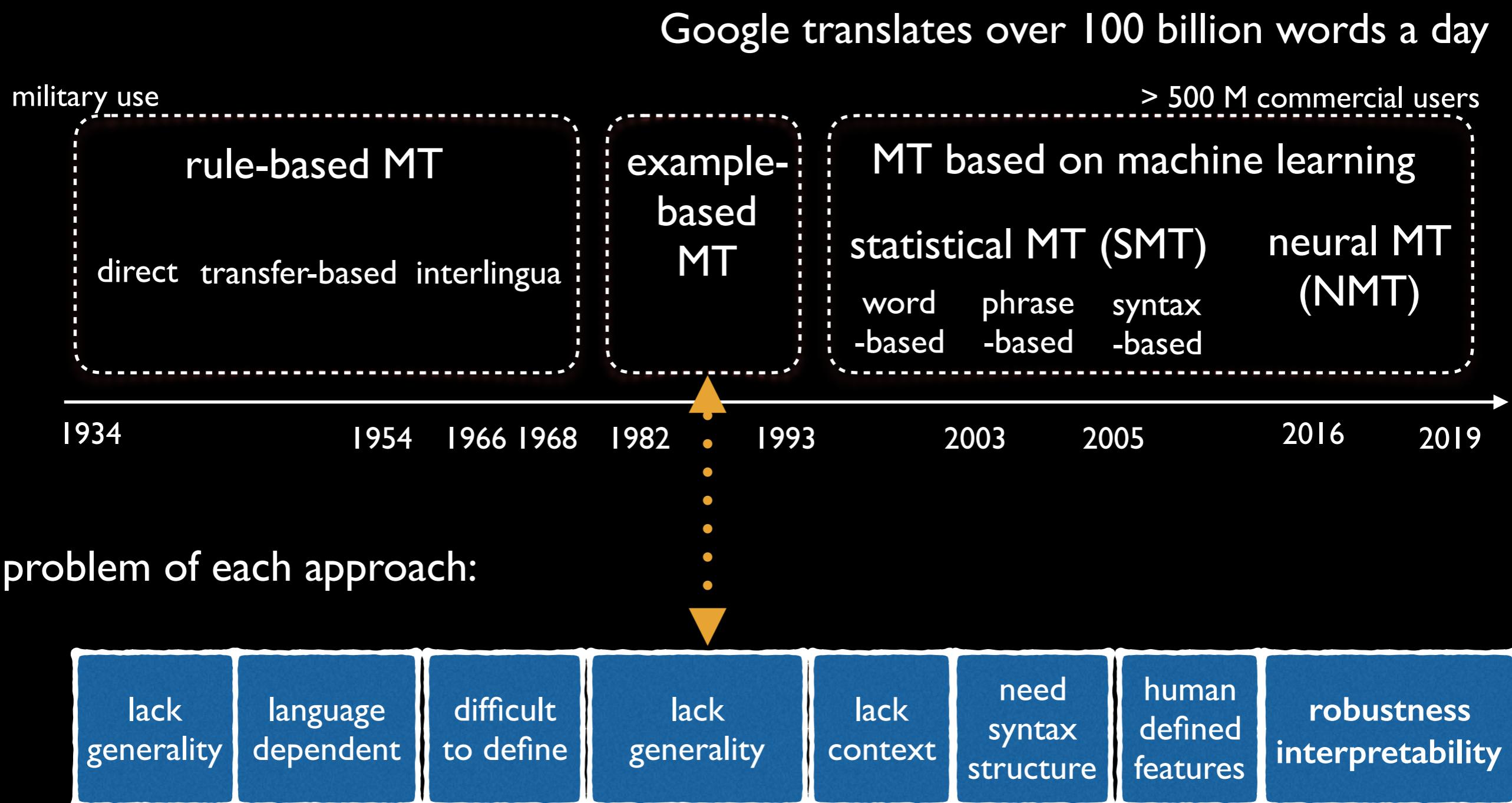
machine translation history & challenges



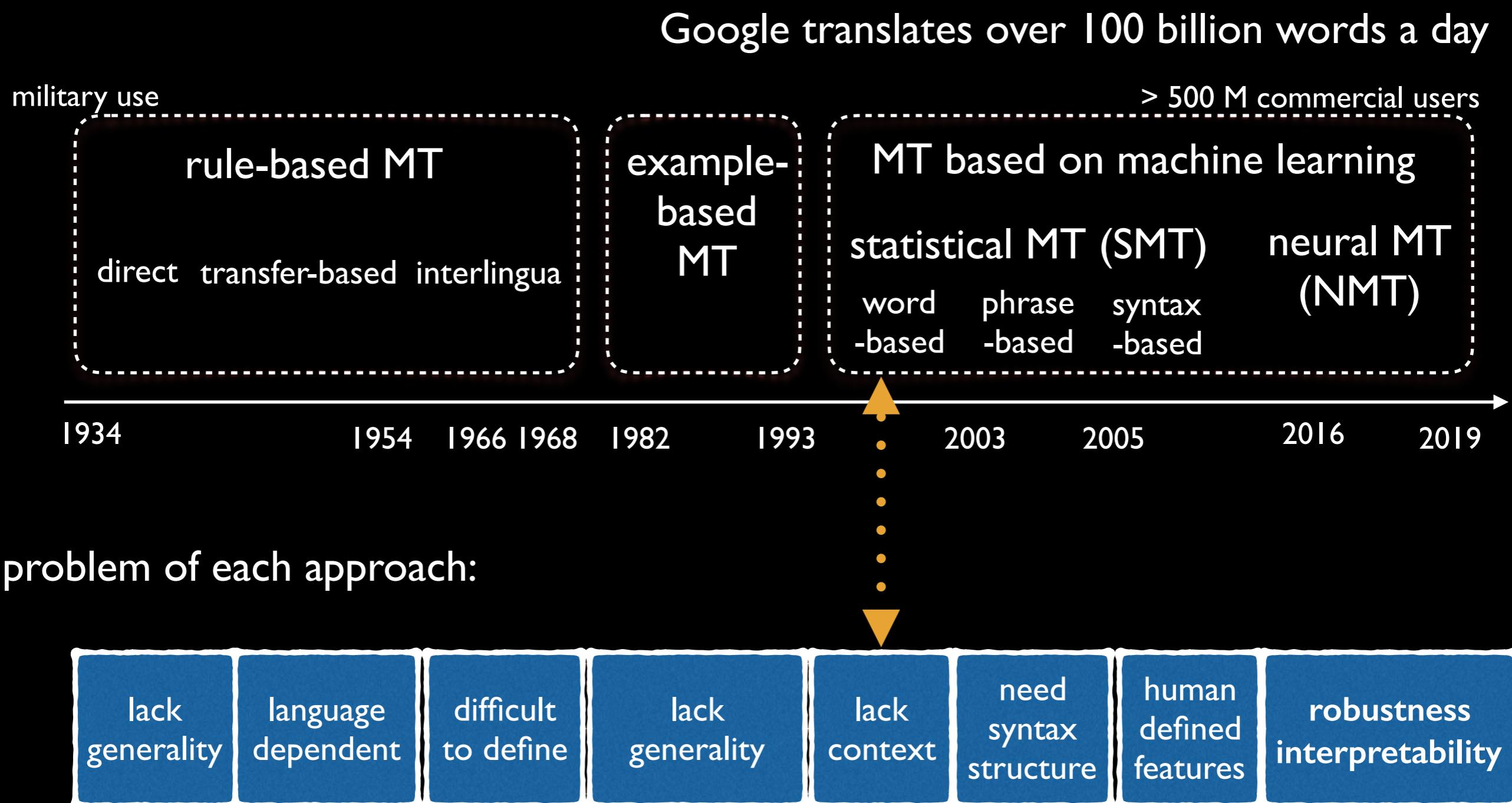
machine translation history & challenges



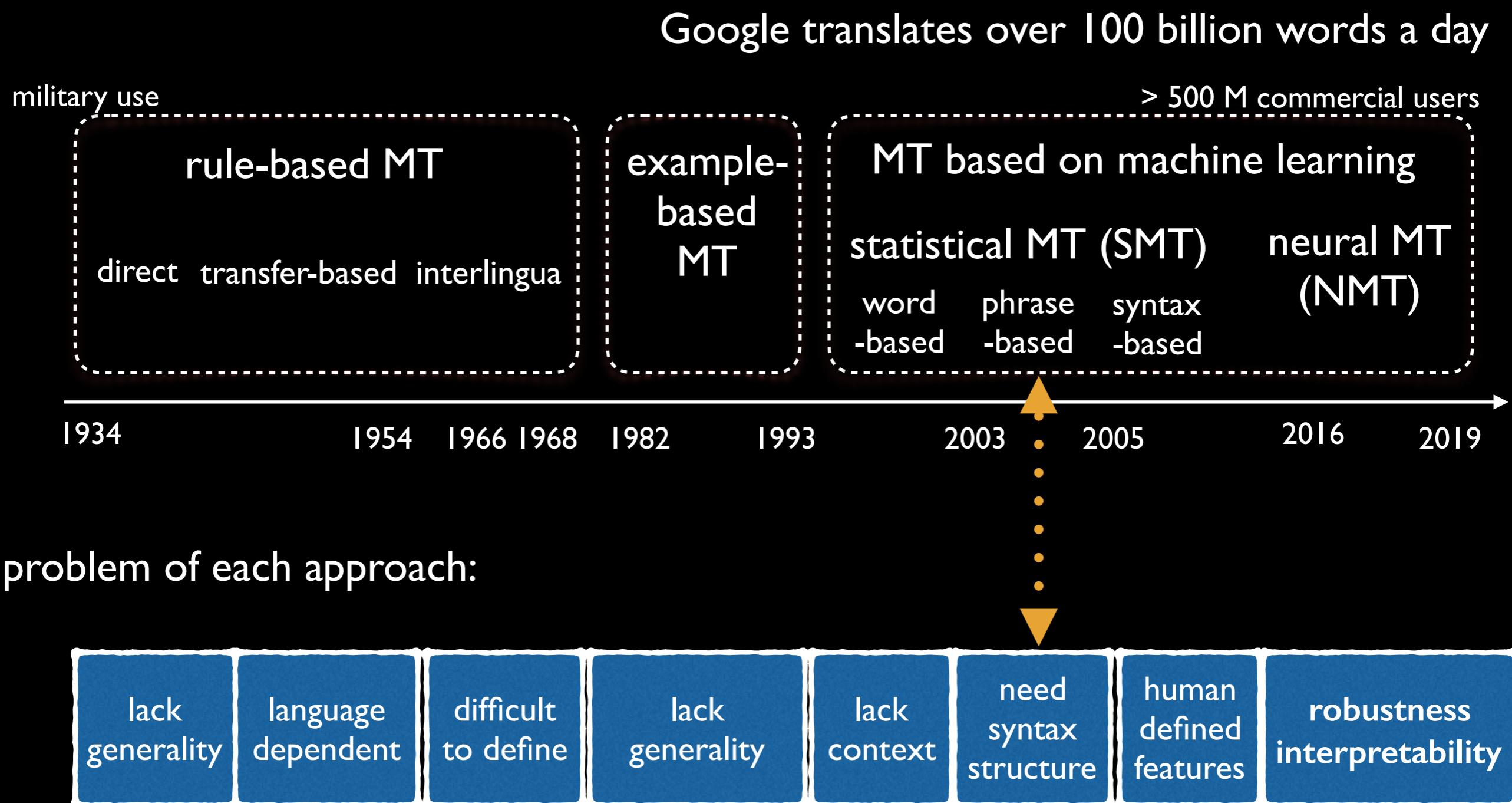
machine translation history & challenges



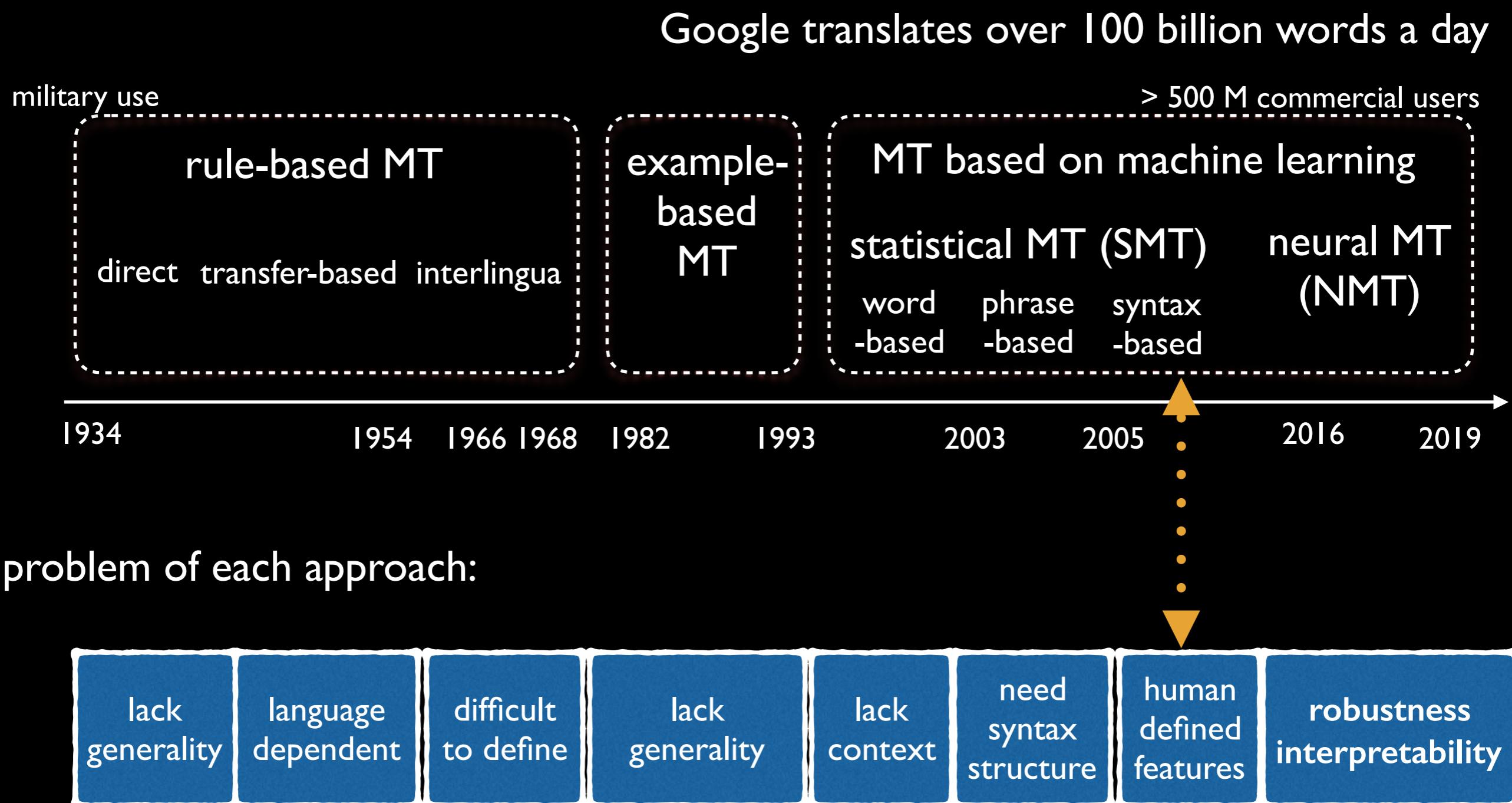
machine translation history & challenges



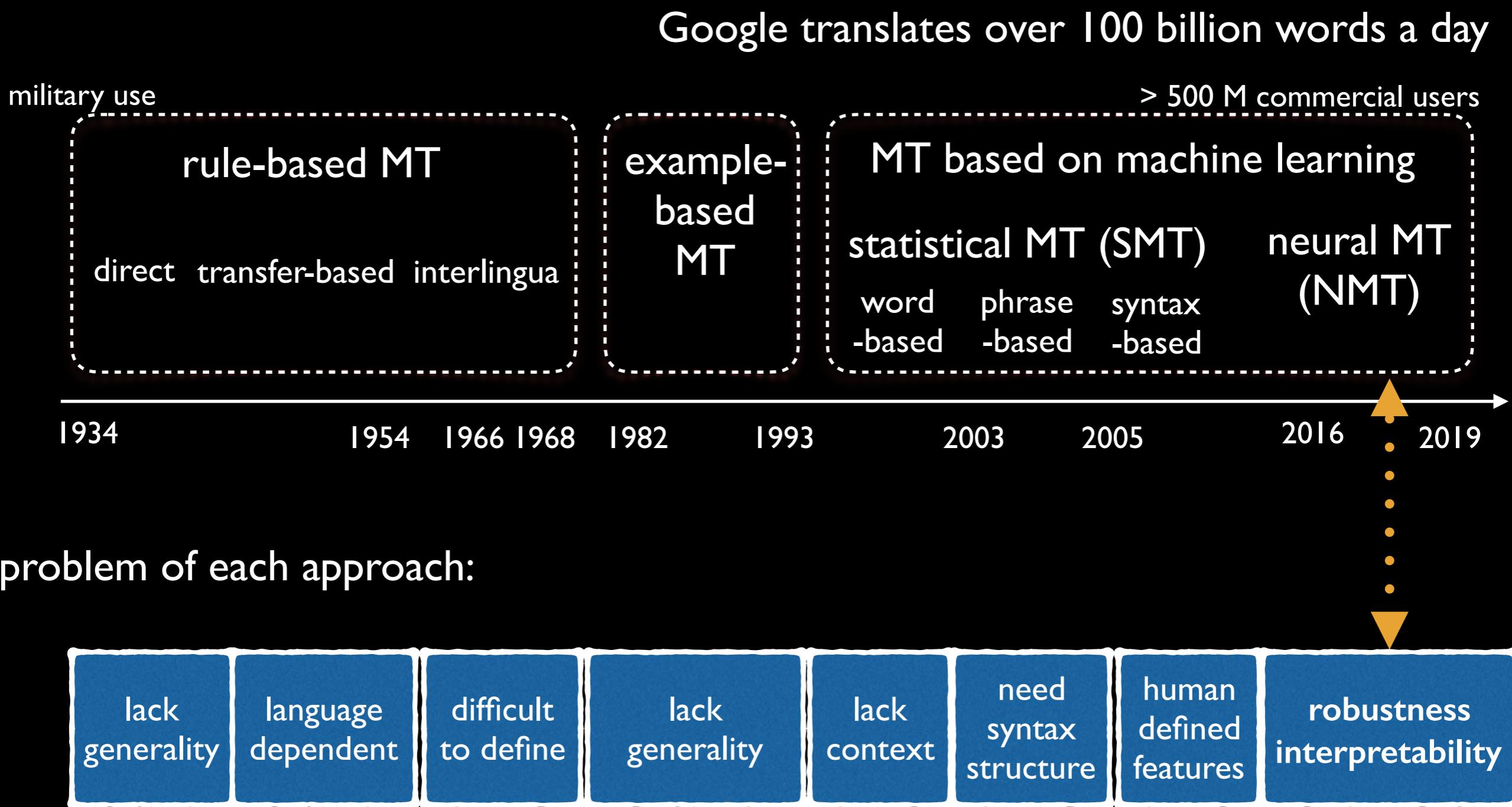
machine translation history & challenges



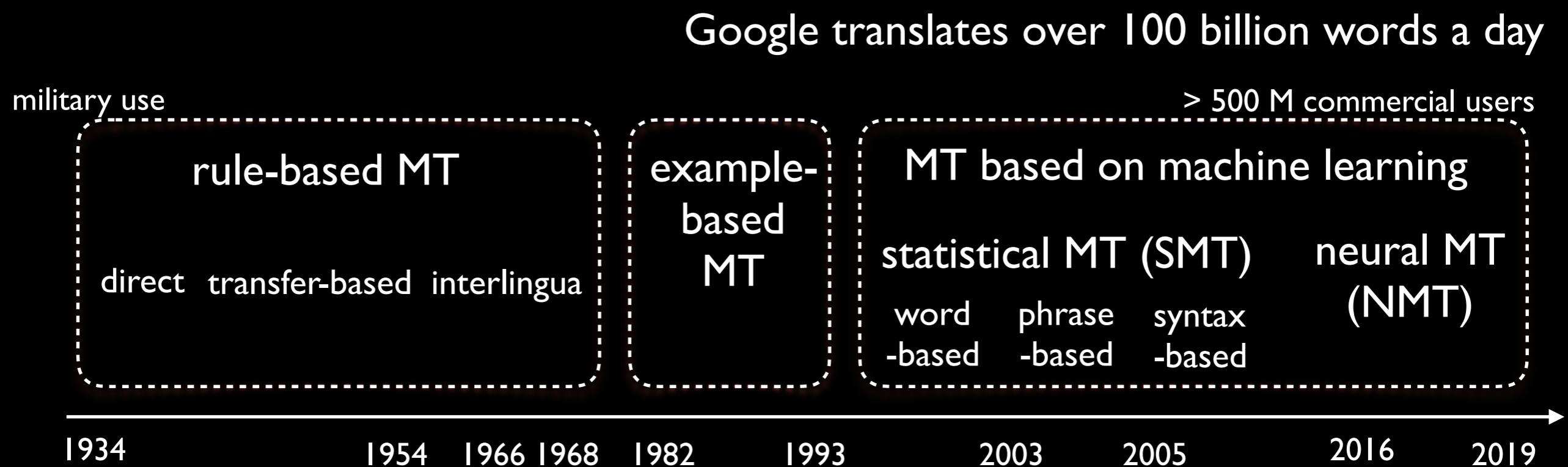
machine translation history & challenges



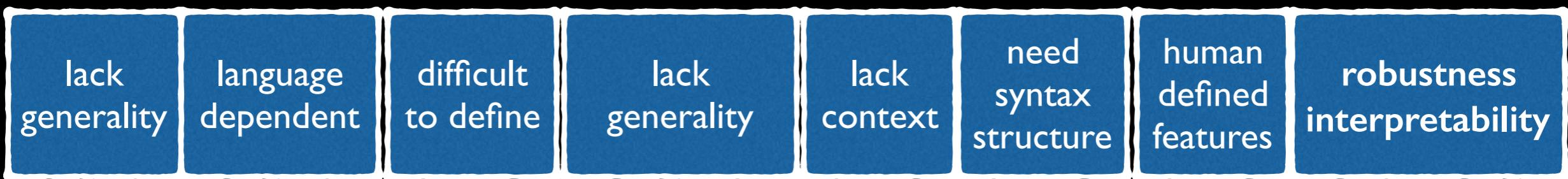
machine translation history & challenges



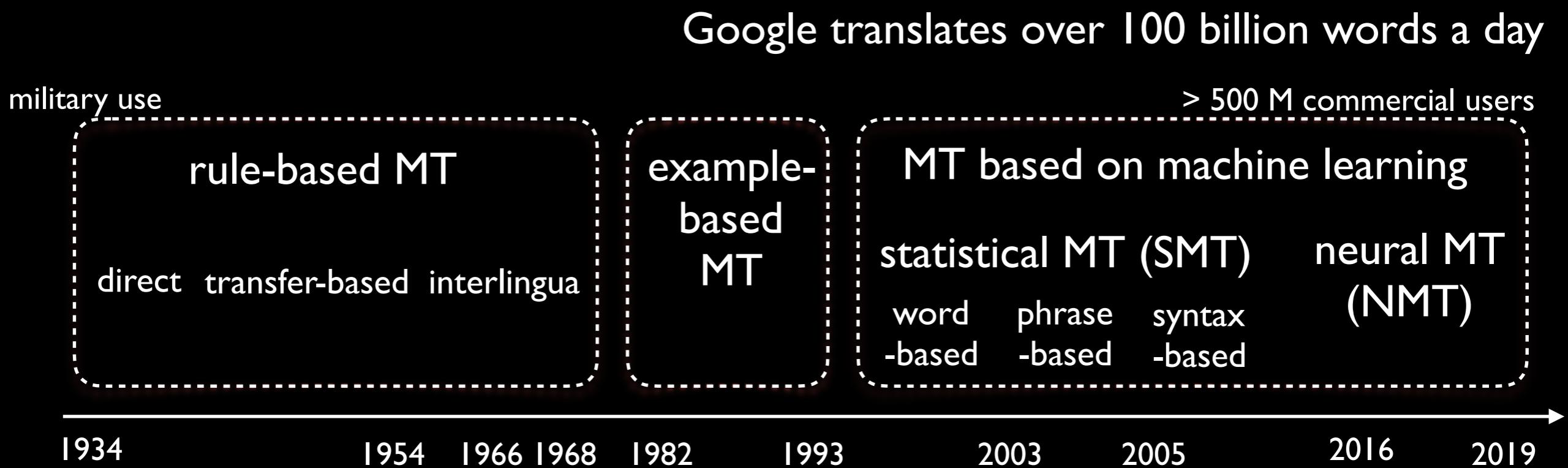
machine translation history & challenges



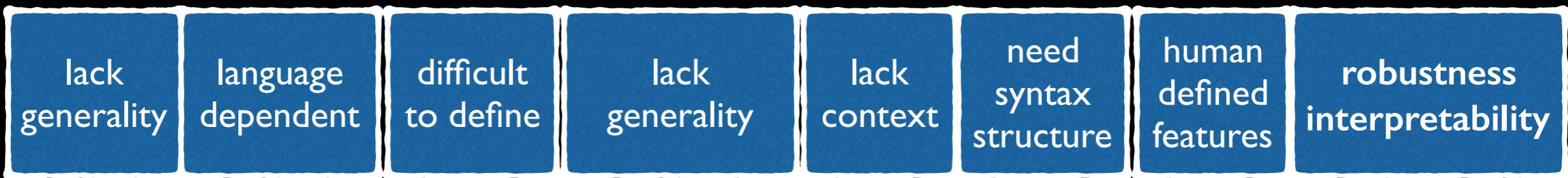
problem of each approach:



machine translation history & challenges

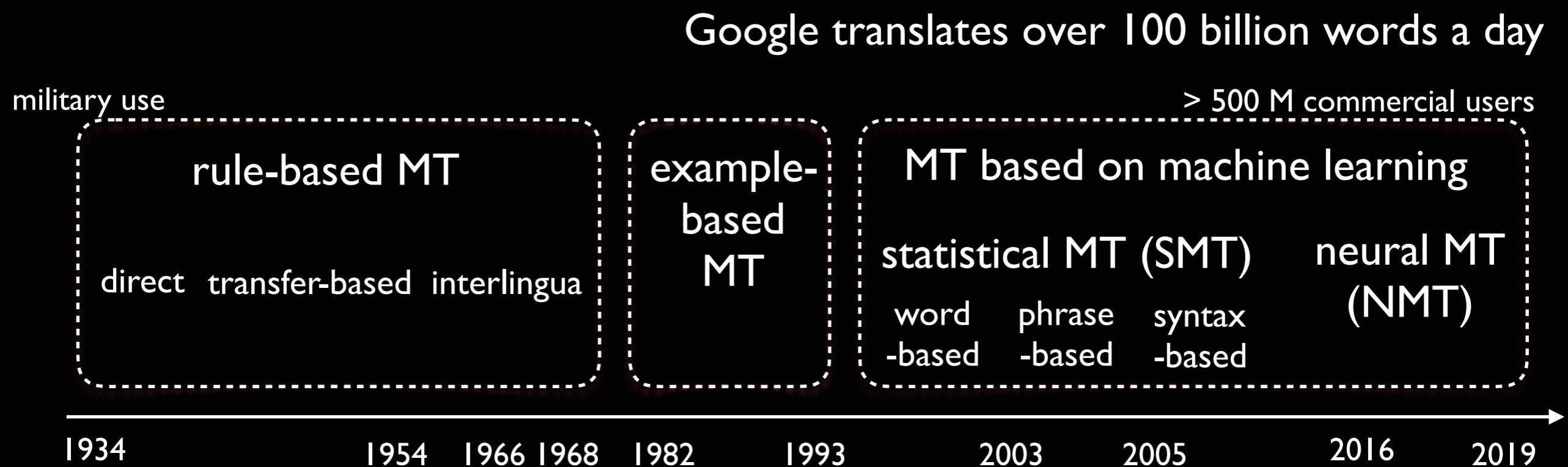


problem of each approach:

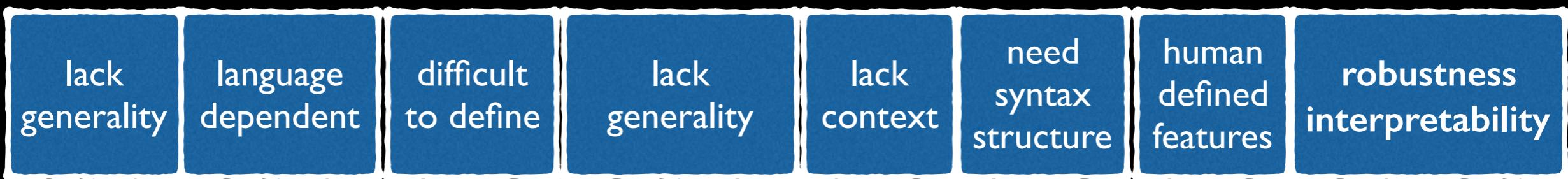


Question #1: how to enhance NMT robustness?

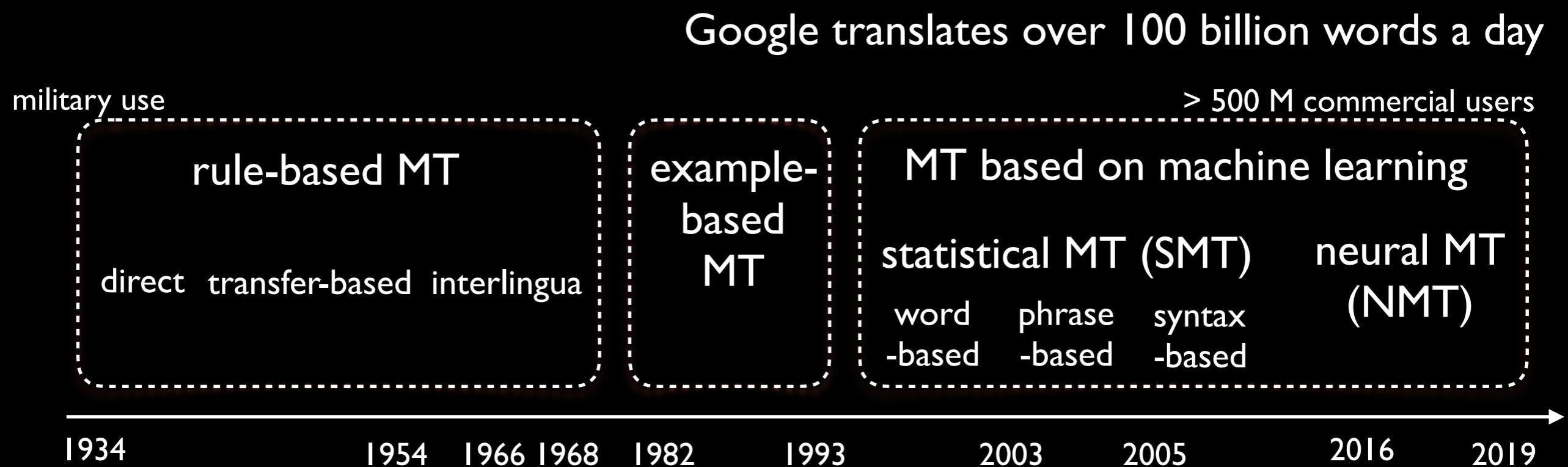
machine translation history & challenges



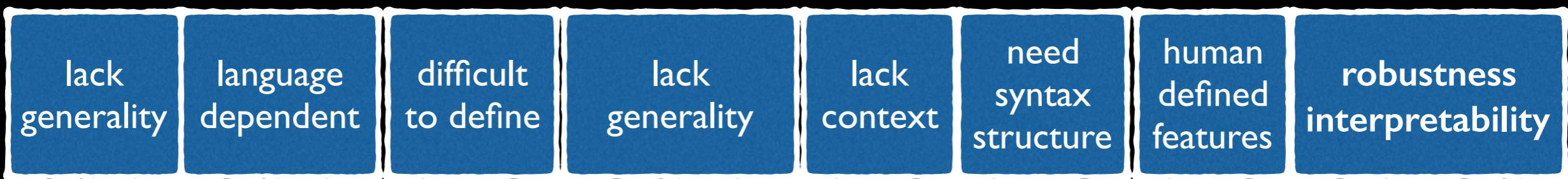
problem of each approach:



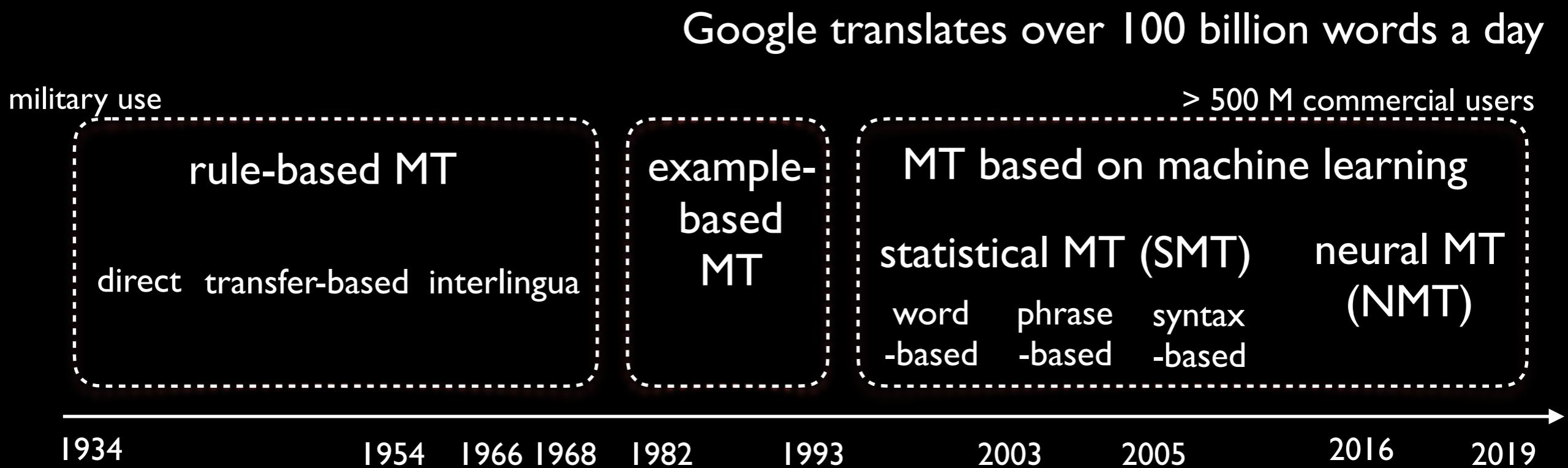
machine translation history & challenges



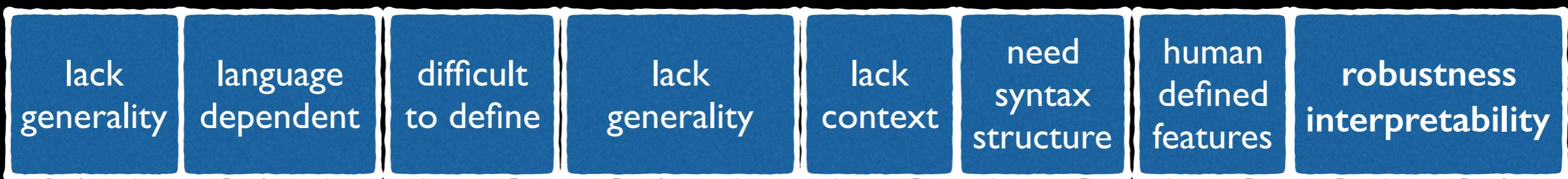
problem of each approach:



machine translation history & challenges

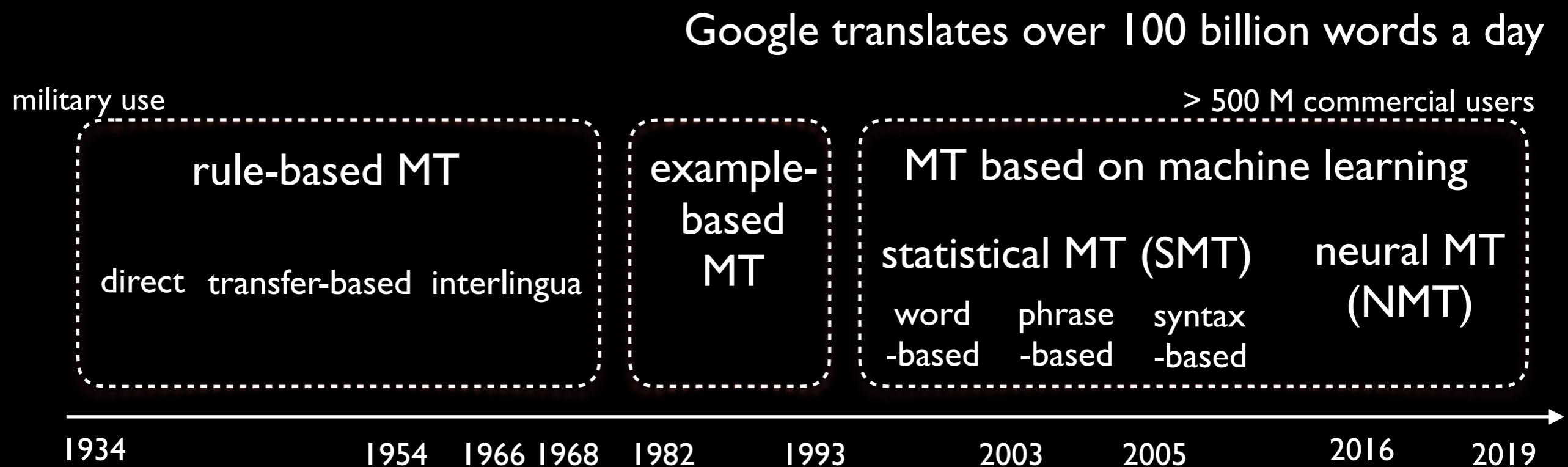


problem of each approach:

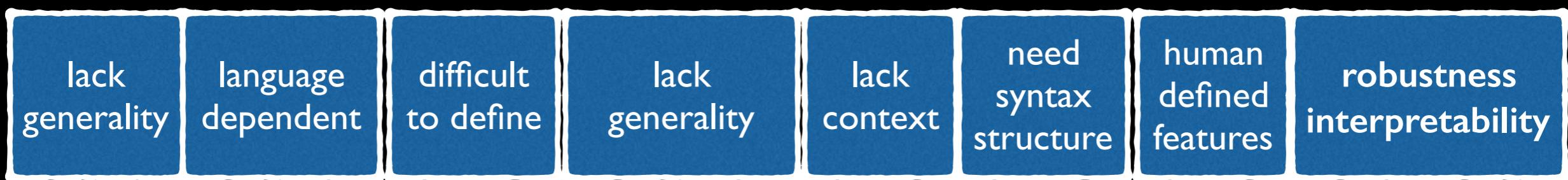


Question #2: how to increase interpretability?

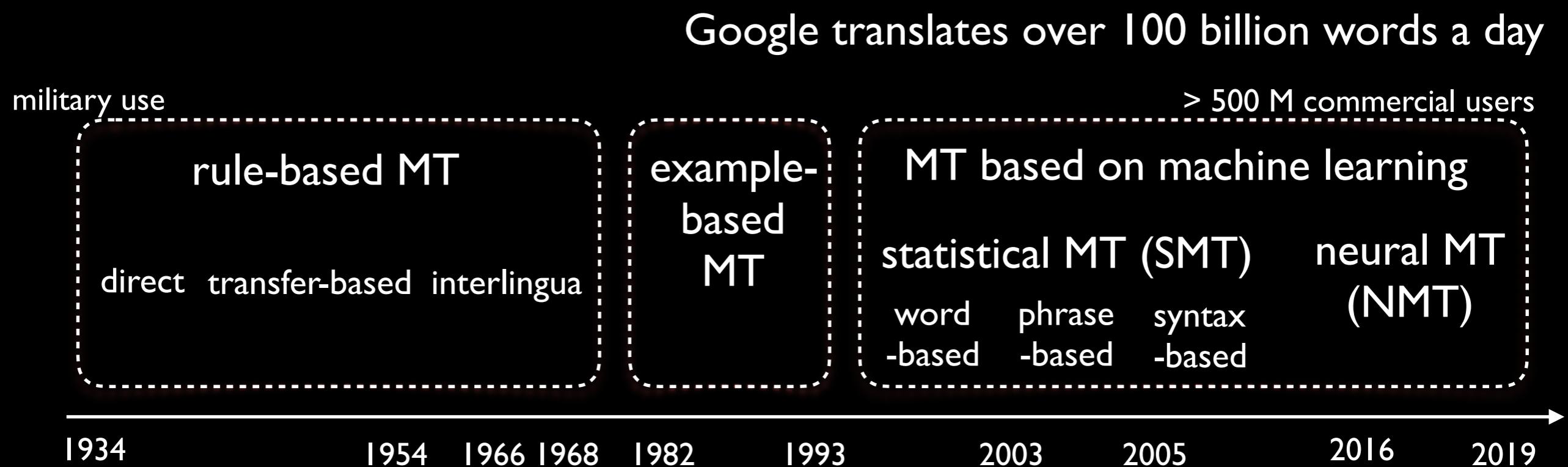
machine translation history & challenges



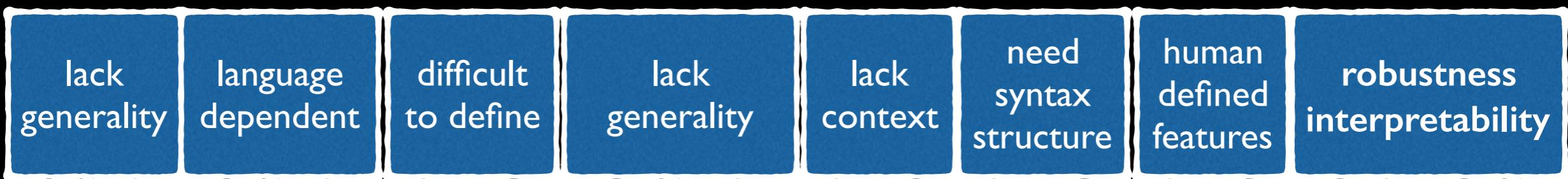
problem of each approach:



machine translation history & challenges



problem of each approach:



machine translation using machine learning

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
word phrase syntax (NMT)
-based -based -based

machine translation using machine learning

noisy channel model

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
word phrase syntax (NMT)
-based -based -based

machine translation using machine learning

noisy channel model

[Shannon, 1948] Information Theory

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
(NMT)

word phrase syntax
-based -based -based

machine translation using machine learning

noisy channel model

[Shannon, 1948] Information Theory

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
word phrase syntax
-based -based -based (NMT)

- input: source sentence (observation) f
- output: target sentence (decision) e
- Bayes decision rule

machine translation using machine learning

noisy channel model

[Shannon, 1948] Information Theory

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
word phrase syntax
-based -based -based
(NMT)

- input: source sentence (observation) f

- output: target sentence (decision) e

- Bayes decision rule

$$\hat{e} = \operatorname{argmax}_e \{Pr(e|f)\}$$

$$= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}$$

machine translation using machine learning

noisy channel model

[Shannon, 1948] Information Theory

> 500 M commercial users

MT based on machine learning

statistical MT (SMT) neural MT
word phrase syntax
-based -based -based
(NMT)

- input: source sentence (observation) f

- output: target sentence (decision) e

- Bayes decision rule

$$\hat{e} = \operatorname{argmax}_e \{Pr(e|f)\}$$

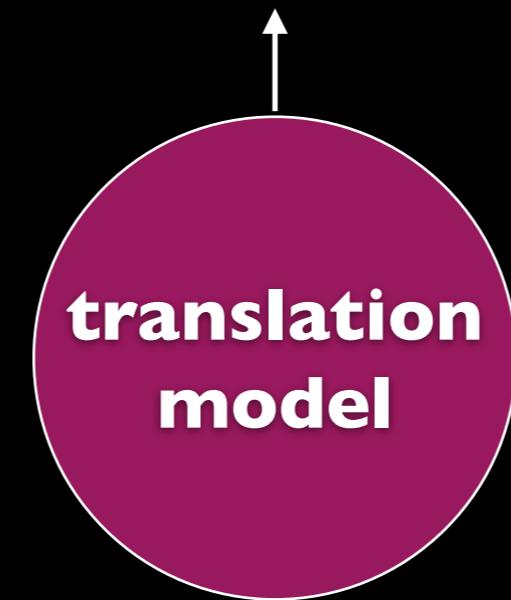
$$= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}$$

machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

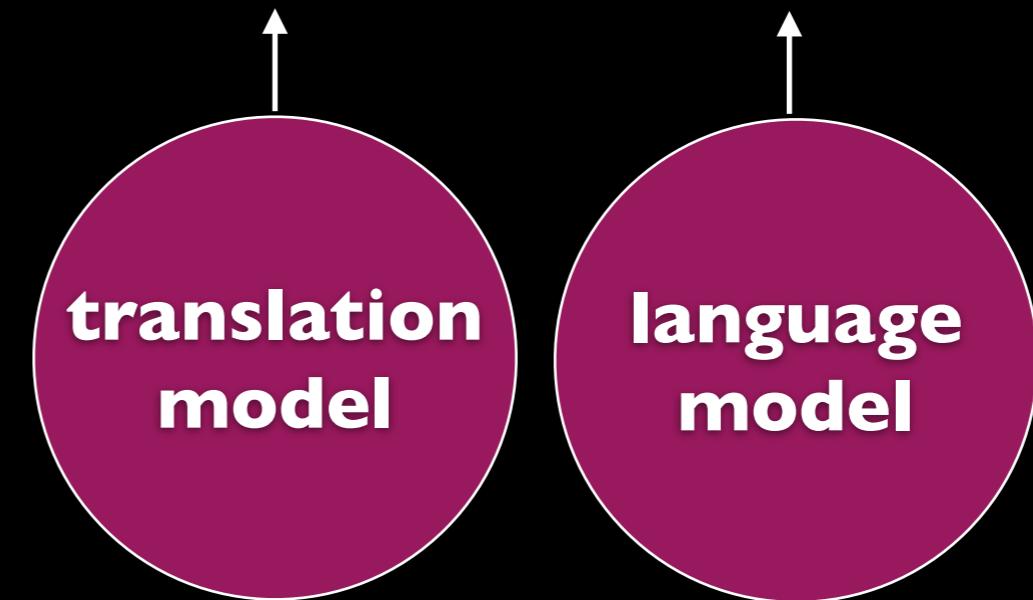
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



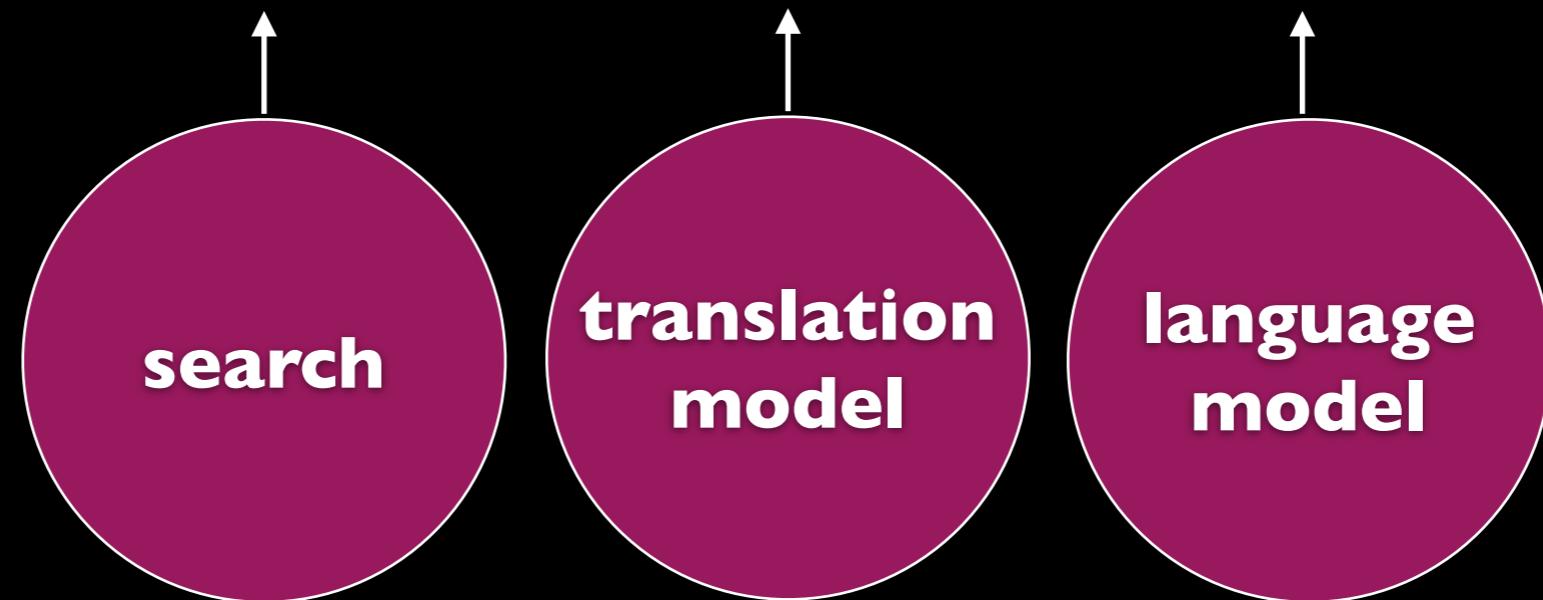
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



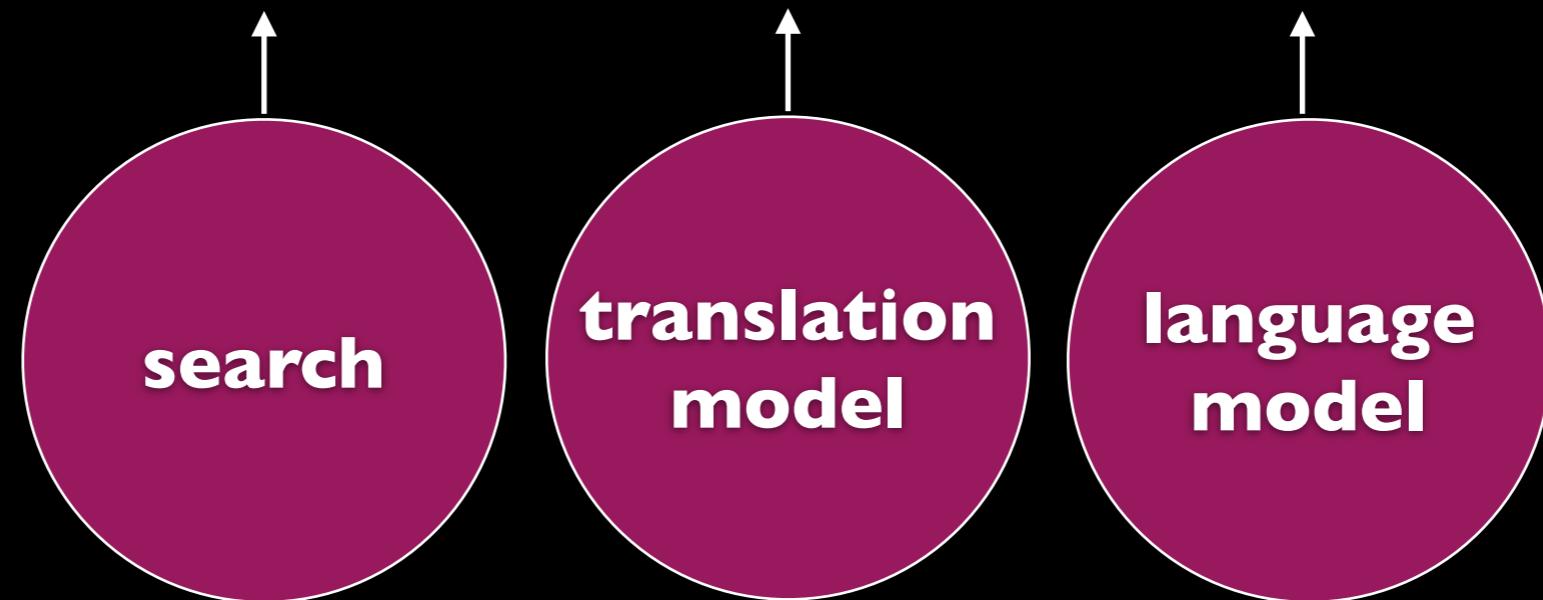
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



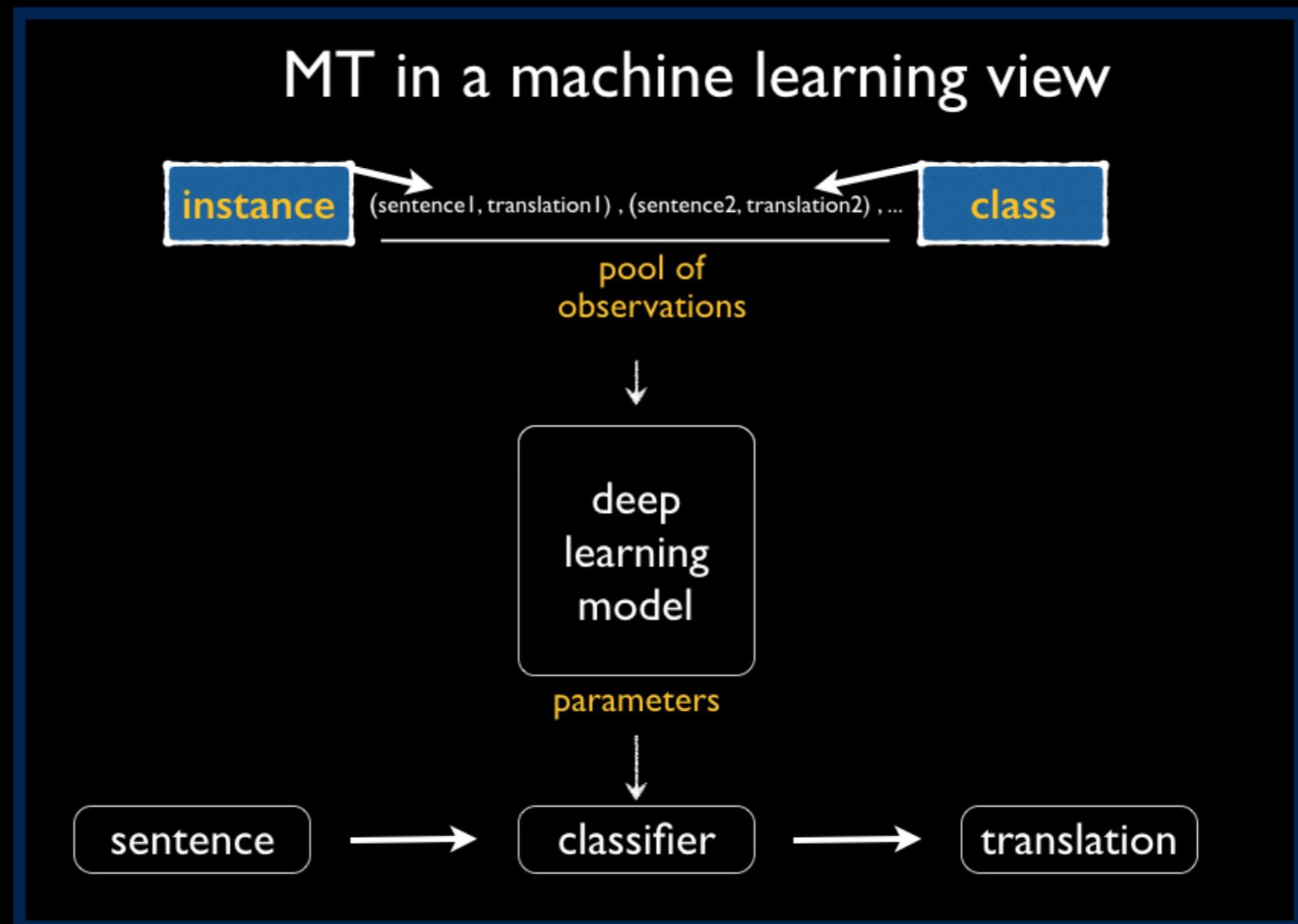
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



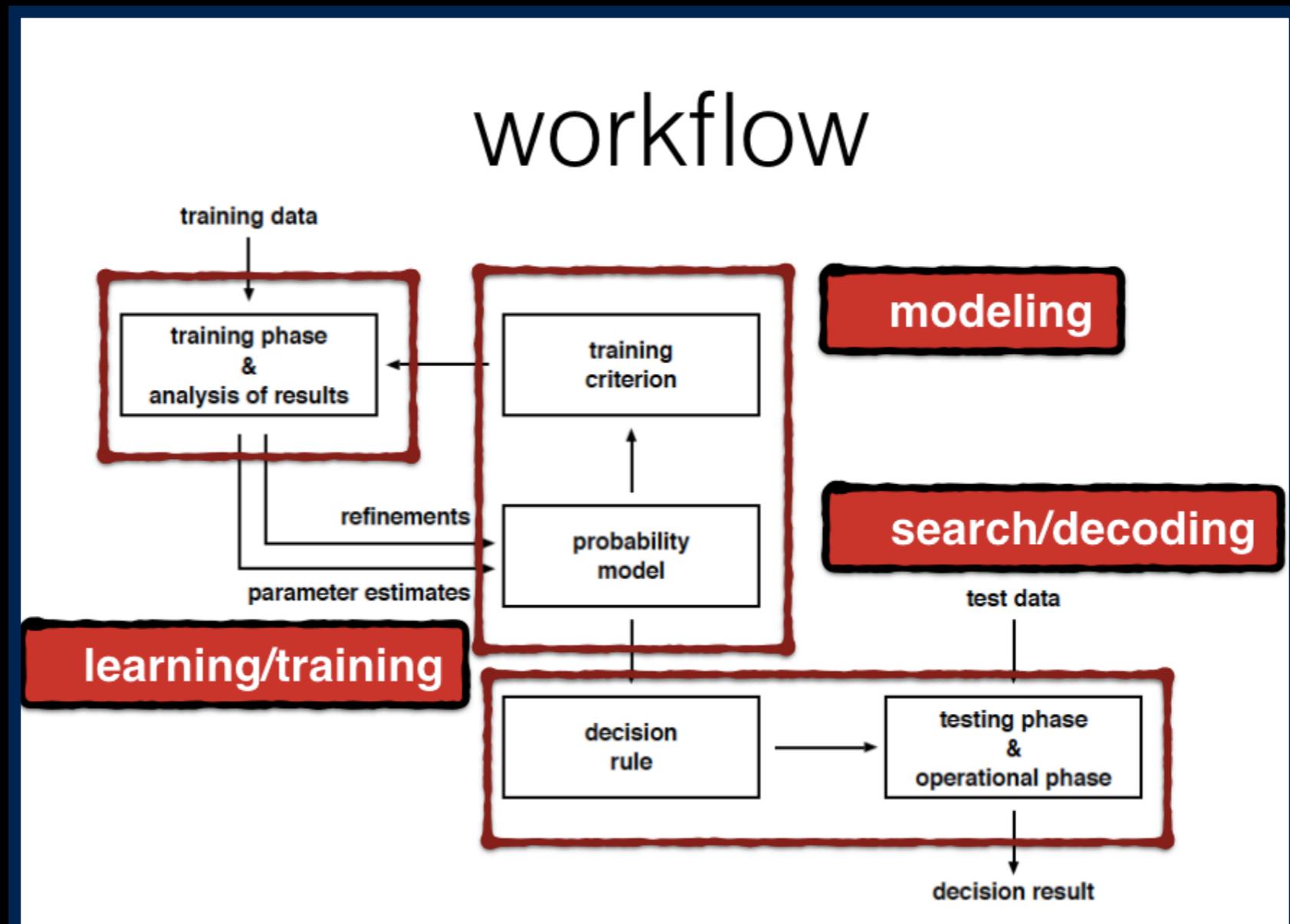
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



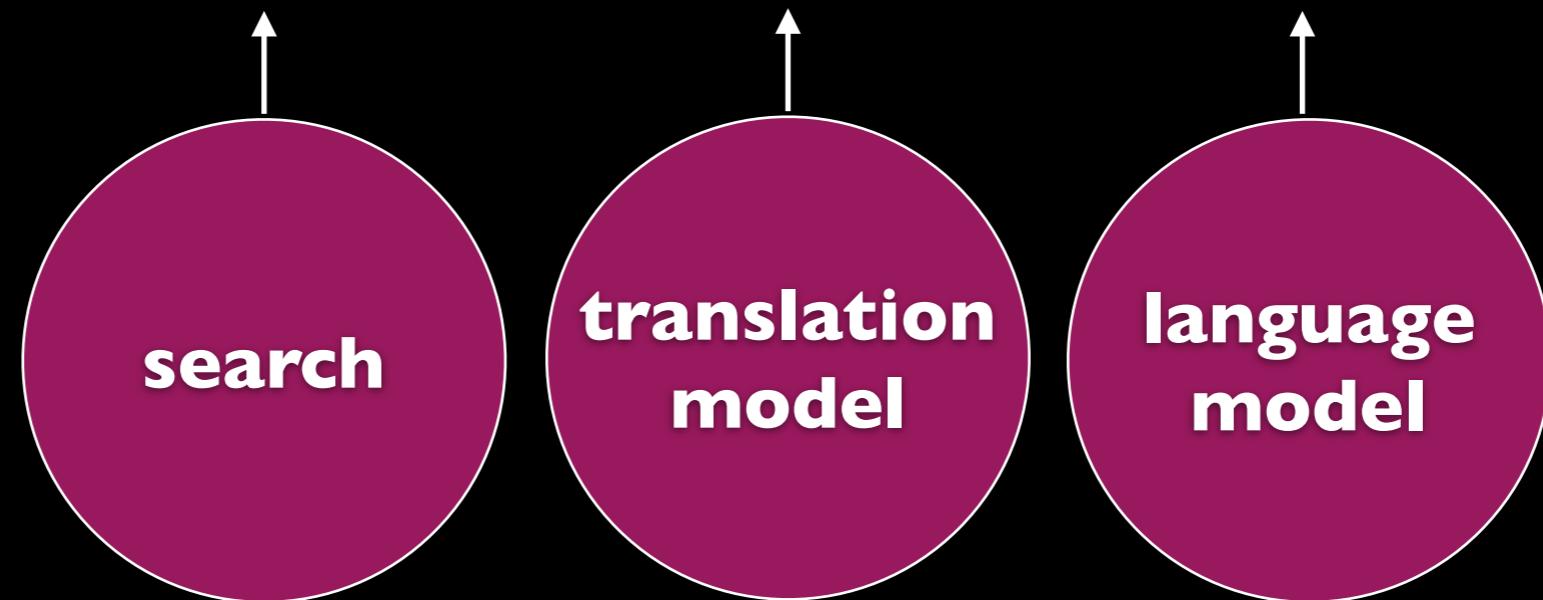
machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



machine translation using machine learning

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



machine translation

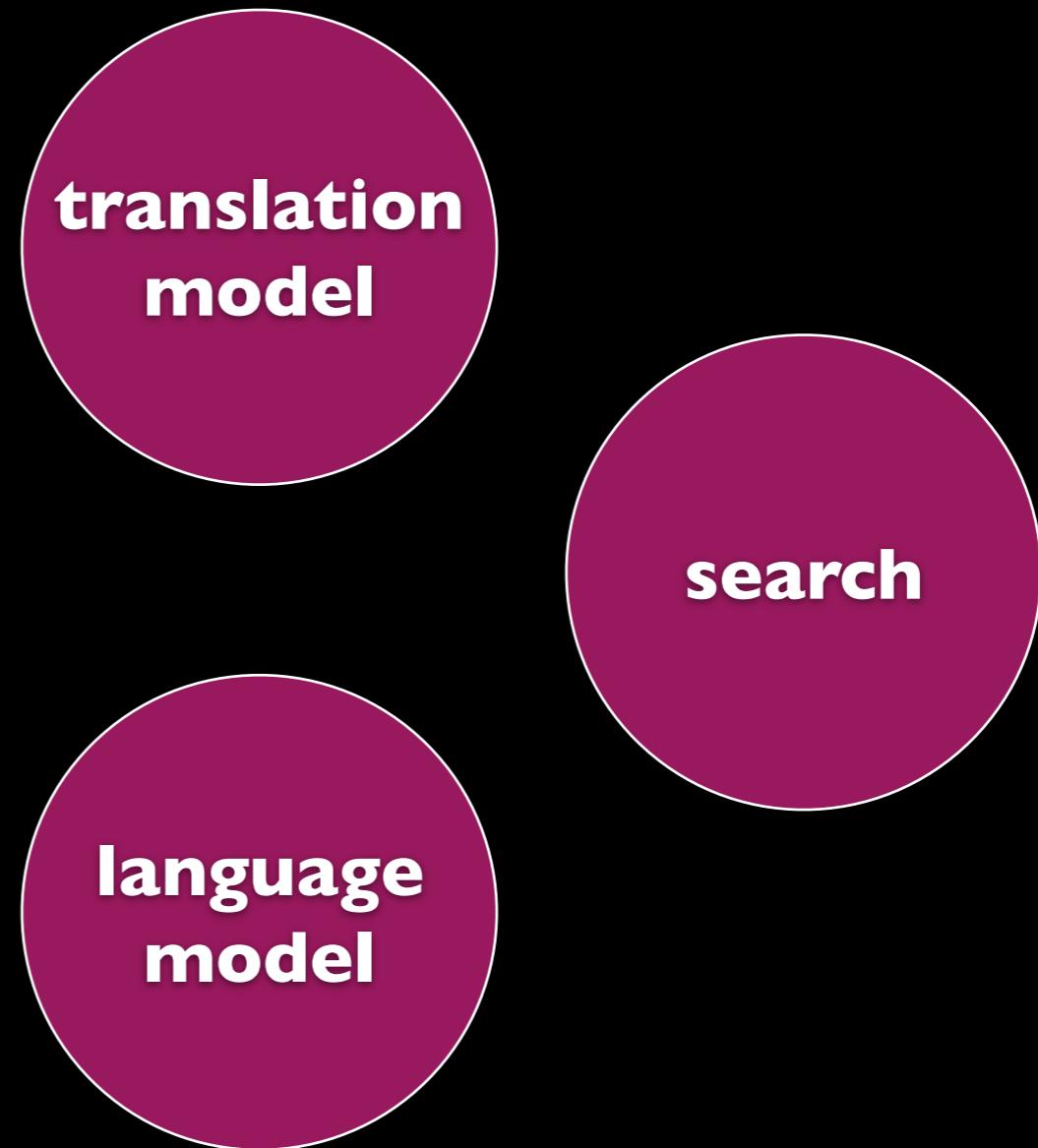
A diagram illustrating the components of machine translation. At the top center is the title "machine translation". Below it are three circular nodes arranged in a triangle. The top-left node is labeled "translation model", the top-right node is labeled "search", and the bottom node is labeled "language model". All nodes are dark red circles with white text.

**translation
model**

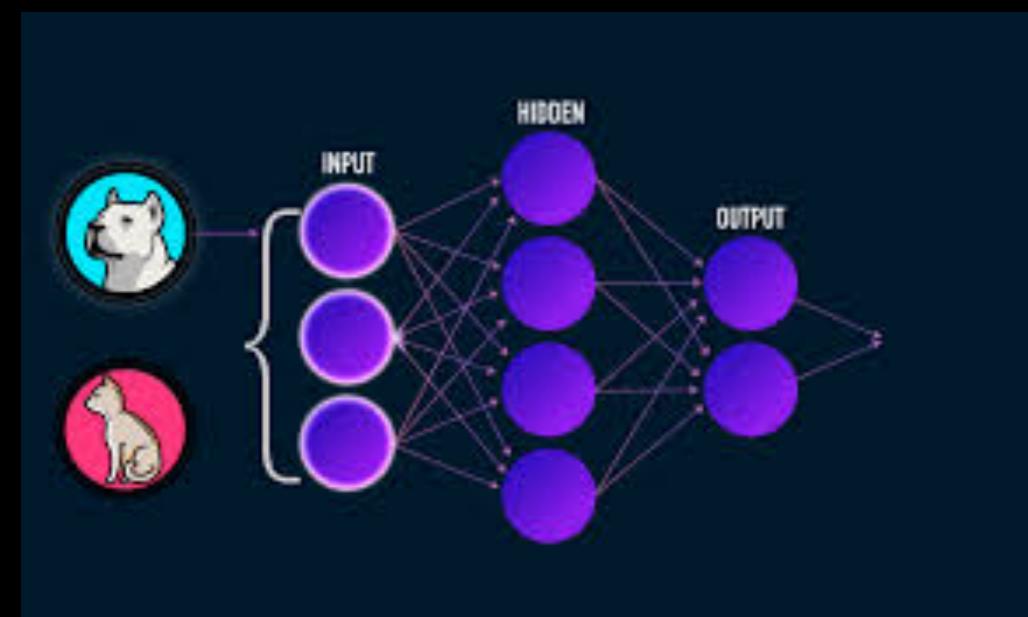
search

**language
model**

neural machine translation



neural machine translation



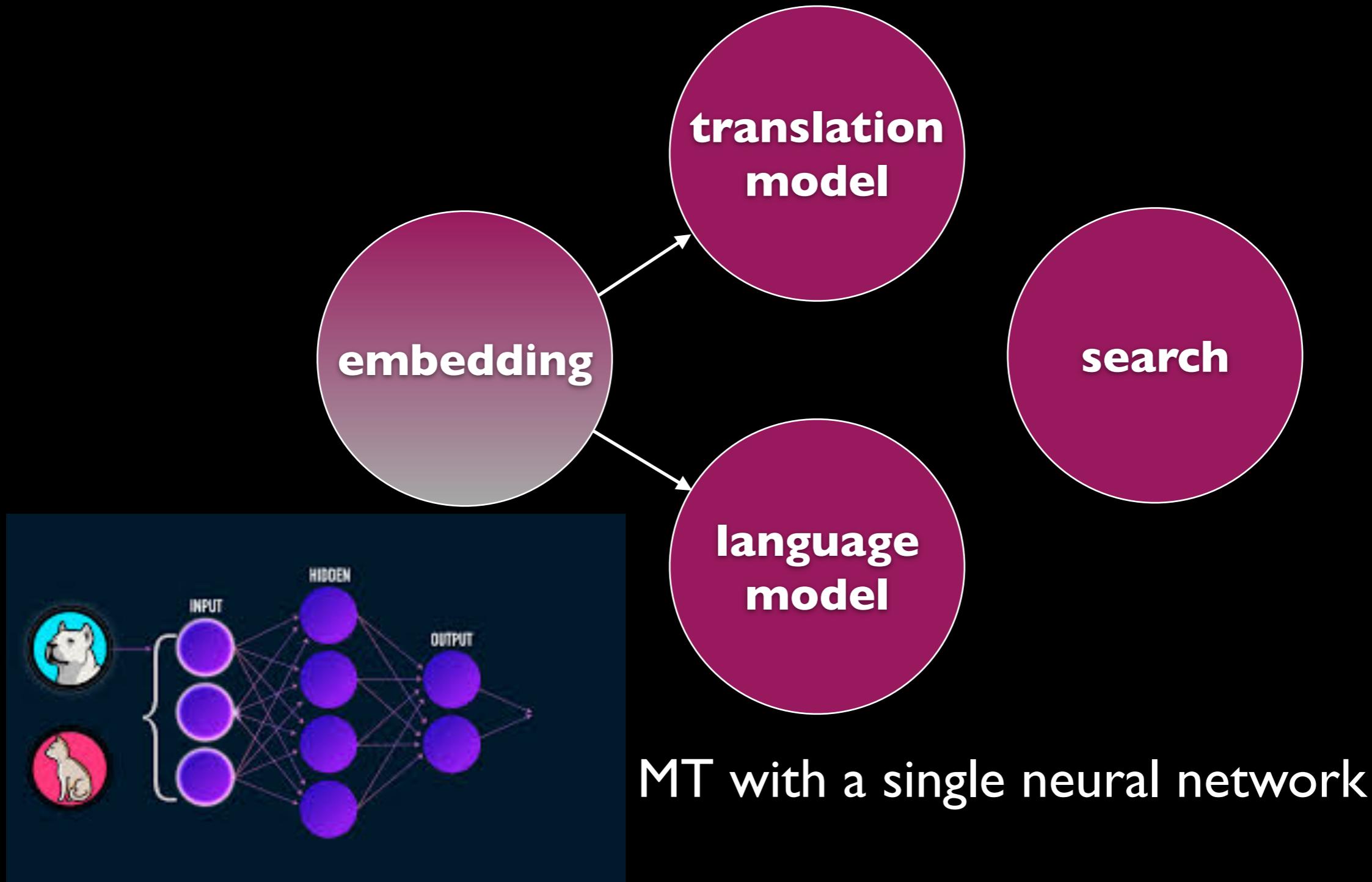
**translation
model**

**language
model**

search

MT with a single neural network

neural machine translation



neural machine translation



embedding

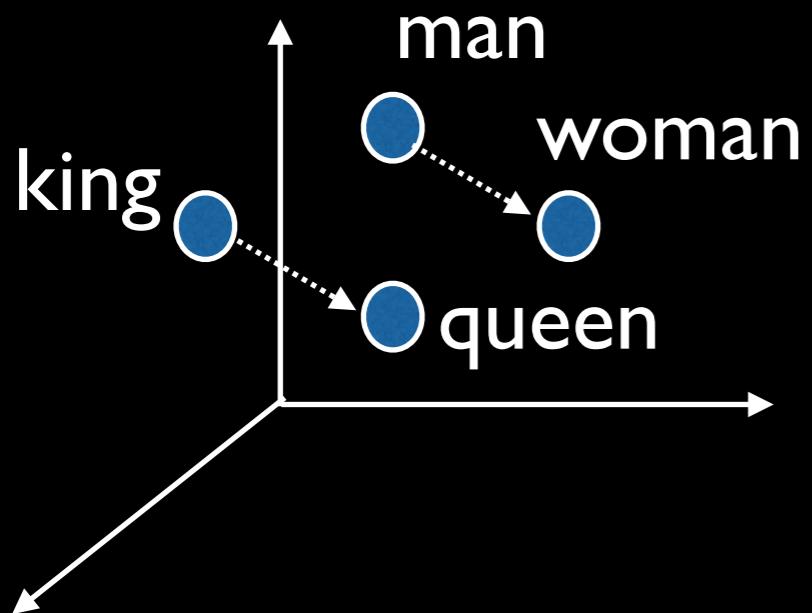
neural machine translation

introduced in the previous tutorial

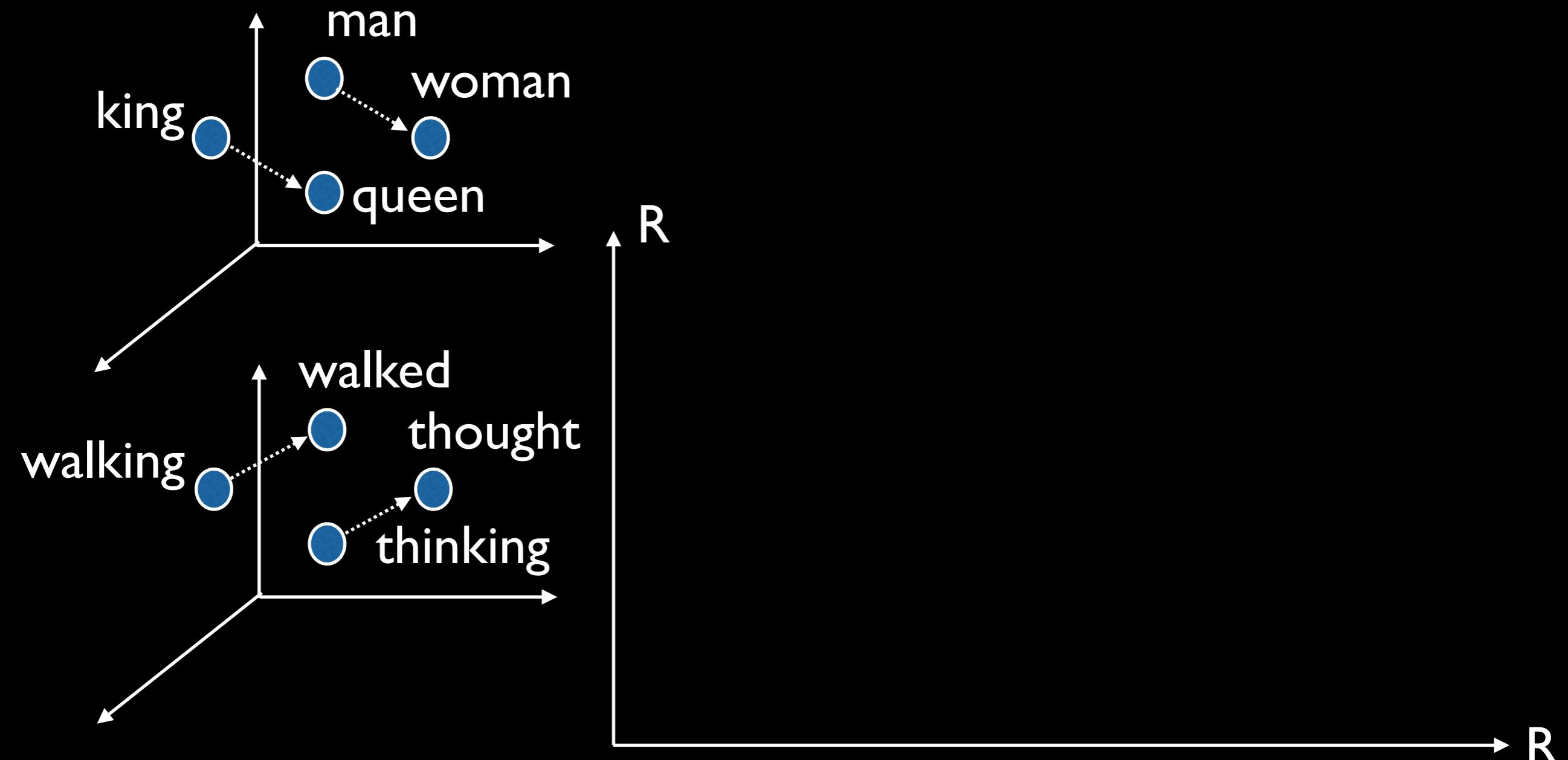


word embedding maps words to vectors

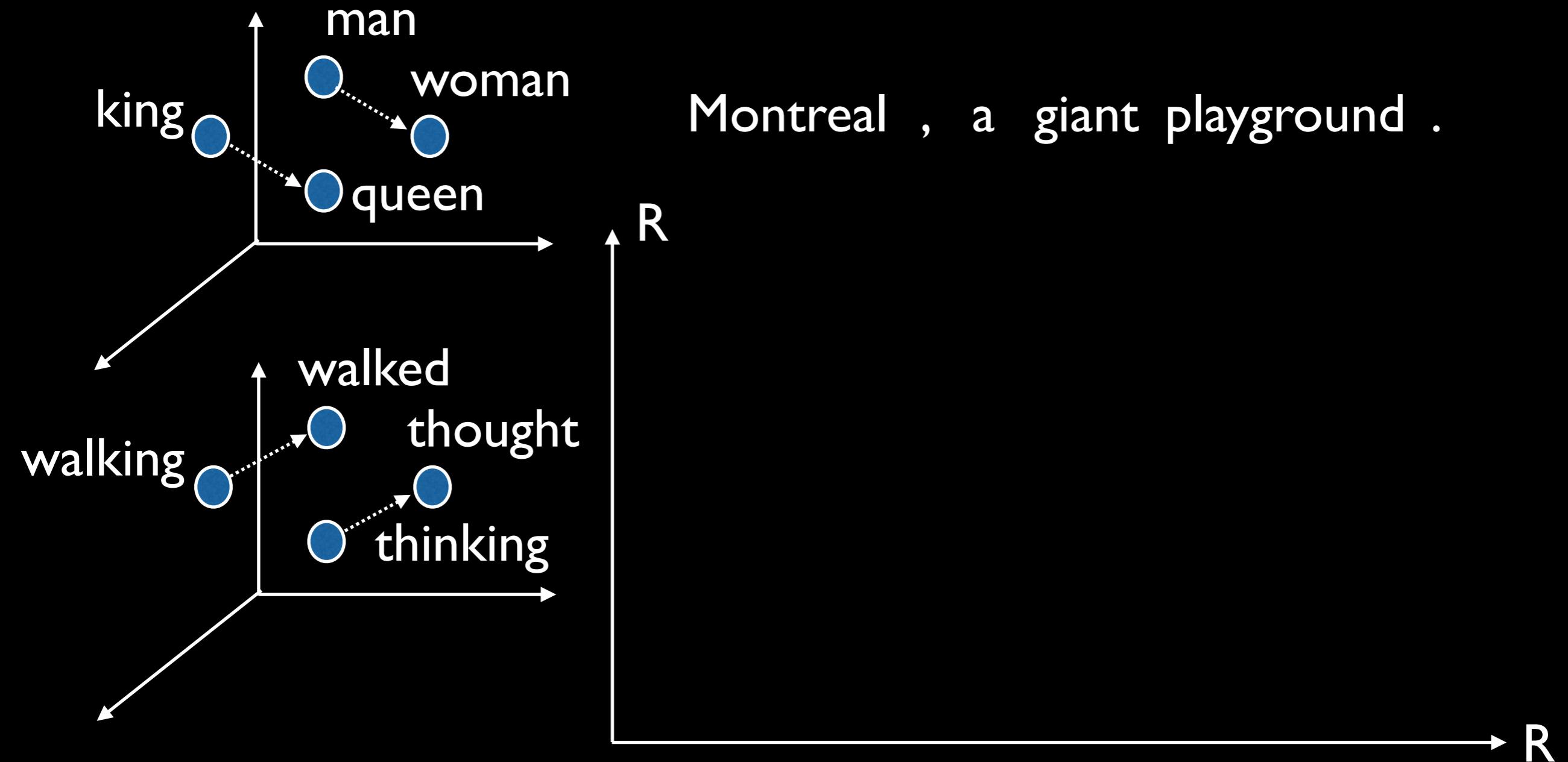
word embedding maps words to vectors



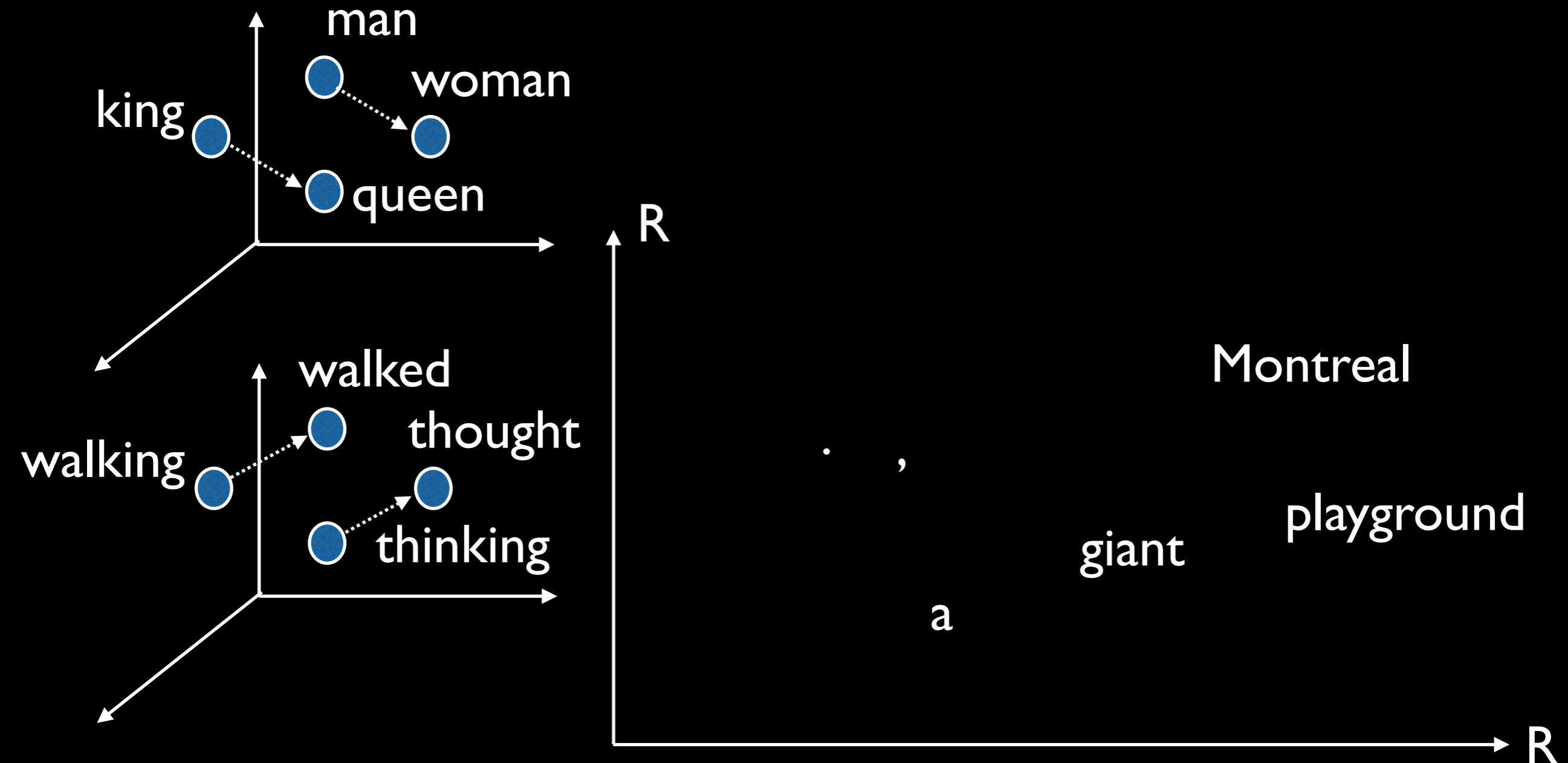
word embedding maps words to vectors



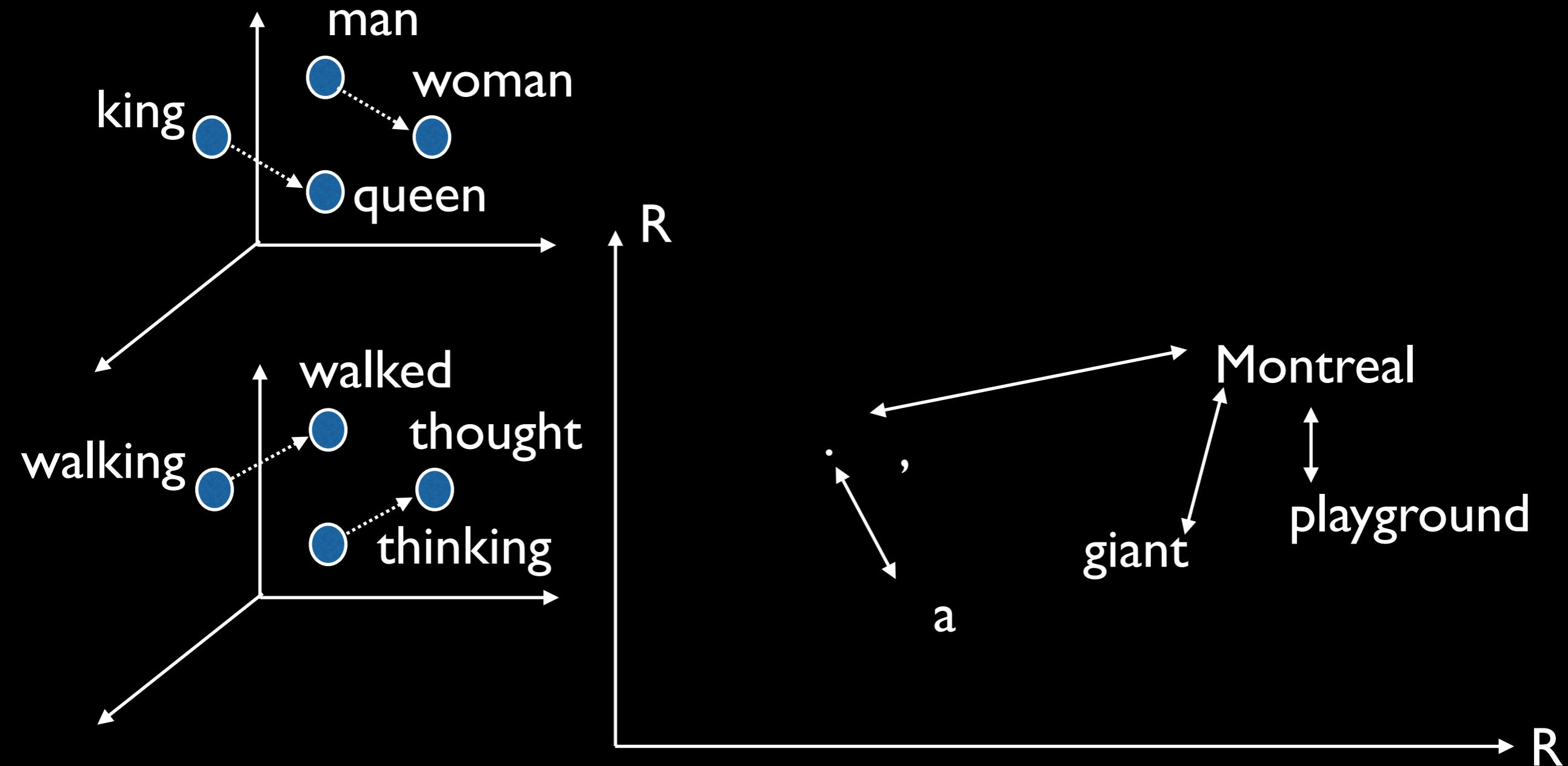
word embedding maps words to vectors



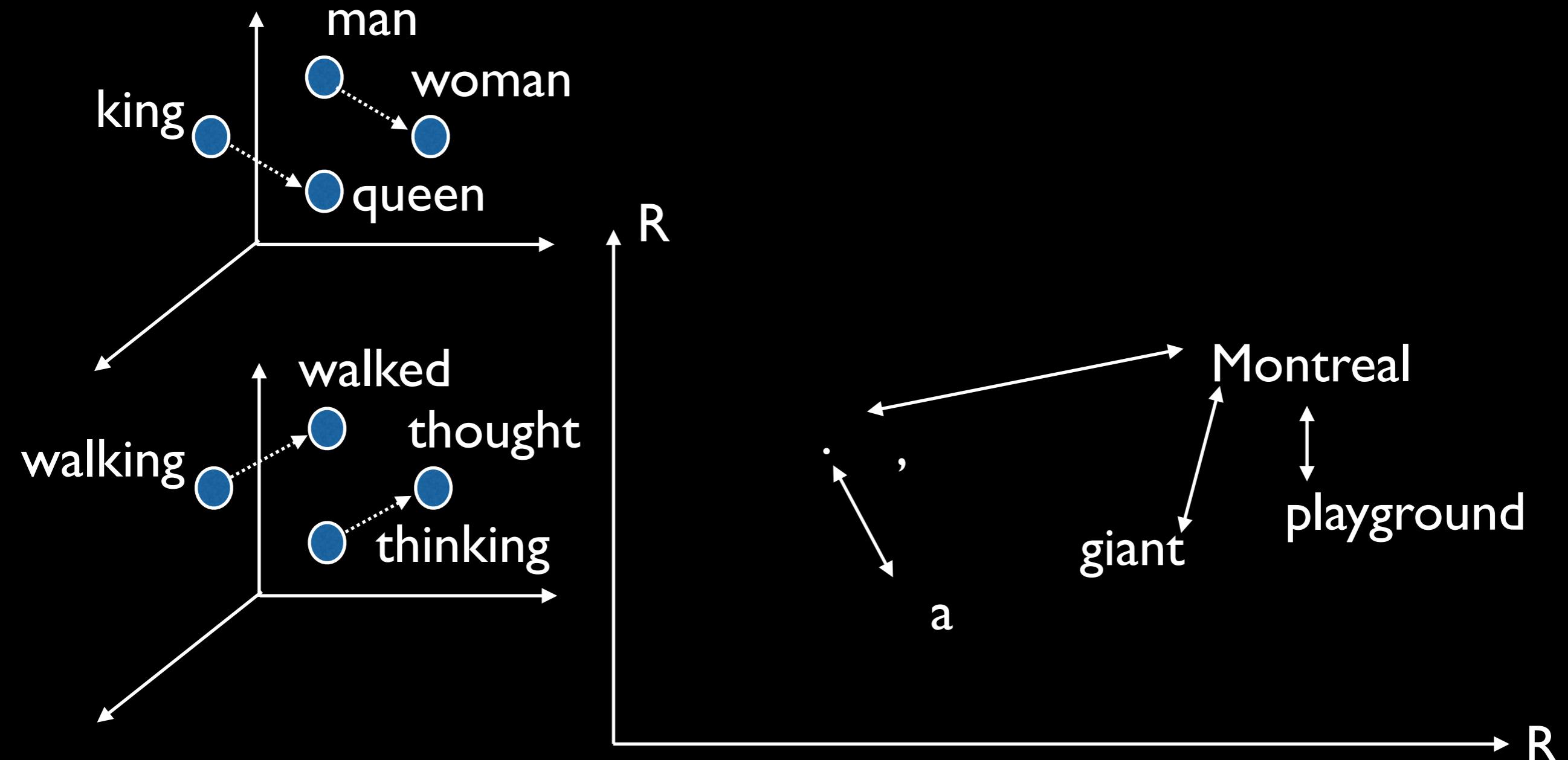
word embedding maps words to vectors



word embedding maps words to vectors

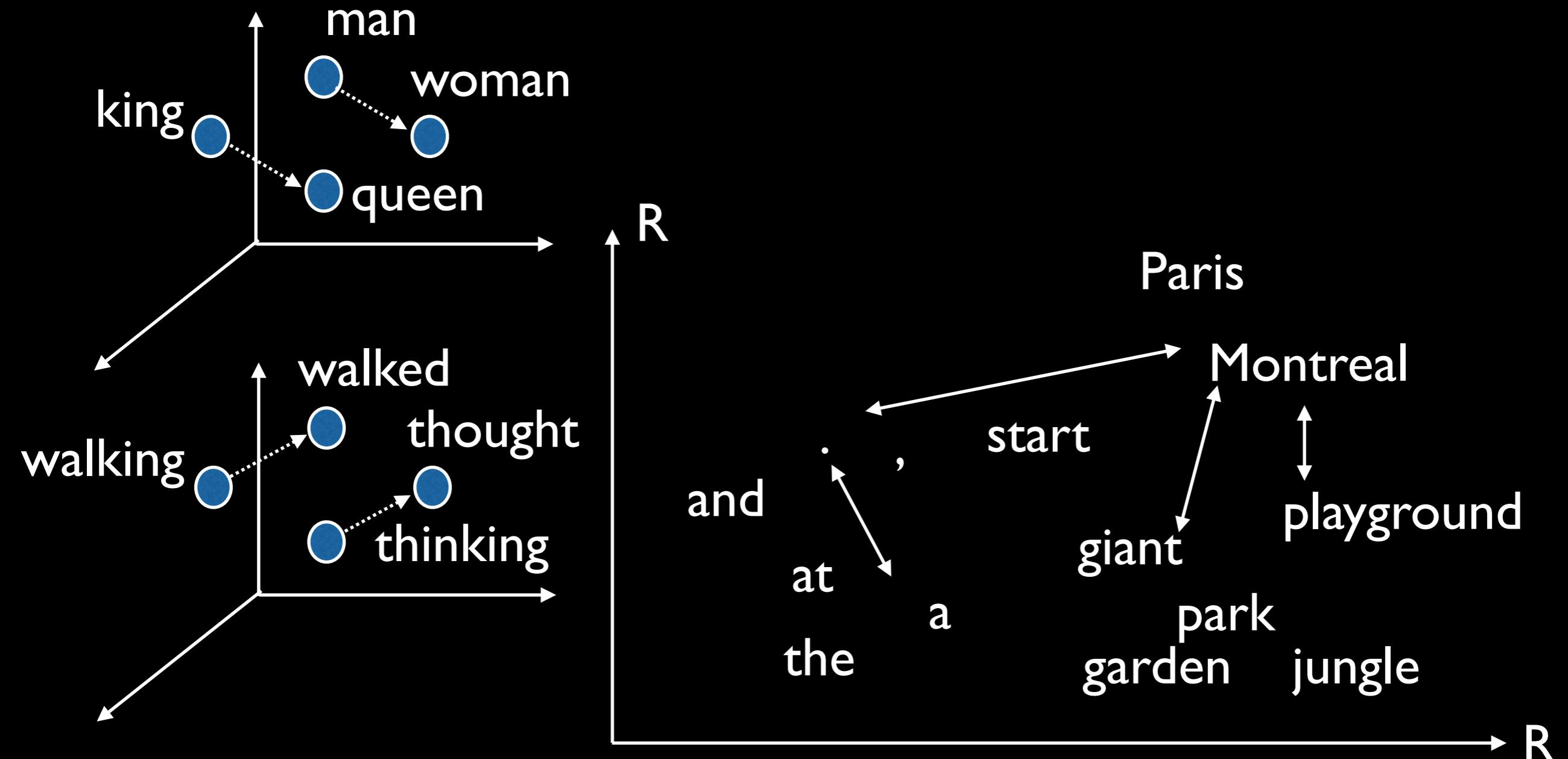


word embedding maps words to vectors



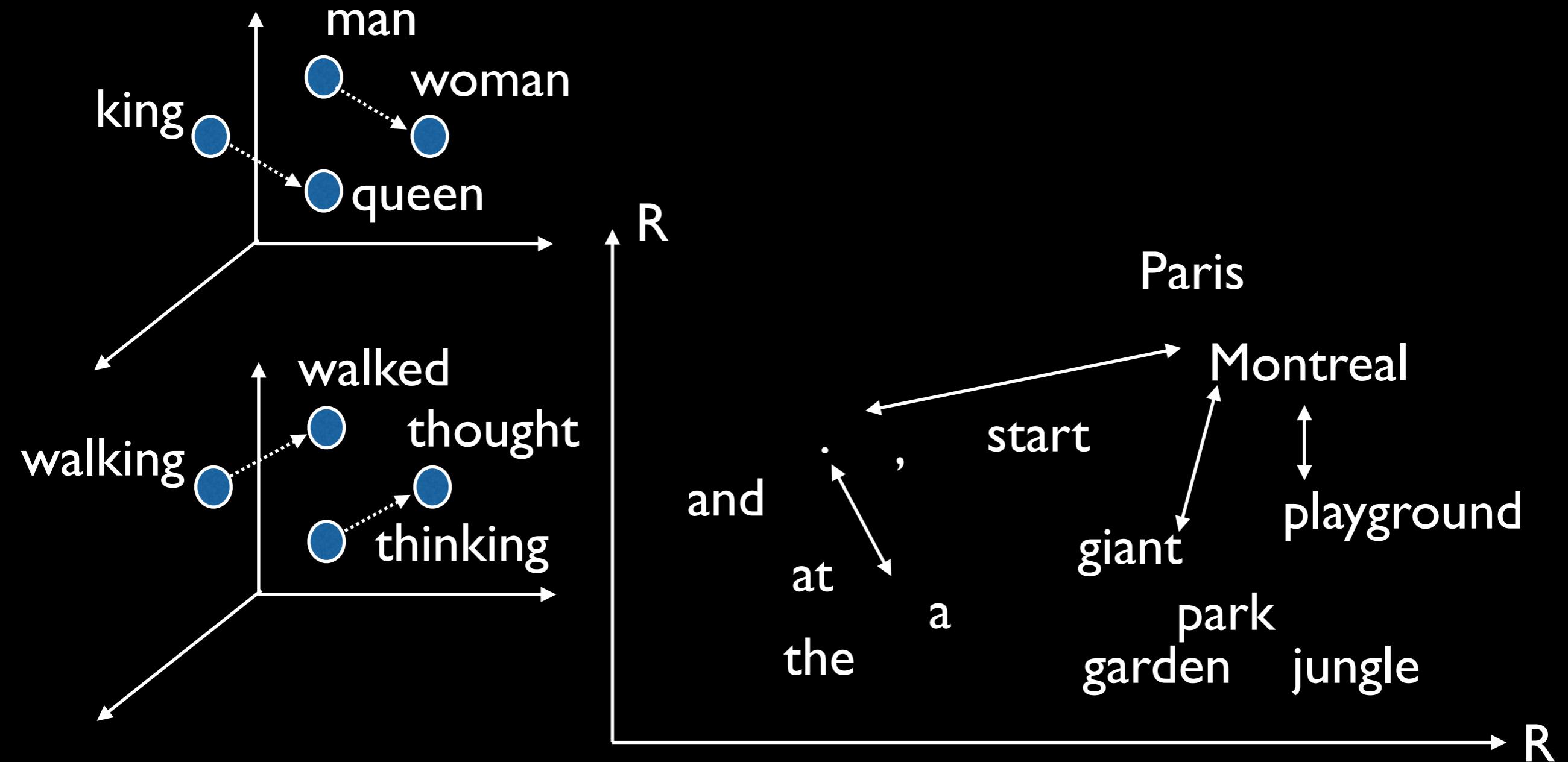
word pair semantic distance preserved

word embedding maps words to vectors



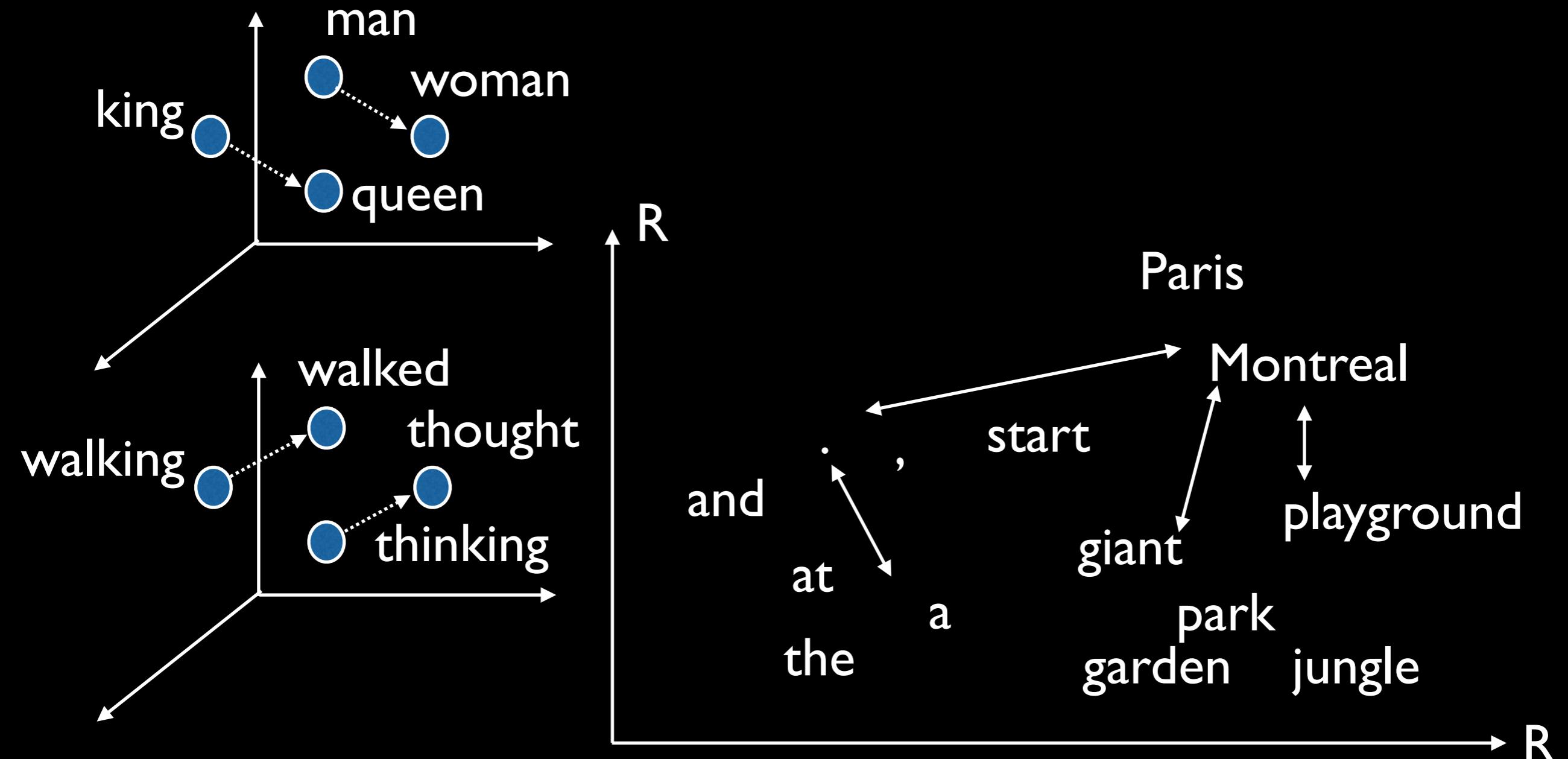
word pair semantic distance preserved

word embedding maps words to vectors



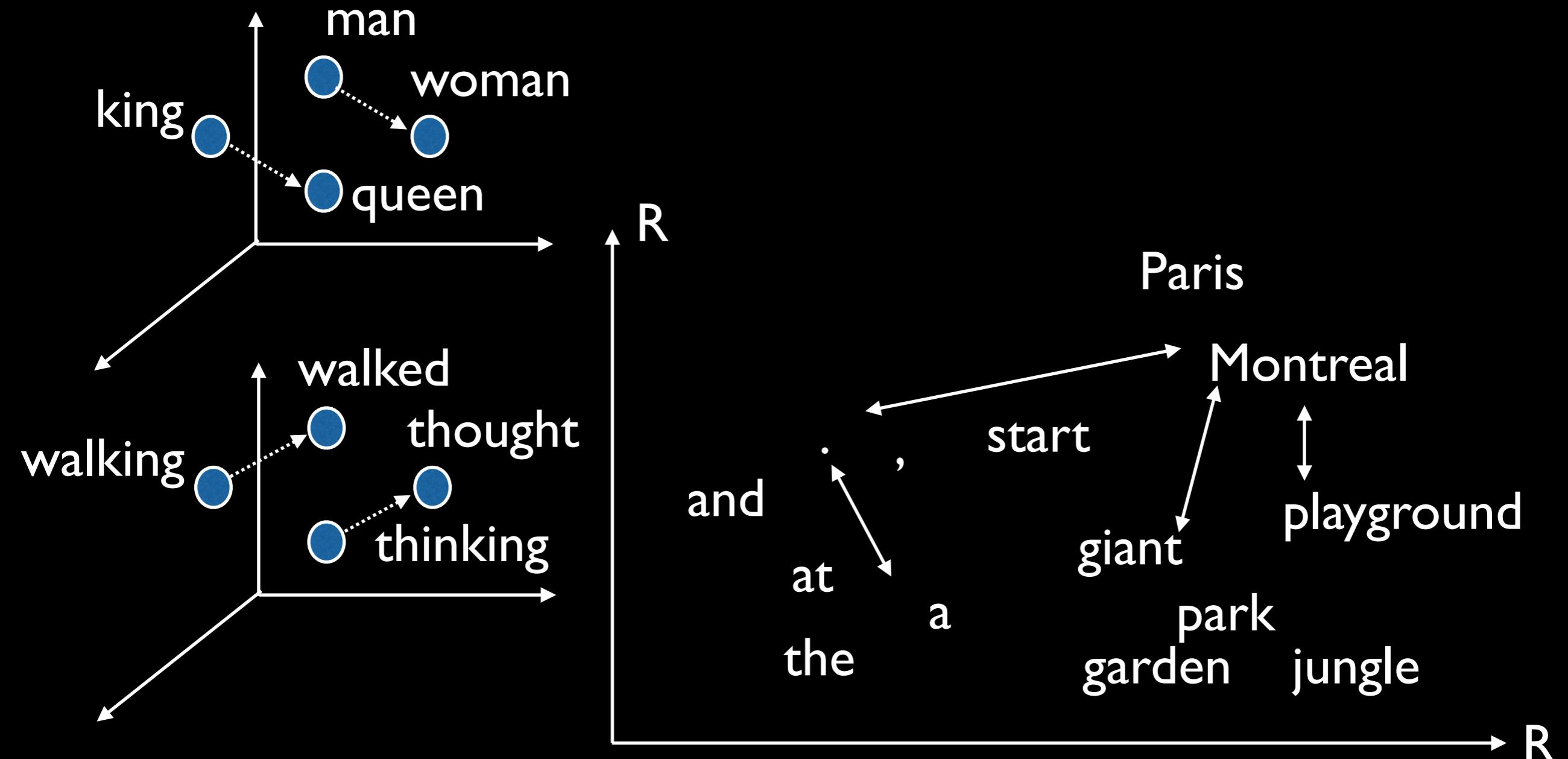
Question #3: better text embedding/representation?
BERT, ELMO, GloVec, FastText, ...

word embedding maps words to vectors



word pair semantic distance preserved

word embedding maps words to vectors

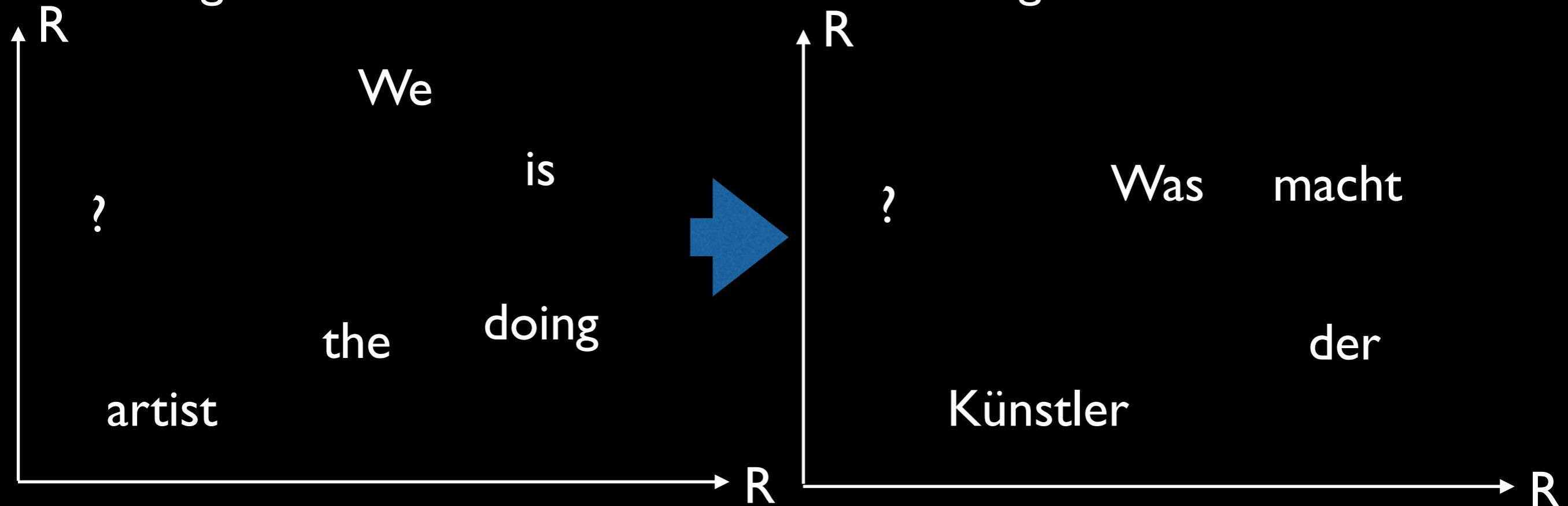


word pair semantic distance preserved

source to target language embedding

Source: Was macht der Künstler ?

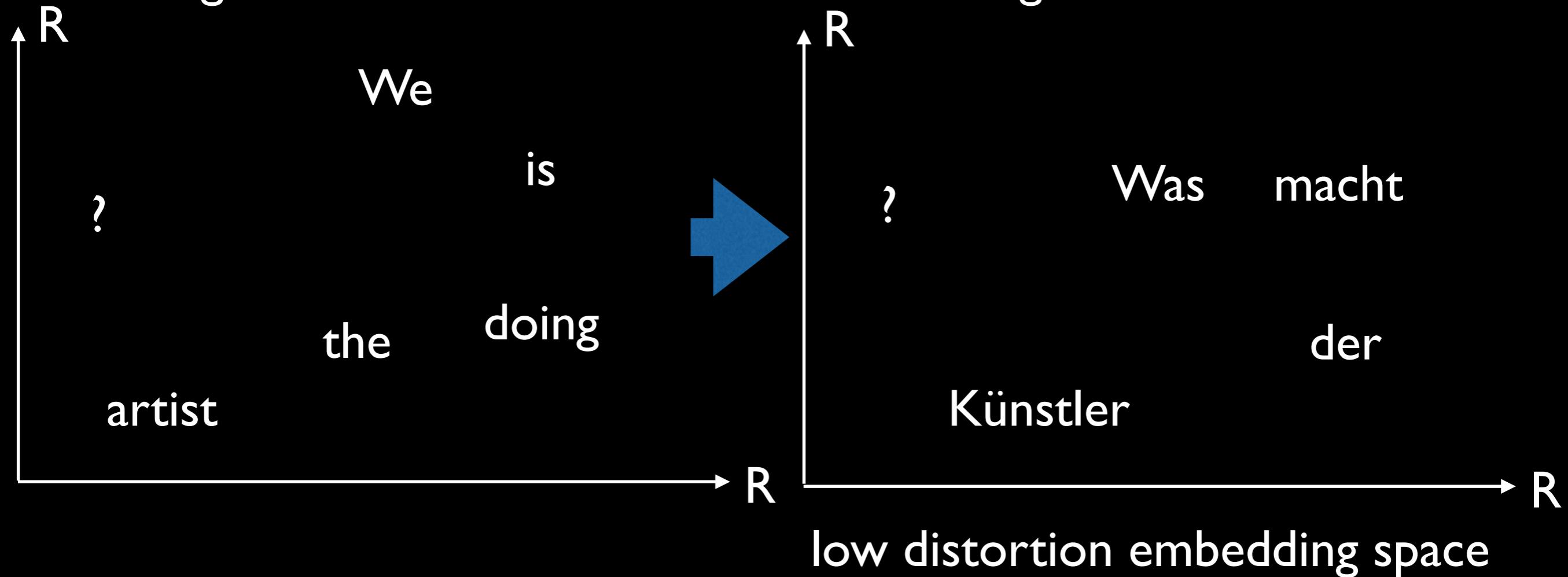
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

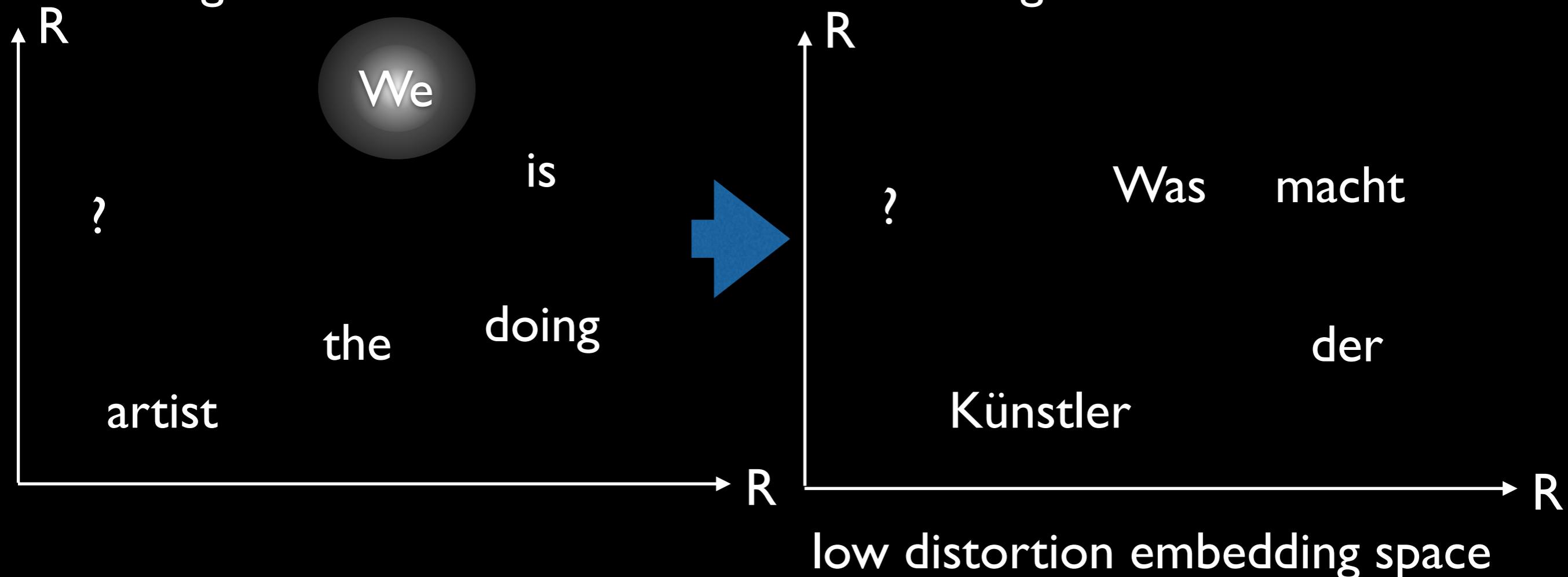
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

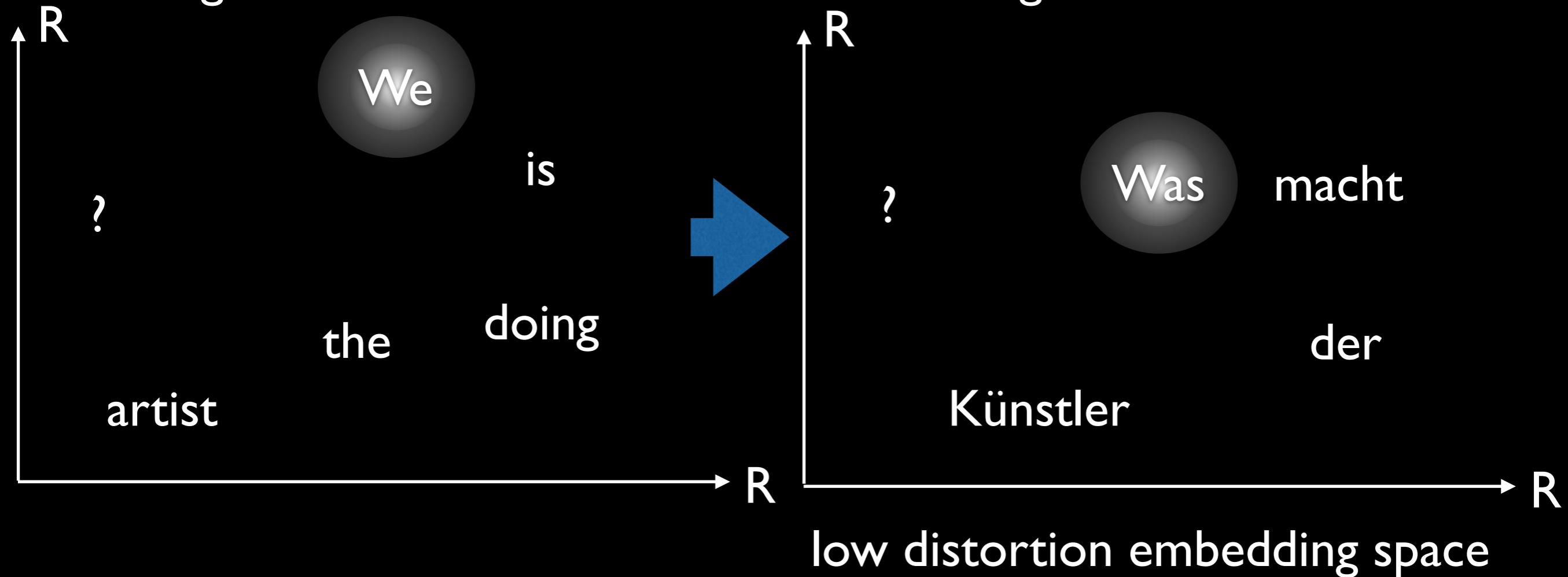
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

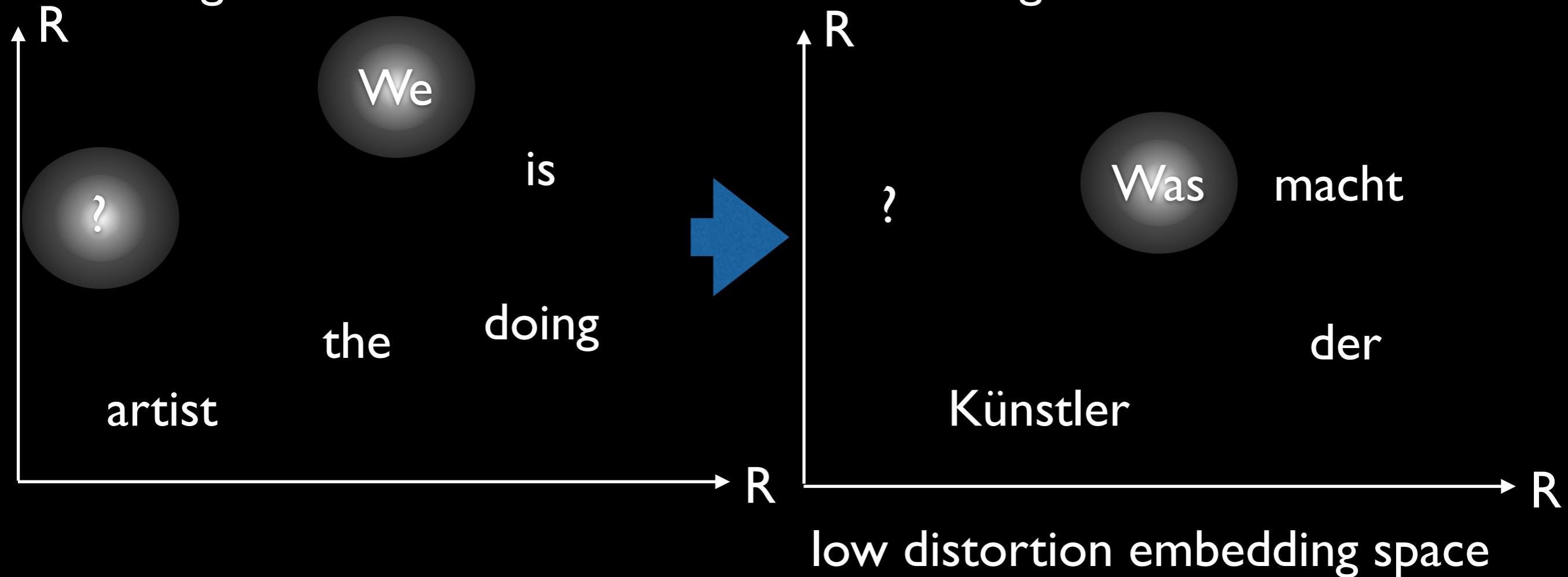
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

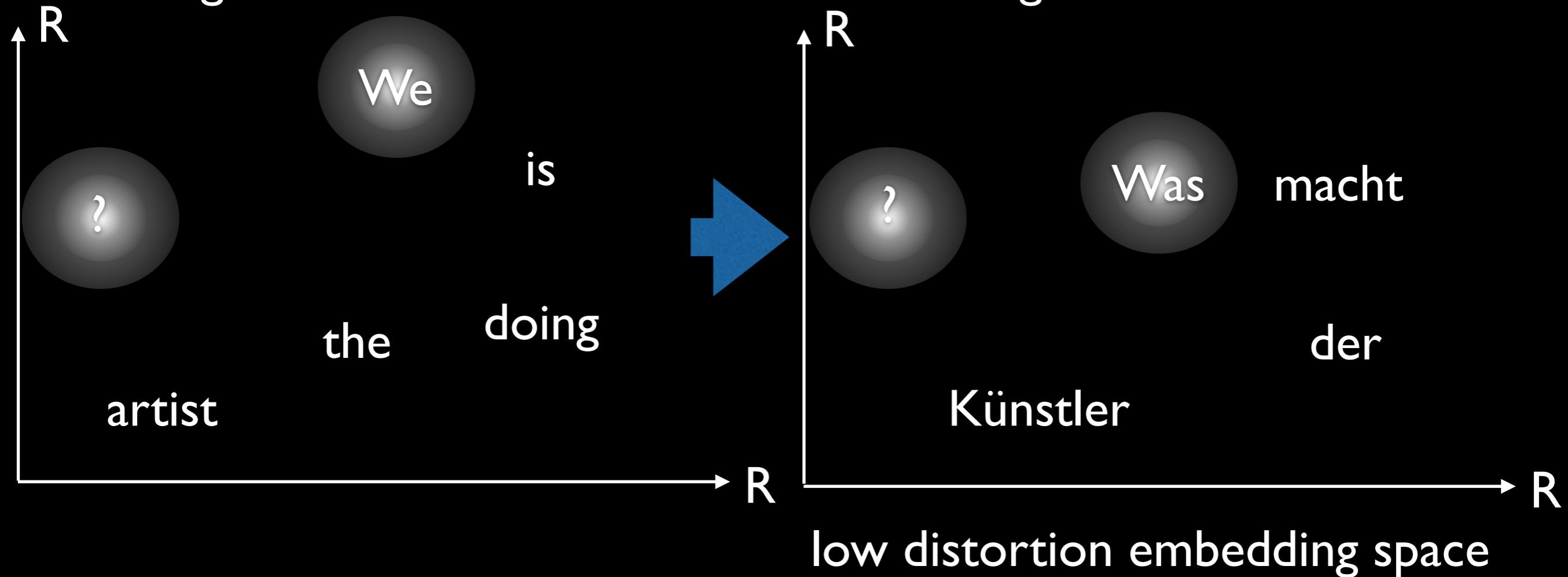
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

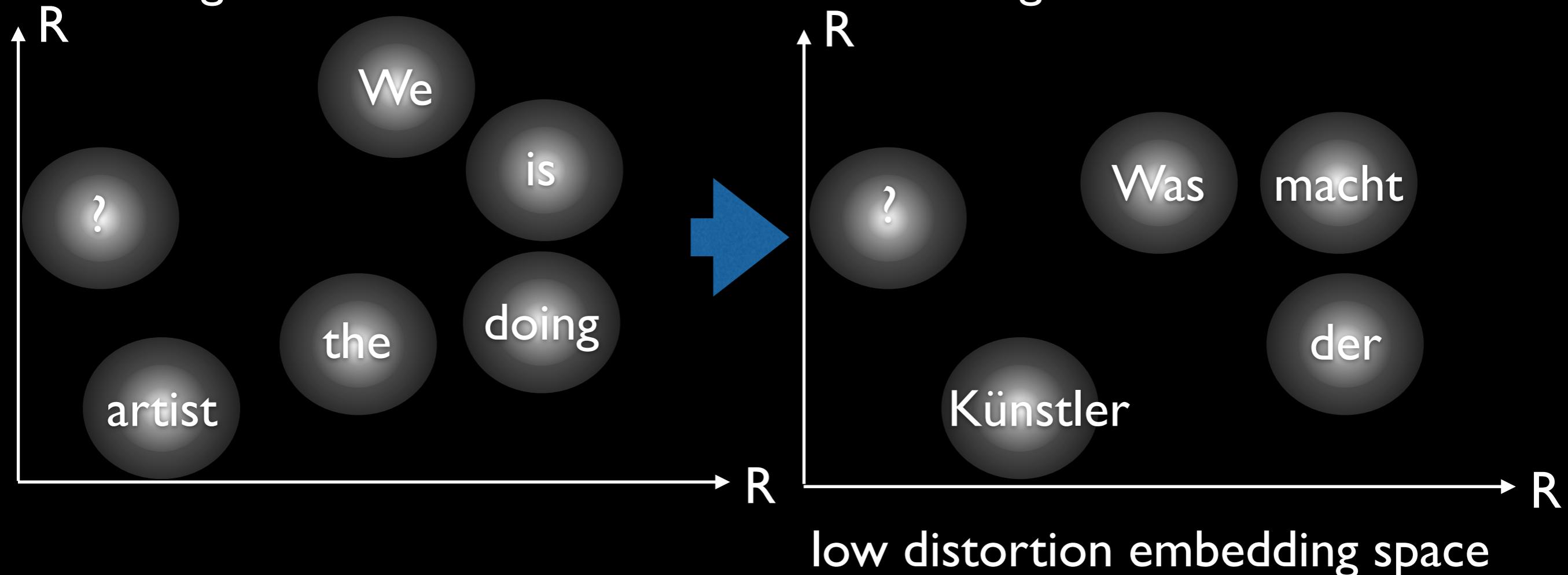
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

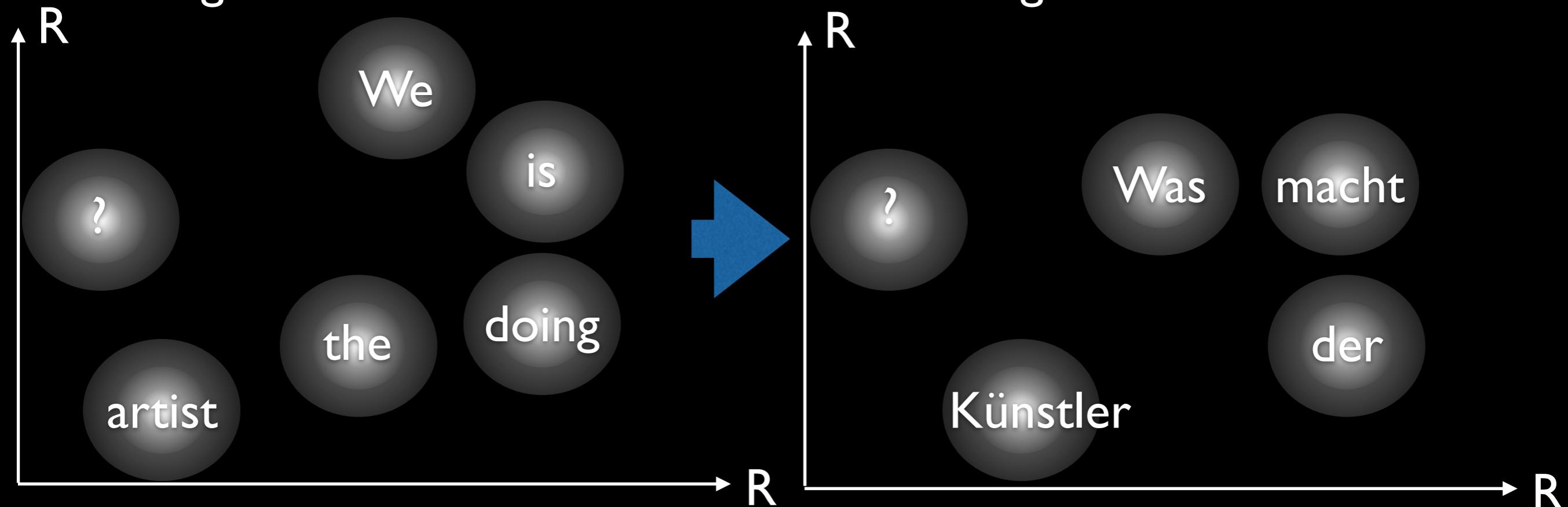
Target: What is the artist doing ?



source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?

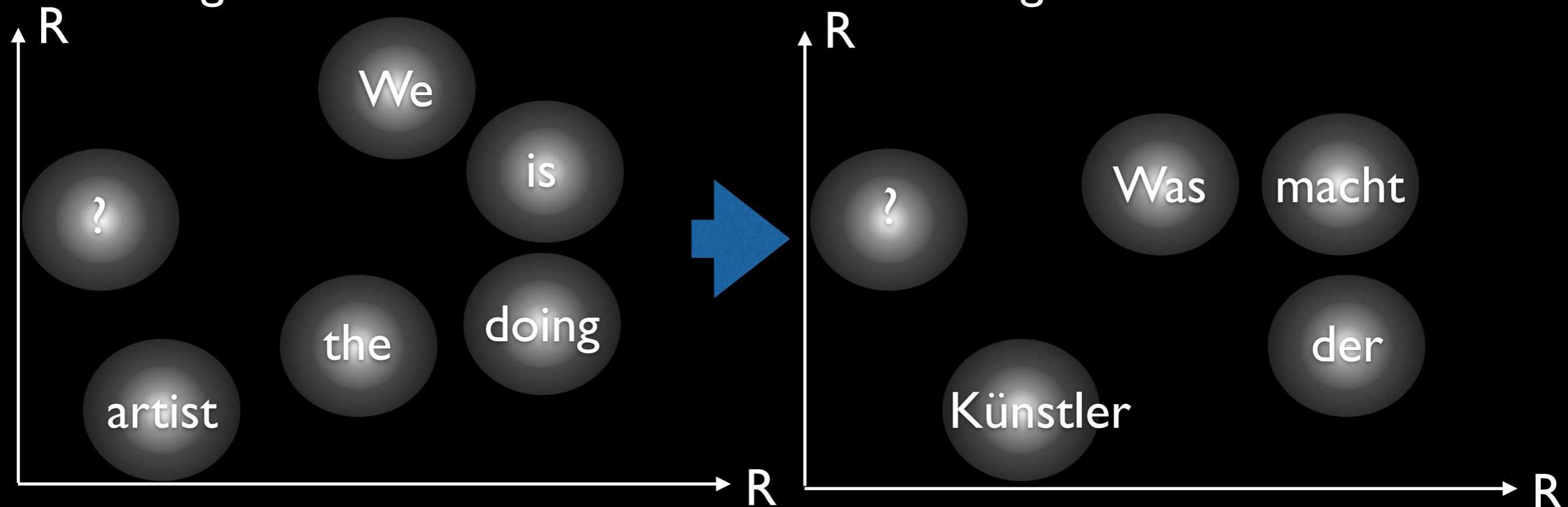


Can we learn only from distance in low distortion embedding space ?

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?



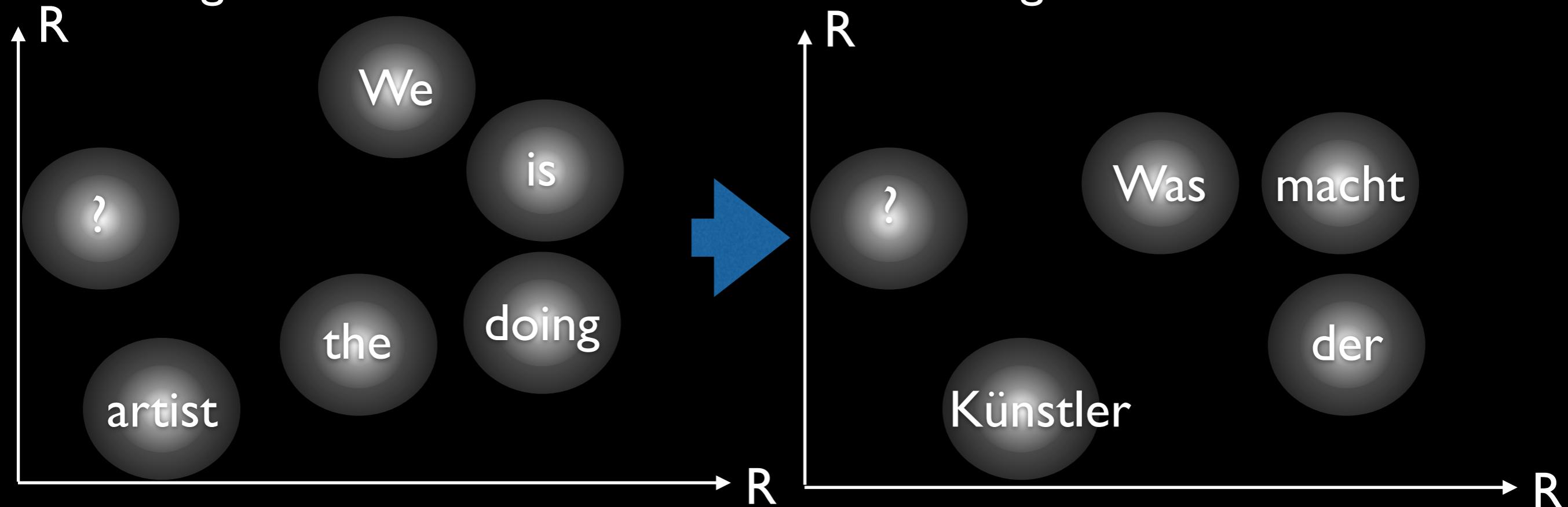
Can we learn only from distance in low distortion embedding space ?

No, in NLP we need language models

source to target language embedding

Source: Was macht der Künstler ?

Target: What is the artist doing ?



Can we learn only from distance in low distortion embedding space ?

No, in NLP we need language models

setting: multi-class classification + metric structure + 2 experts

we cannot combine them to a better one by querying them [PYX, 16]

machine translation components

what are
language models

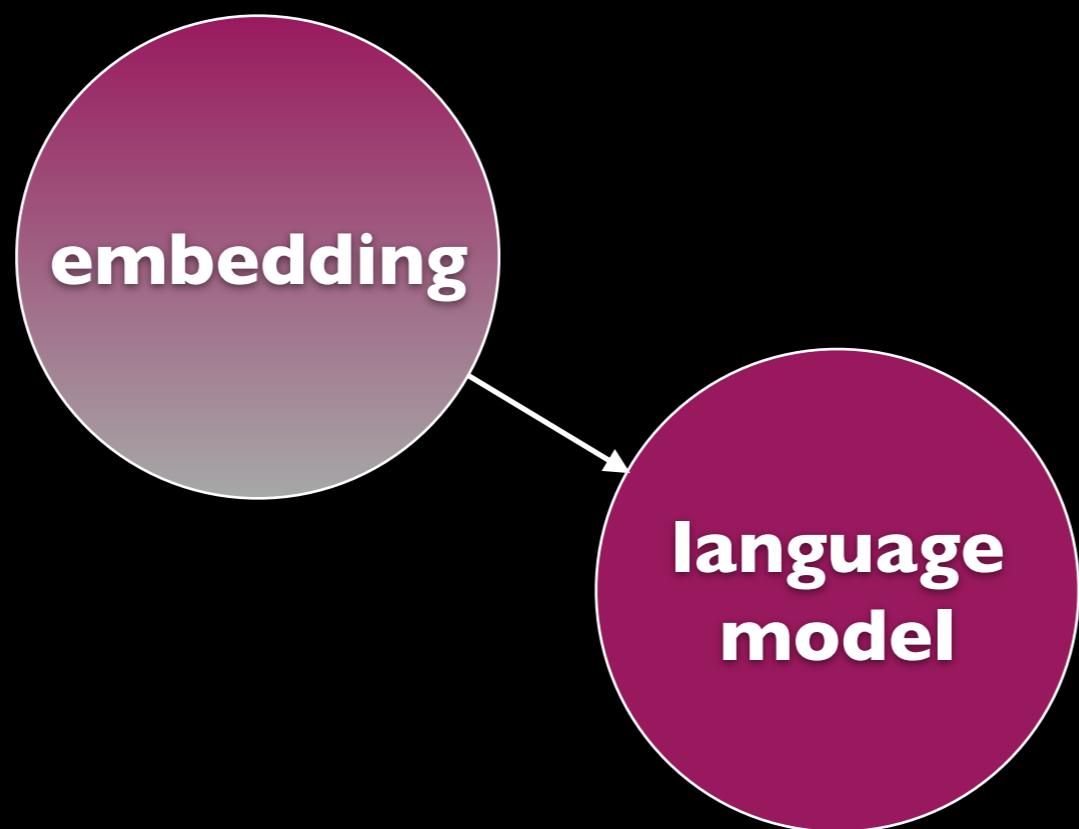
machine translation components

what are
language models



machine translation components

what are
language models



n-gram language model



language
model

n-gram language model

previously



language
model

n-gram language model

previously

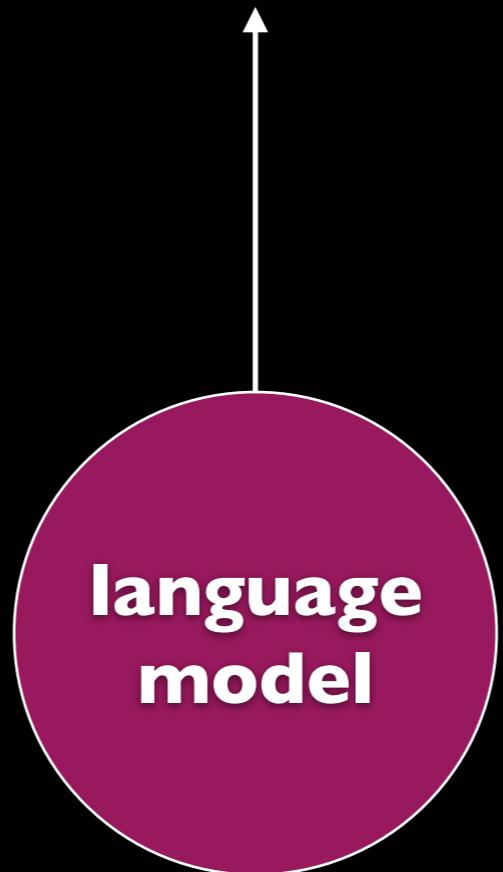
$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



n-gram language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$



evaluating language models: perplexity

let e_i be a word in the document that contains N words

PPL (perplexity) is measured as

$$\log \text{PPL} = -\frac{1}{N} \sum_{i=1}^N \log P(e_i | h_i)$$

h_i is the history of word e_i

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

language model

previously

$$\hat{e} = \operatorname{argmax}_e \{Pr(e|f)\}$$

$$= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}$$

let e_1^I be a sentence of a sequence of words $e_1, e_2 \dots e_I$

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule} \\ &\approx P(e_i | e_1, \dots, e_{i-n+1}) \quad \text{estimated as relative frequency}\end{aligned}$$

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule} \\ &\approx P(e_i | e_1, \dots, e_{i-n+1}) \quad \begin{matrix} \text{estimated as} \\ \text{relative frequency} \end{matrix}\end{aligned}$$

n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$



n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$



n-gram based; if n-gram unseen, then back-off

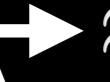
language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

statistical  $\approx P(e_i | e_1, \dots, e_{i-n+1})$ estimated as relative frequency

 n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$

n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$



n-gram based; if n-gram unseen, then back-off

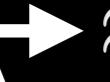
language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

statistical  $\approx P(e_i | \text{[redacted]} \dots, e_{i-n+1})$ estimated as relative frequency

 n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$



n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical  $\approx P(e_i | e_1, \dots, e_{i-n+1})$

n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$

n-gram based; if n-gram unseen, then back-off

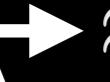
language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

statistical  $\approx P(e_i | e_1, \dots, e_{i-n+1})$ estimated as relative frequency

 n-gram based; if n-gram unseen, then back-off

language model

previously

$$\begin{aligned}\hat{e} &= \operatorname{argmax}_e \{Pr(e|f)\} \\ &= \operatorname{argmax}_e \{Pr(f|e) \cdot Pr(e)\}\end{aligned}$$

let e_1^I be a sentence of a sequence of words e_1, e_2, \dots, e_I

$$\begin{aligned}Pr(e_1^I) &= Pr(e_1, e_2, \dots, e_I) \\ &= \prod_i Pr(e_i | e_1, \dots, e_{i-1}) \quad \text{chain rule}\end{aligned}$$

estimated as
relative frequency

statistical 

$$\approx P(e_i | e_1, \dots, e_{i-n+1})$$

n-gram based; if n-gram unseen, then back-off

long-term memory for language models

long-term memory for language models

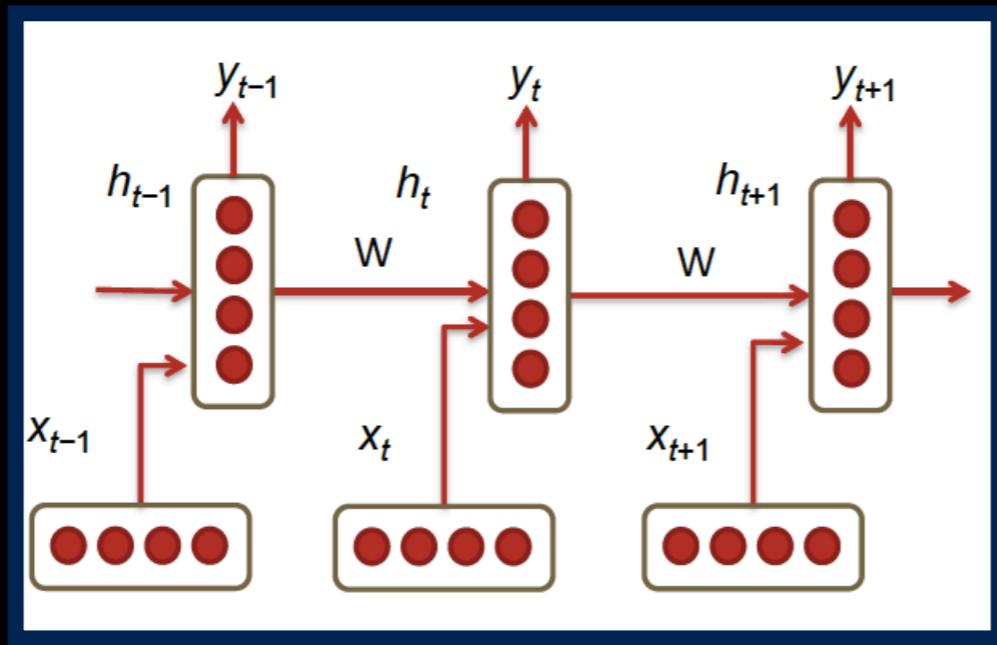
- n-gram LM making prediction on fixed windows
- past n words my not be sufficient to capture the context

long-term memory for language models

- n-gram LM making prediction on fixed windows
- past n words may not be sufficient to capture the context

RNNs are capable of conditioning the model on all previous words

neural language model with RNN



$$h_t = \sigma(W^{(hh)} h_{t-1} + W^{(hx)} x_{[t]})$$

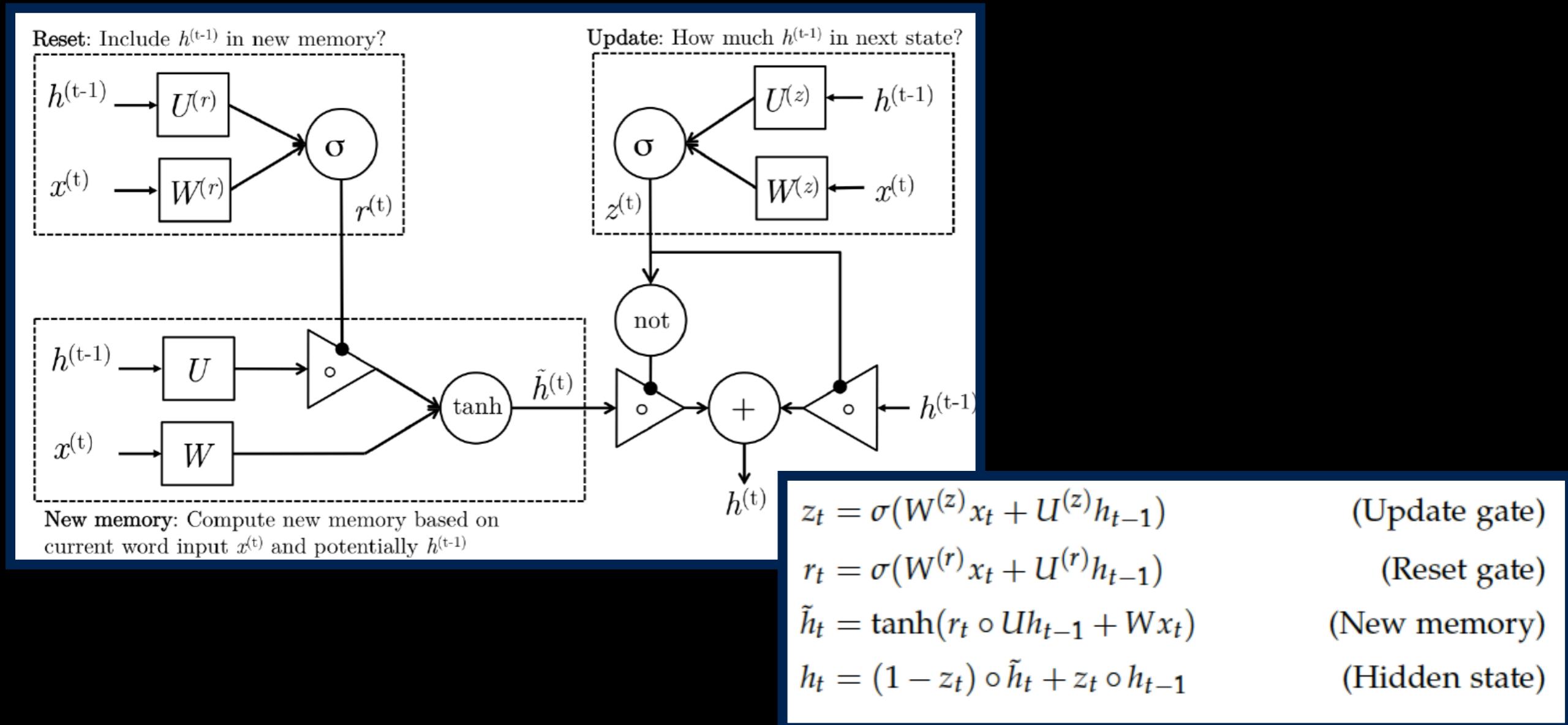
$$\hat{y}_t = softmax(W^{(S)} h_t)$$

$$J = -\frac{1}{T} \sum_{t=1}^T J^{(t)}(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{j=1}^{|V|} y_{t,j} \times \log(\hat{y}_{t,j})$$

- x_t : input word vector at time t
- W : weights matrix to condition t
- h_{t-1} : output of the non-linear function at the previous time step
- σ : the non-linearity function

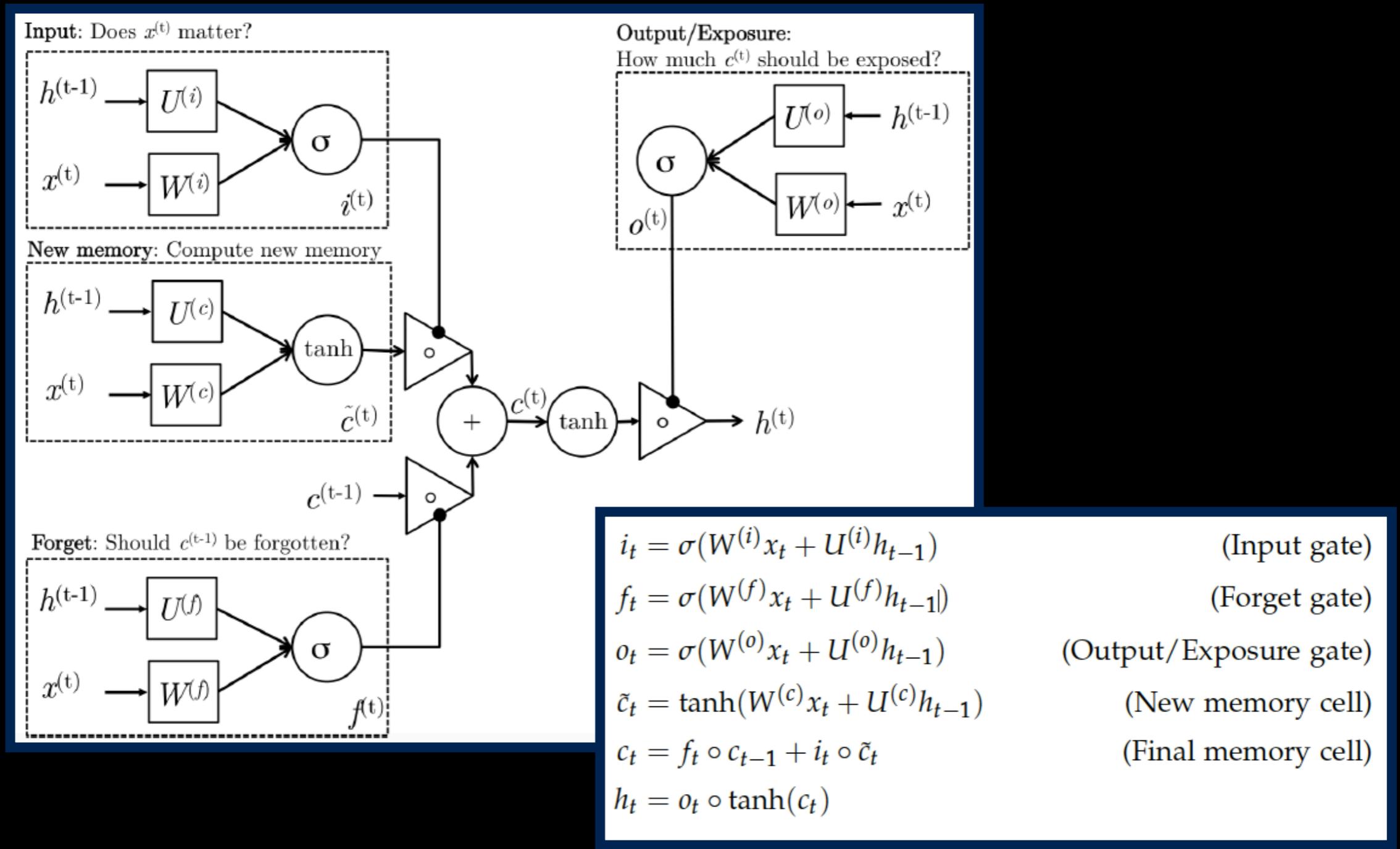
Gated Recurrent Units (GRU)

- problem of vanishing gradients makes RNNs hard to train for long-term dependency
- use more complex units for activation



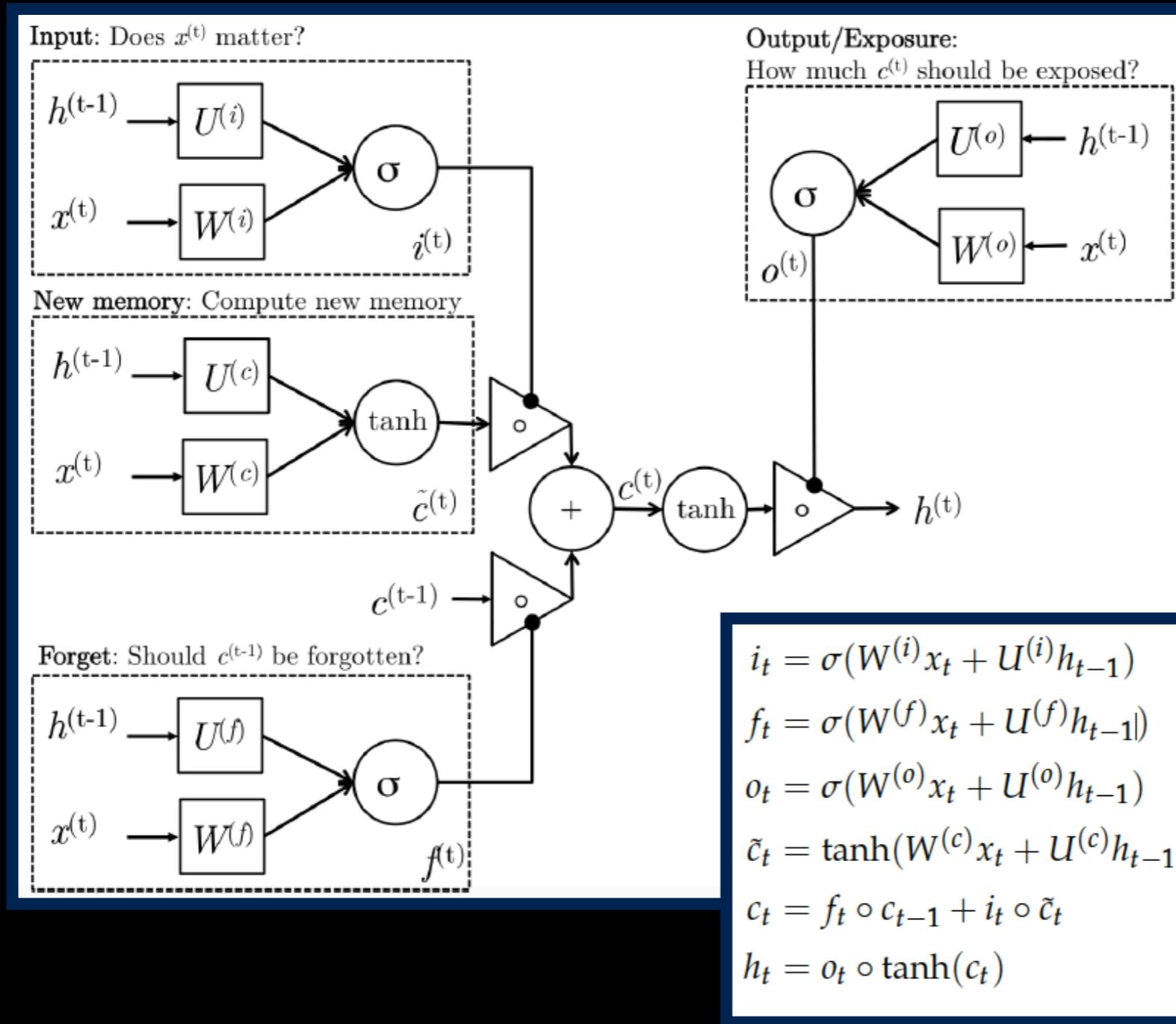
Long-Short-Term-Memories

another type of complex activation unit



Long-Short-Term-Memories

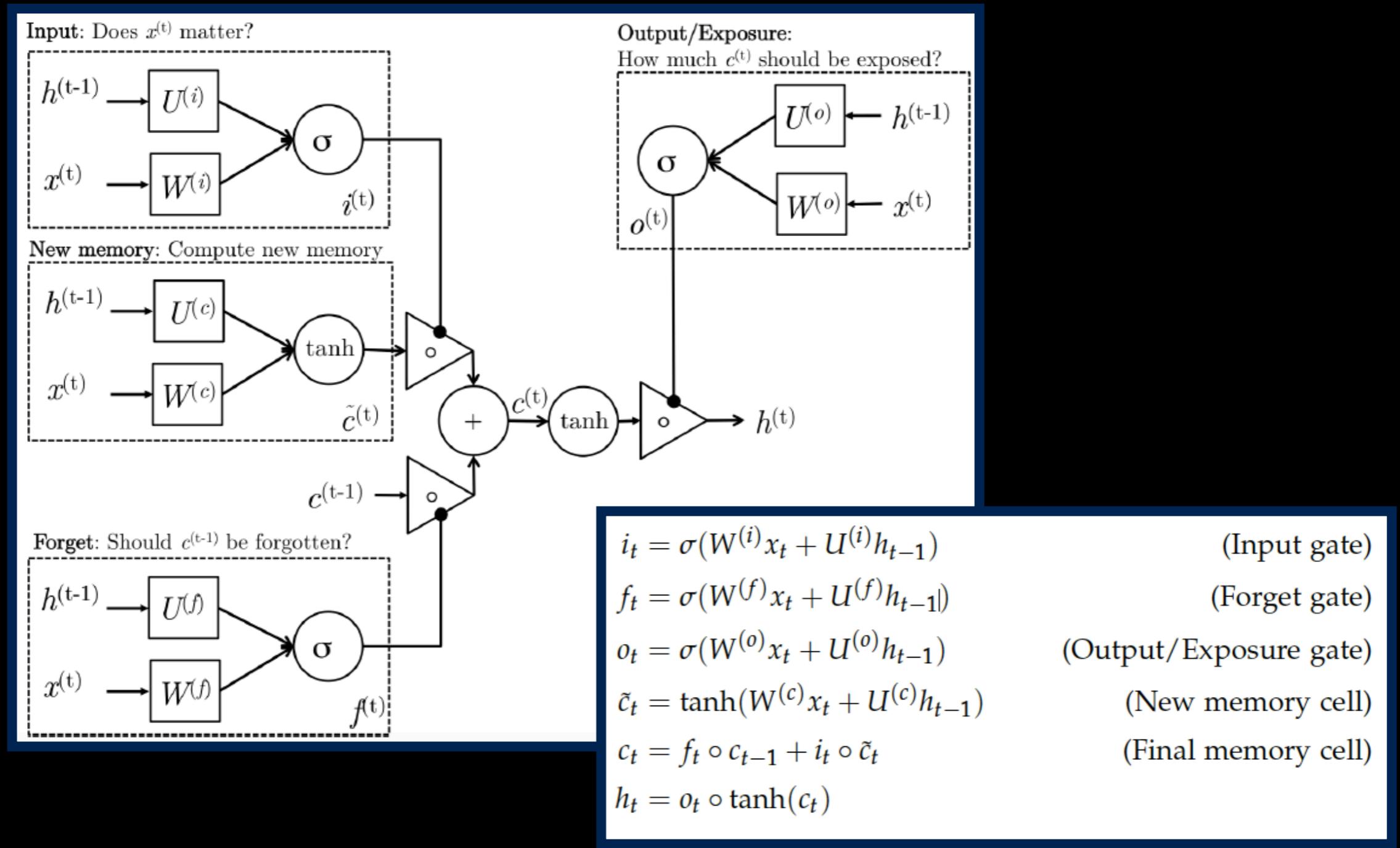
Question #4: contextual memory in language model



$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1})$	(Input gate)
$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1})$	(Forget gate)
$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1})$	(Output/Exposure gate)
$\tilde{c}_t = \tanh(W^{(c)}x_t + U^{(c)}h_{t-1})$	(New memory cell)
$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$	(Final memory cell)
$h_t = o_t \circ \tanh(c_t)$	

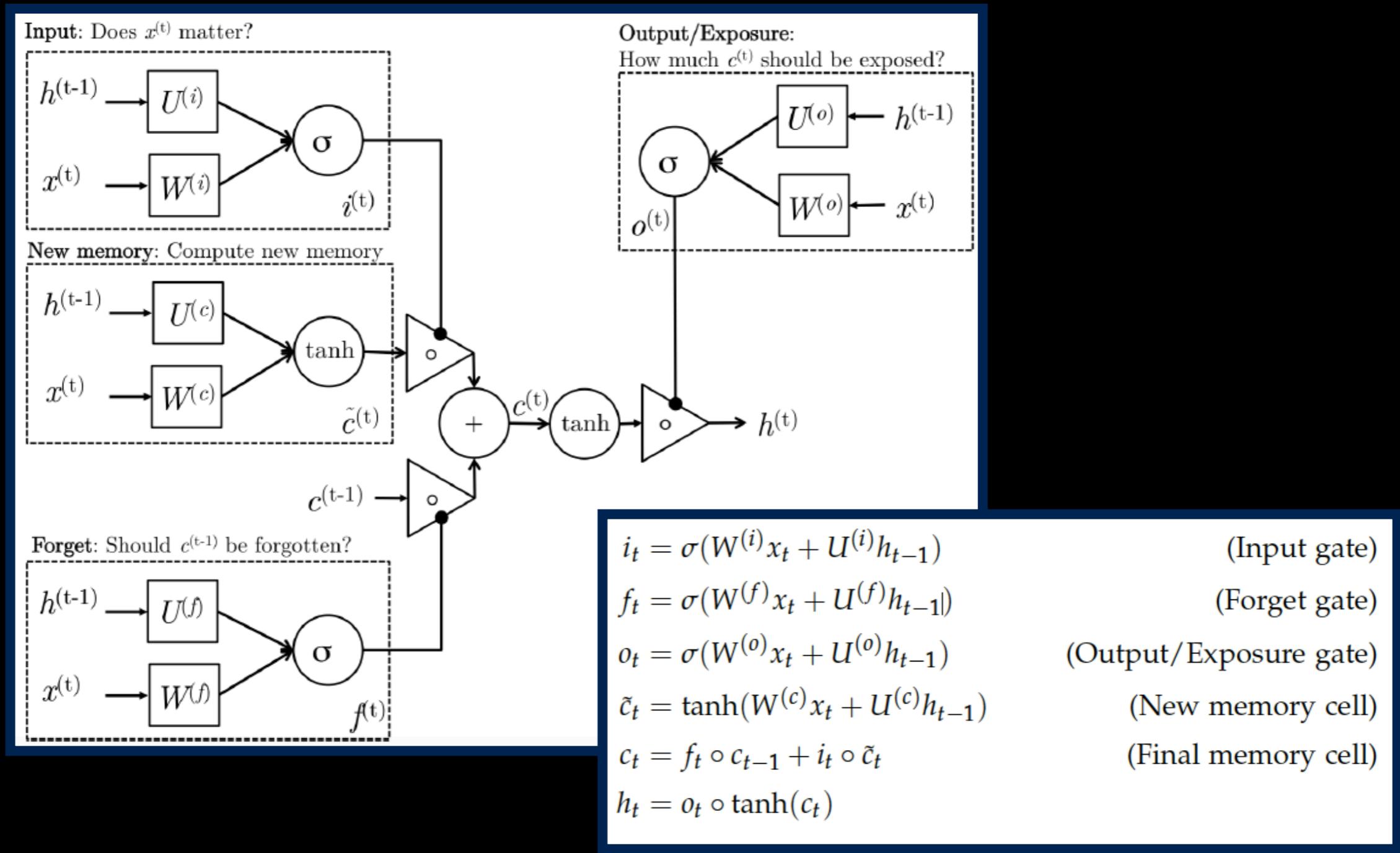
Long-Short-Term-Memories

another type of complex activation unit



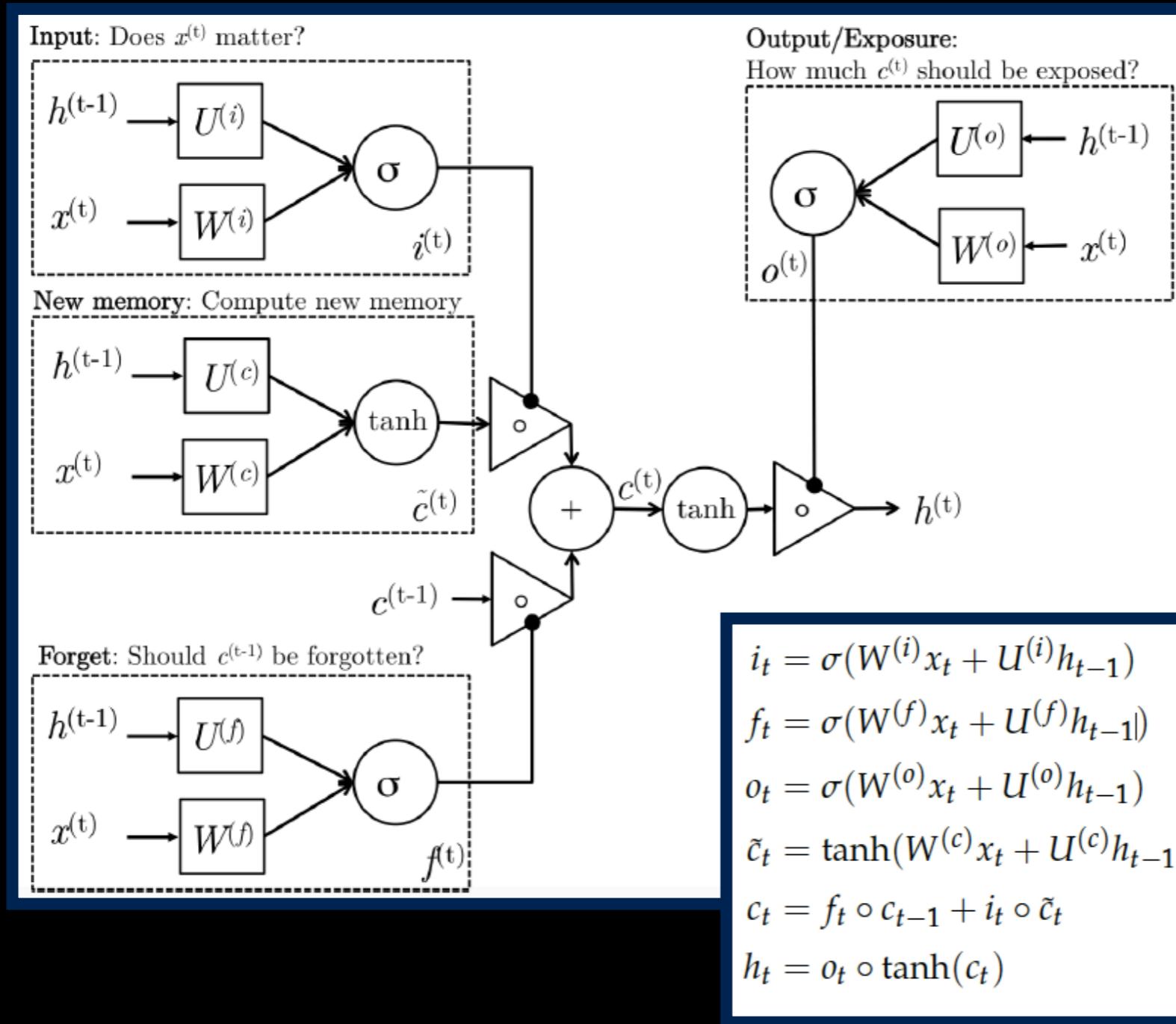
Long-Short-Term-Memories

another type of complex activation unit



Long-Short-Term-Memories

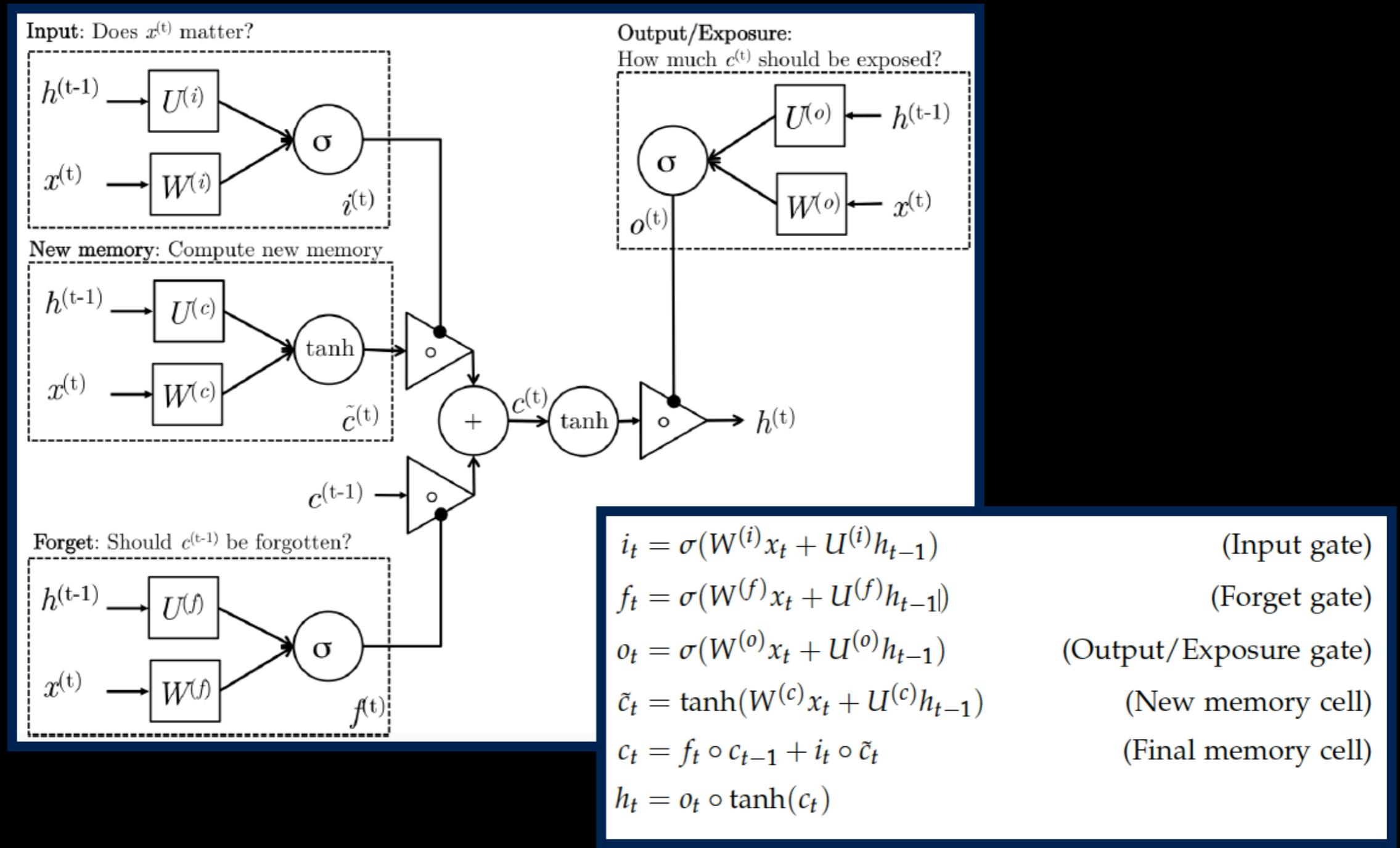
Question #5: affective neuron activation function



$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1})$	(Input gate)
$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1})$	(Forget gate)
$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1})$	(Output/Exposure gate)
$\tilde{c}_t = \tanh(W^{(c)}x_t + U^{(c)}h_{t-1})$	(New memory cell)
$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$	(Final memory cell)
$h_t = o_t \circ \tanh(c_t)$	

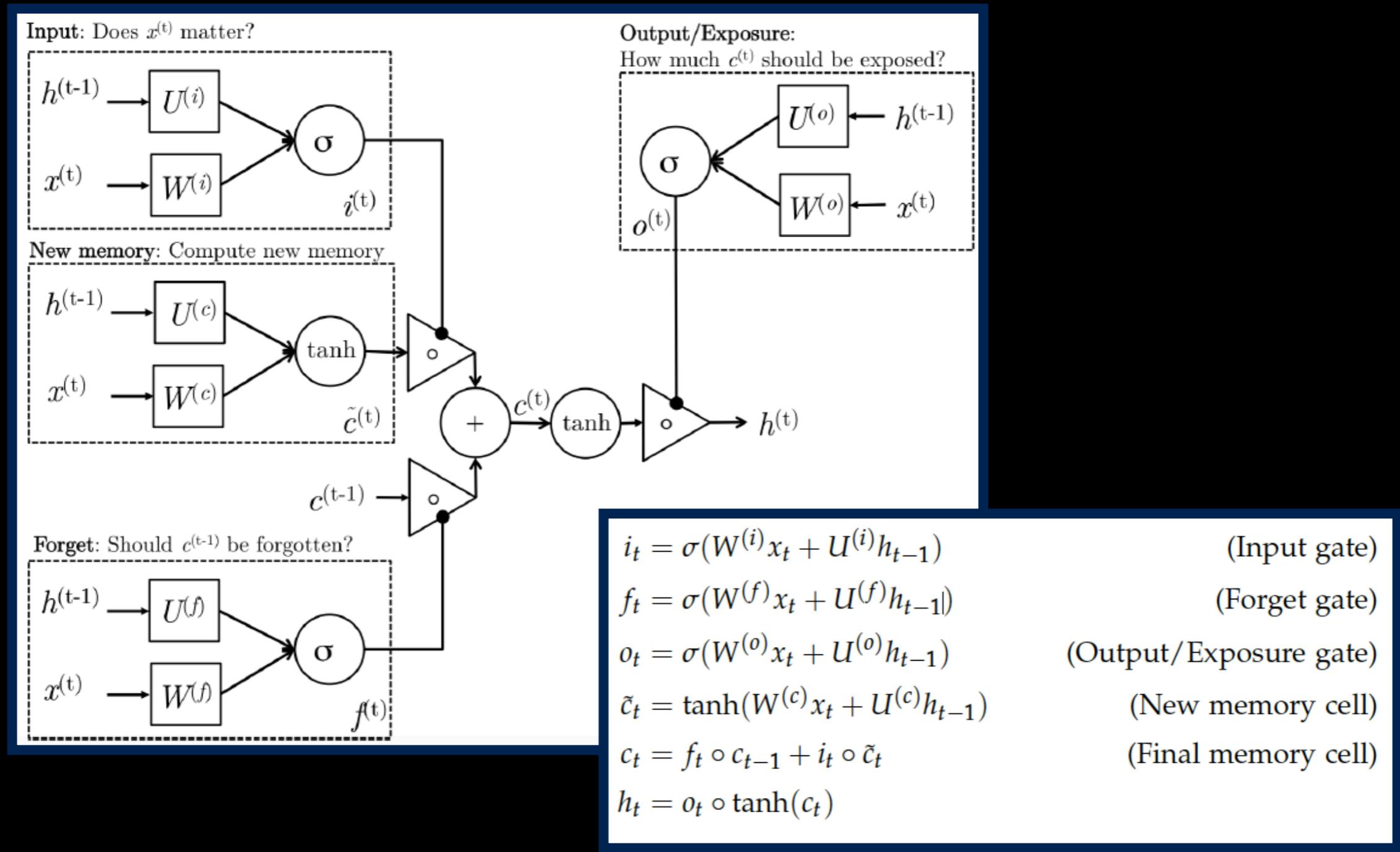
Long-Short-Term-Memories

another type of complex activation unit



Long-Short-Term-Memories

another type of complex activation unit

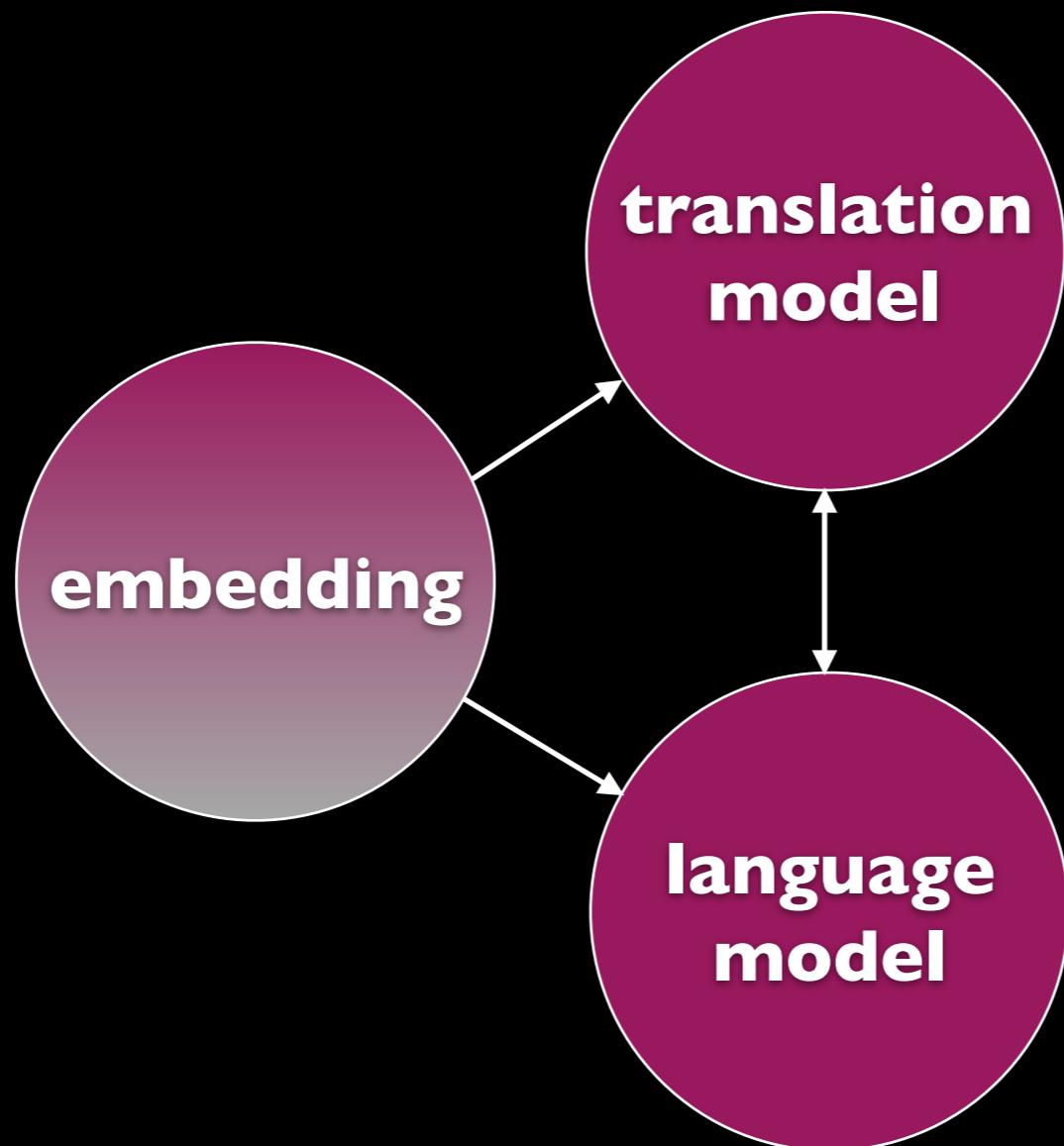


machine translation components

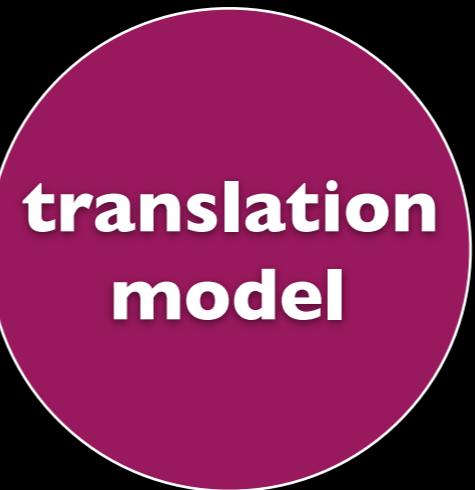


language
model

machine translation components

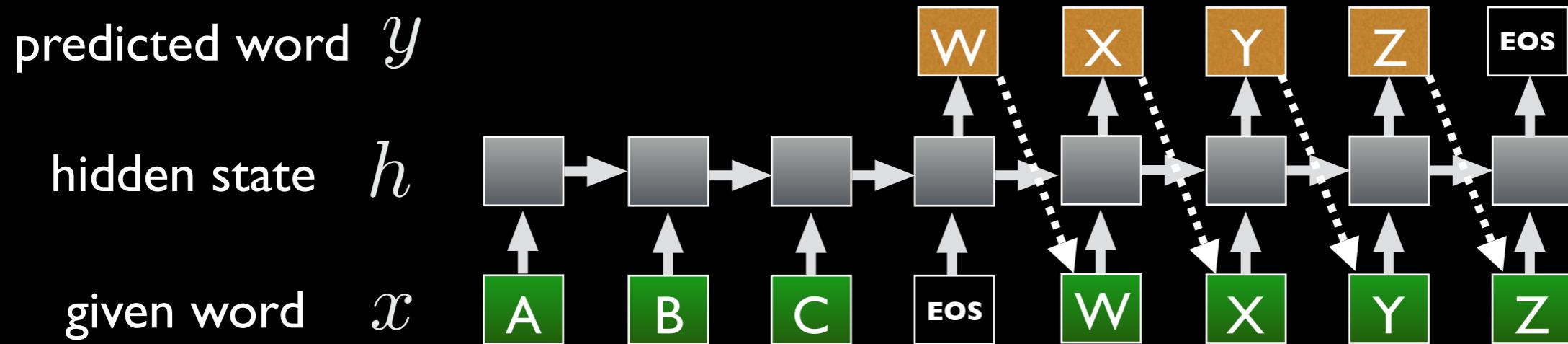


machine translation components



NMT I: sequence-to-sequence with RNN

[Sutskever, 93]



$$P(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} P(y_t | x, y_1, \dots, y_{t-1})$$

$$h_t = \text{sigm}(W^{hx}x_t + W^{hh}h_{t-1})$$

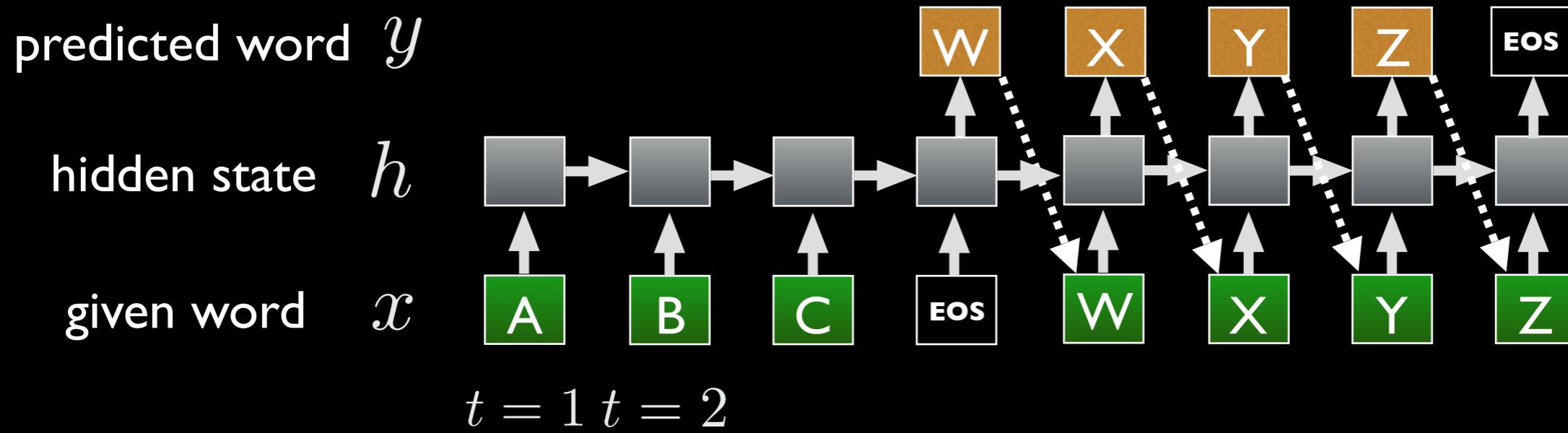
$$y_t = W^{yh}h_t$$

words are in embedded representation

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$

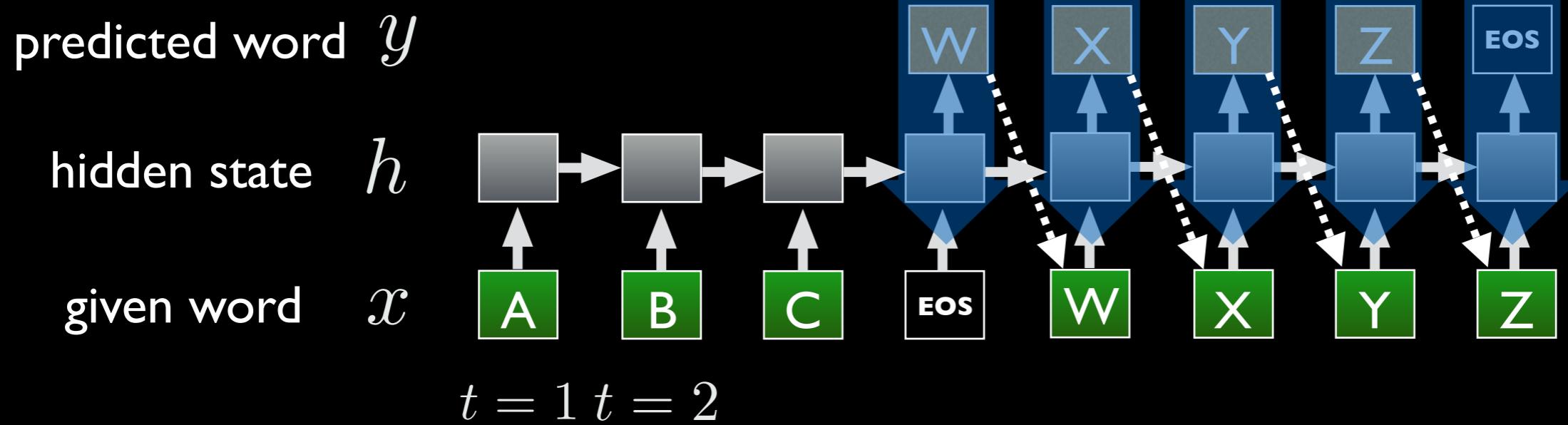


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$

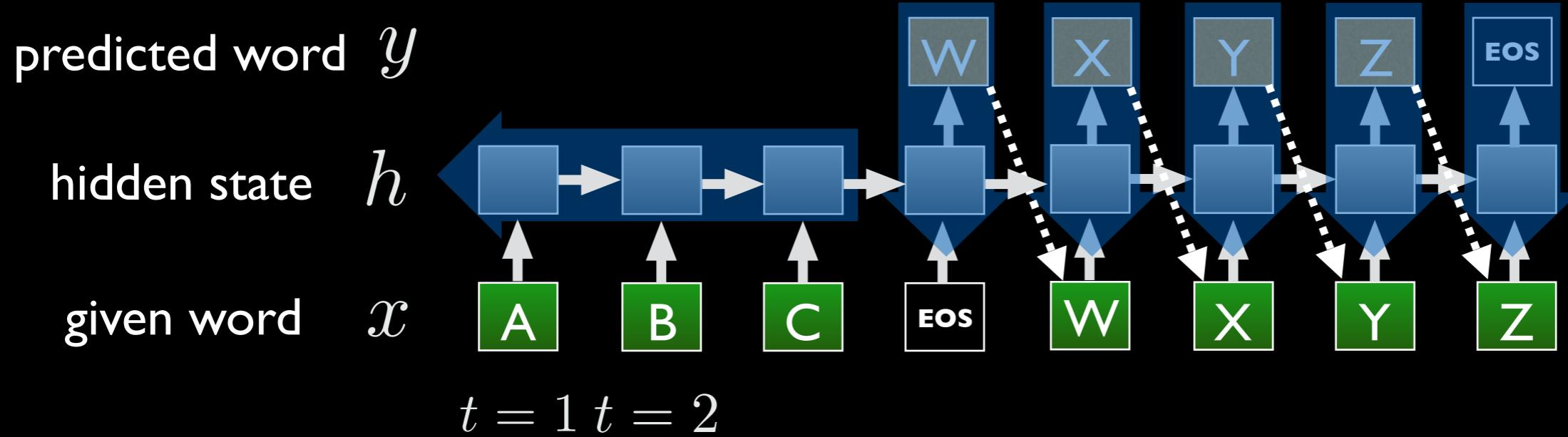


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$

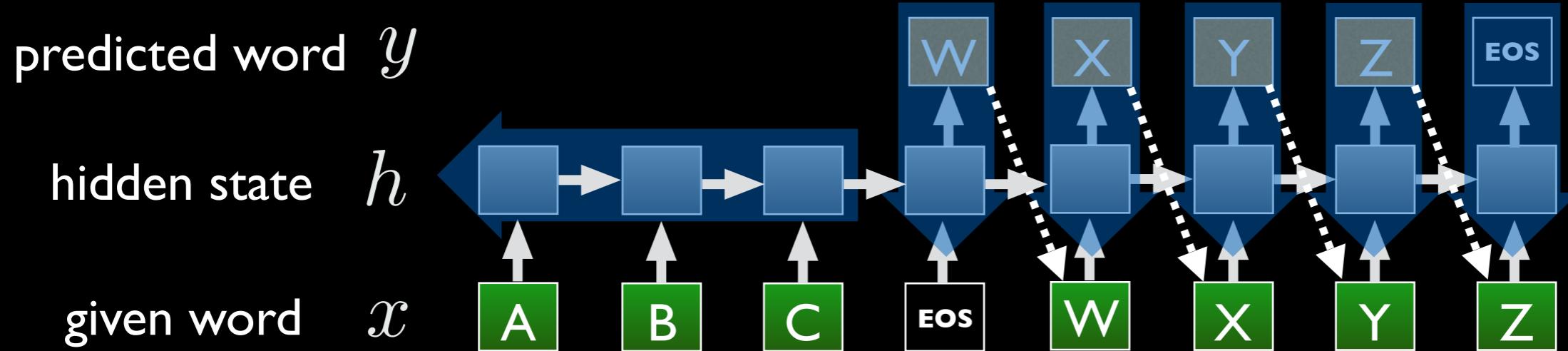


maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



$$t = 1 \quad t = 2$$

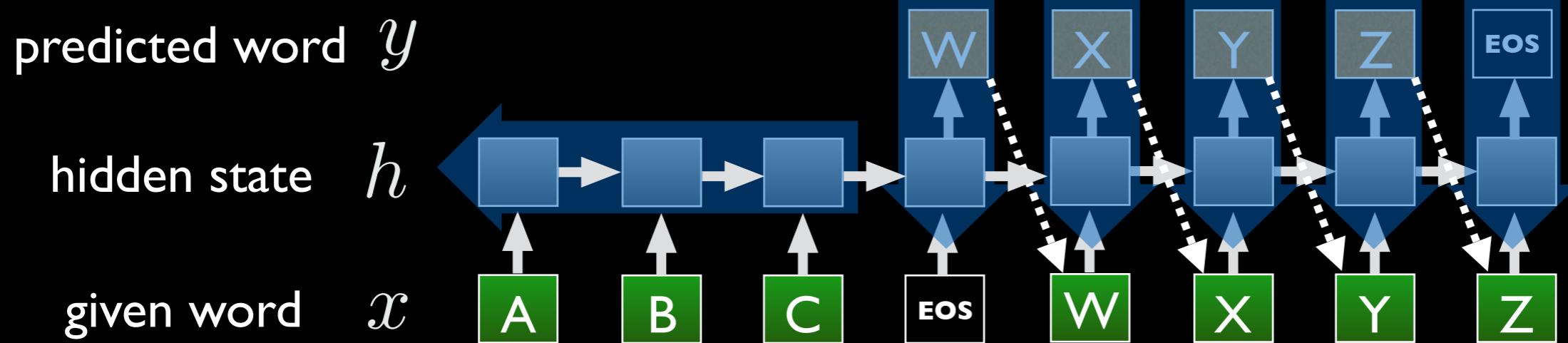
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



$t = 1$ $t = 2$

back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

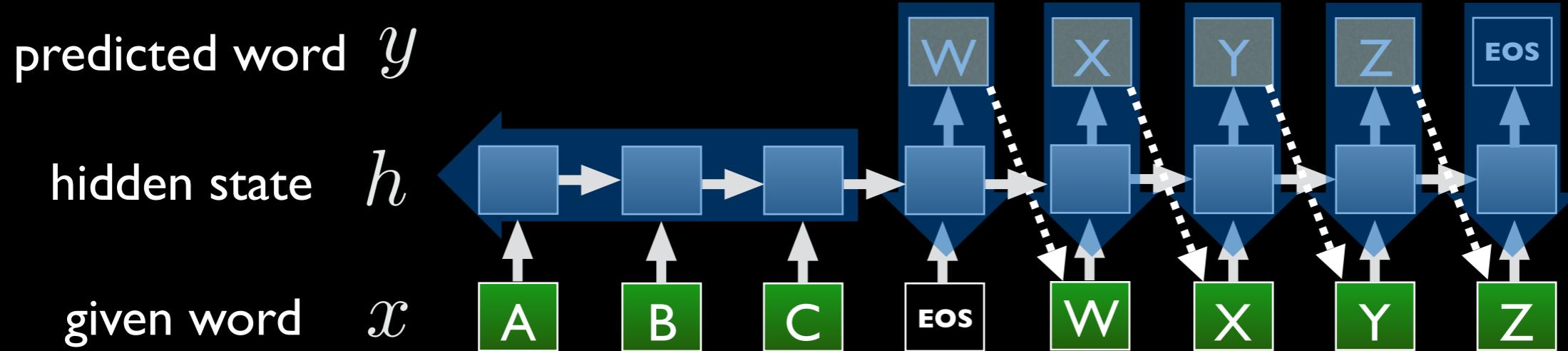
Question #6: better training criterion?

Maximum Likelihood, squared error, MAP, cross-entropy, minimum risk, ..

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



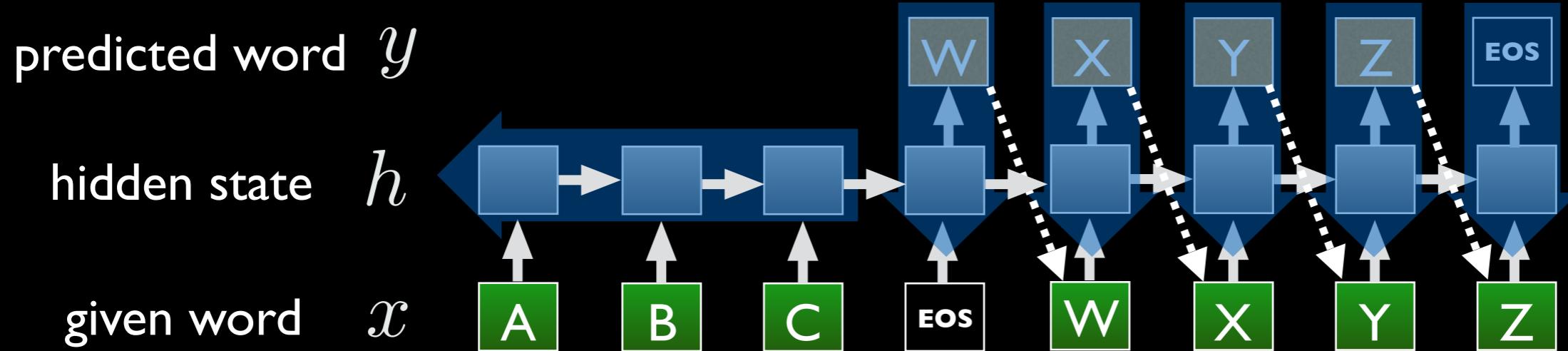
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



$t = 1$ $t = 2$

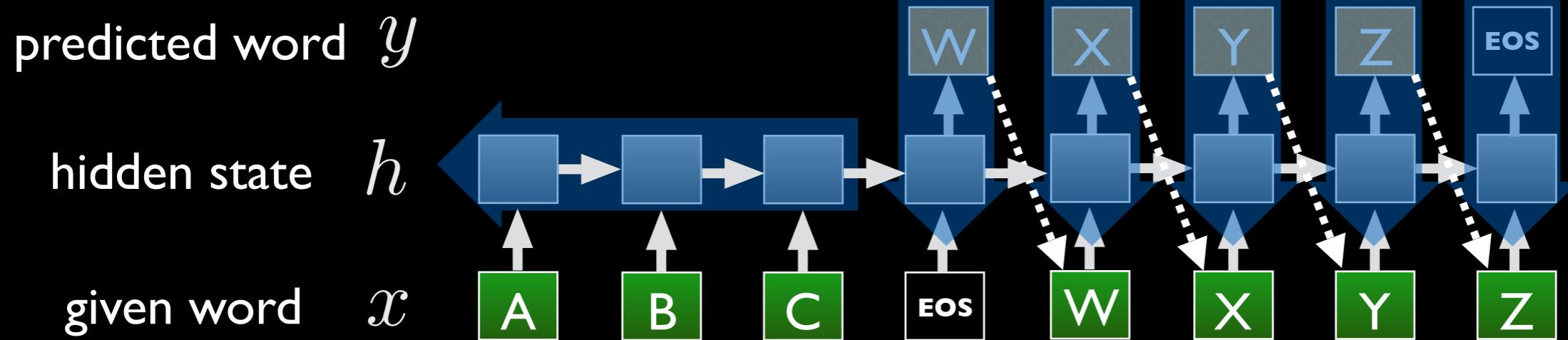
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



$t = 1 \ t = 2$

back propagation operates “end-to-end”

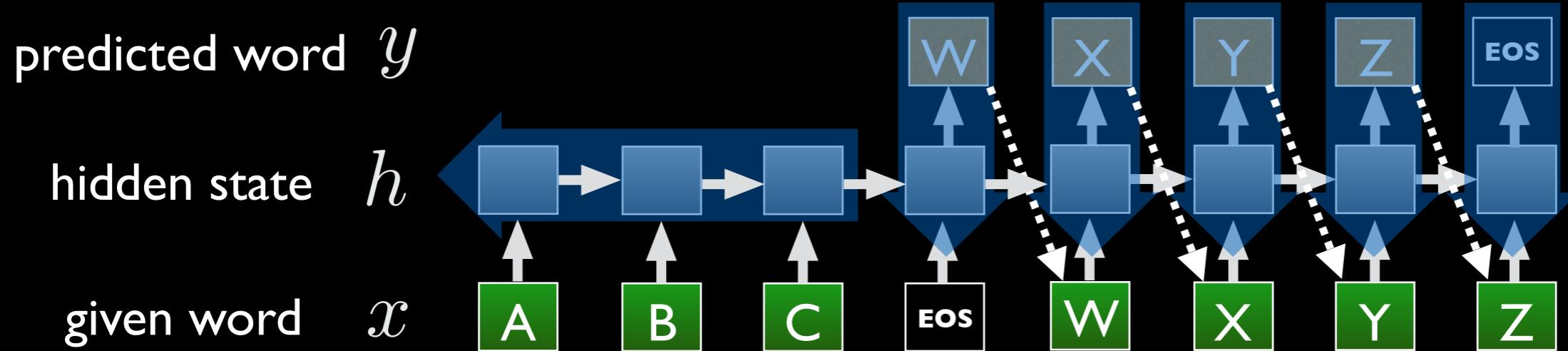
maximize the log probability of a correct translation given the source sentence

Question #7: better training algorithm?
error back propagation, contrastive estimation, ...

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



$$t = 1 \quad t = 2$$

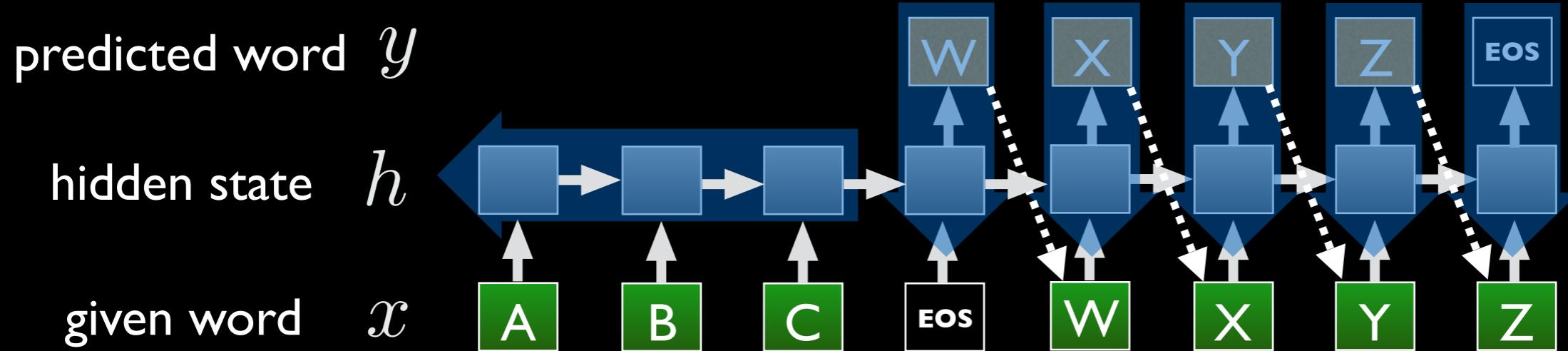
back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

NMT I: training sequence-to-sequence

[Sutskever, 93]

$$\frac{1}{4} [-P(\text{"W"}) - P(\text{"X"}) - P(\text{"Y"}) - P(\text{"Z"})]$$



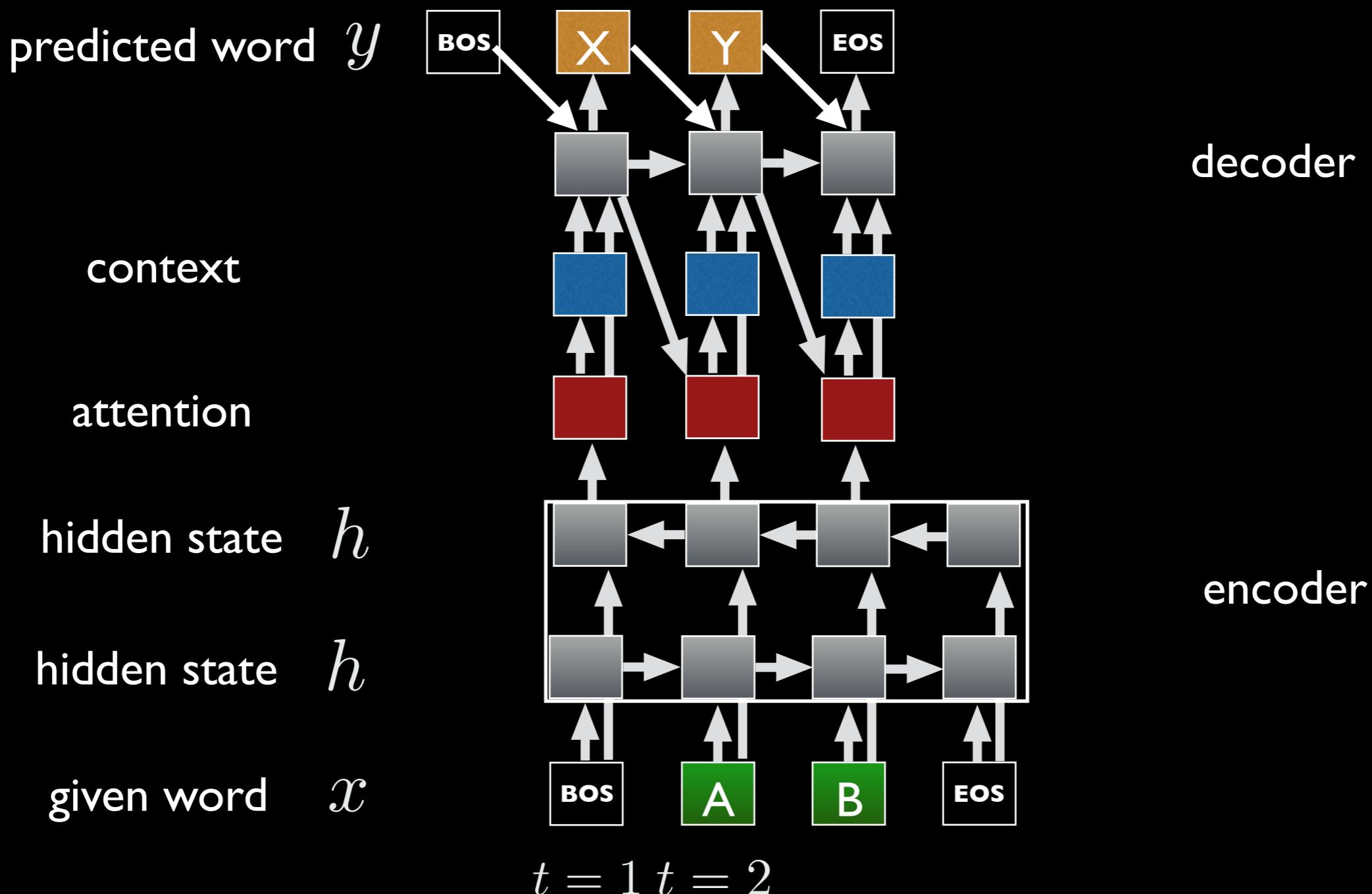
$t = 1$ $t = 2$

back propagation operates “end-to-end”

maximize the log probability of a correct translation given the source sentence

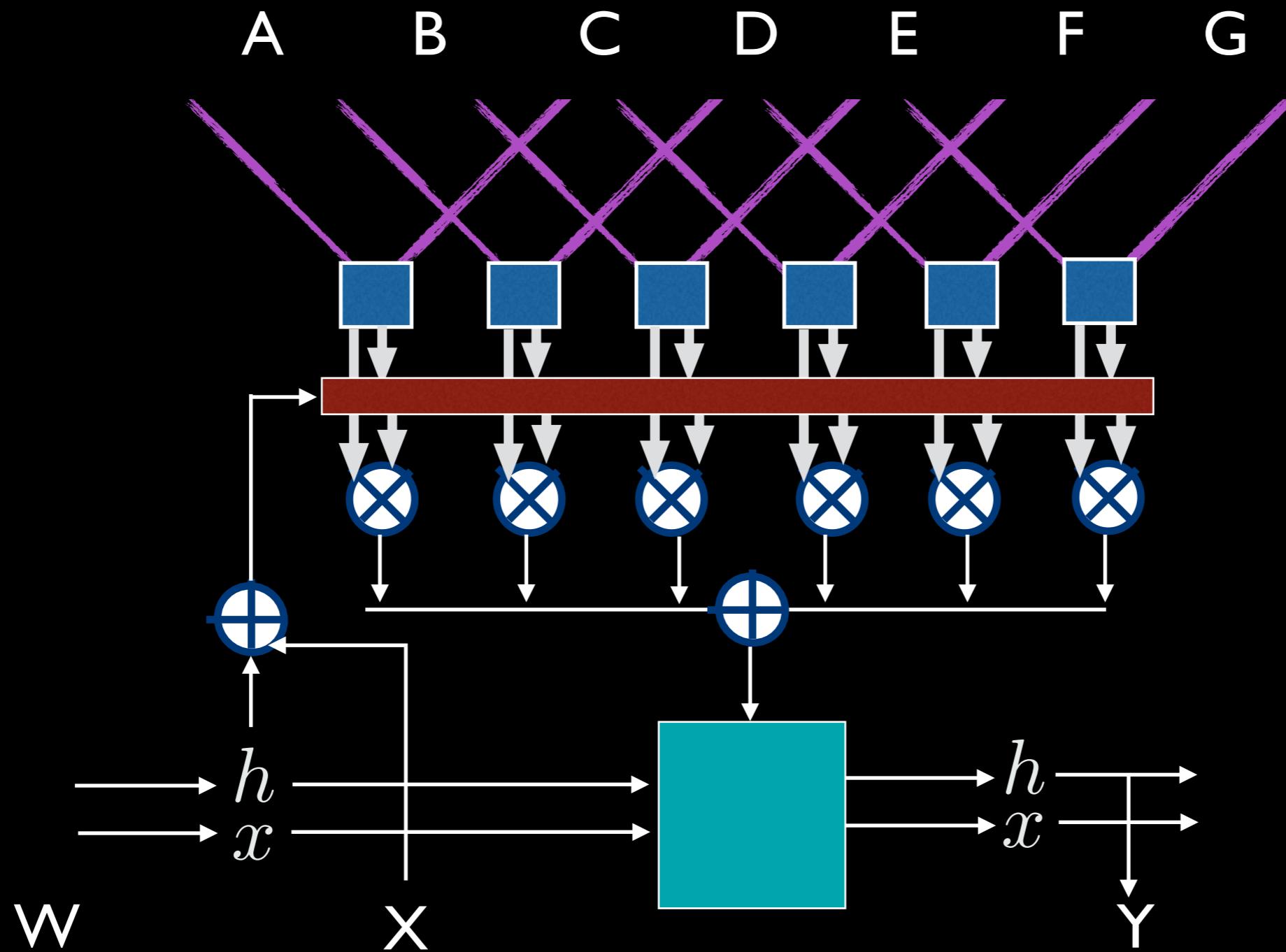
NMT II: encoder & decoder with attention

[Luong et.al., 15]



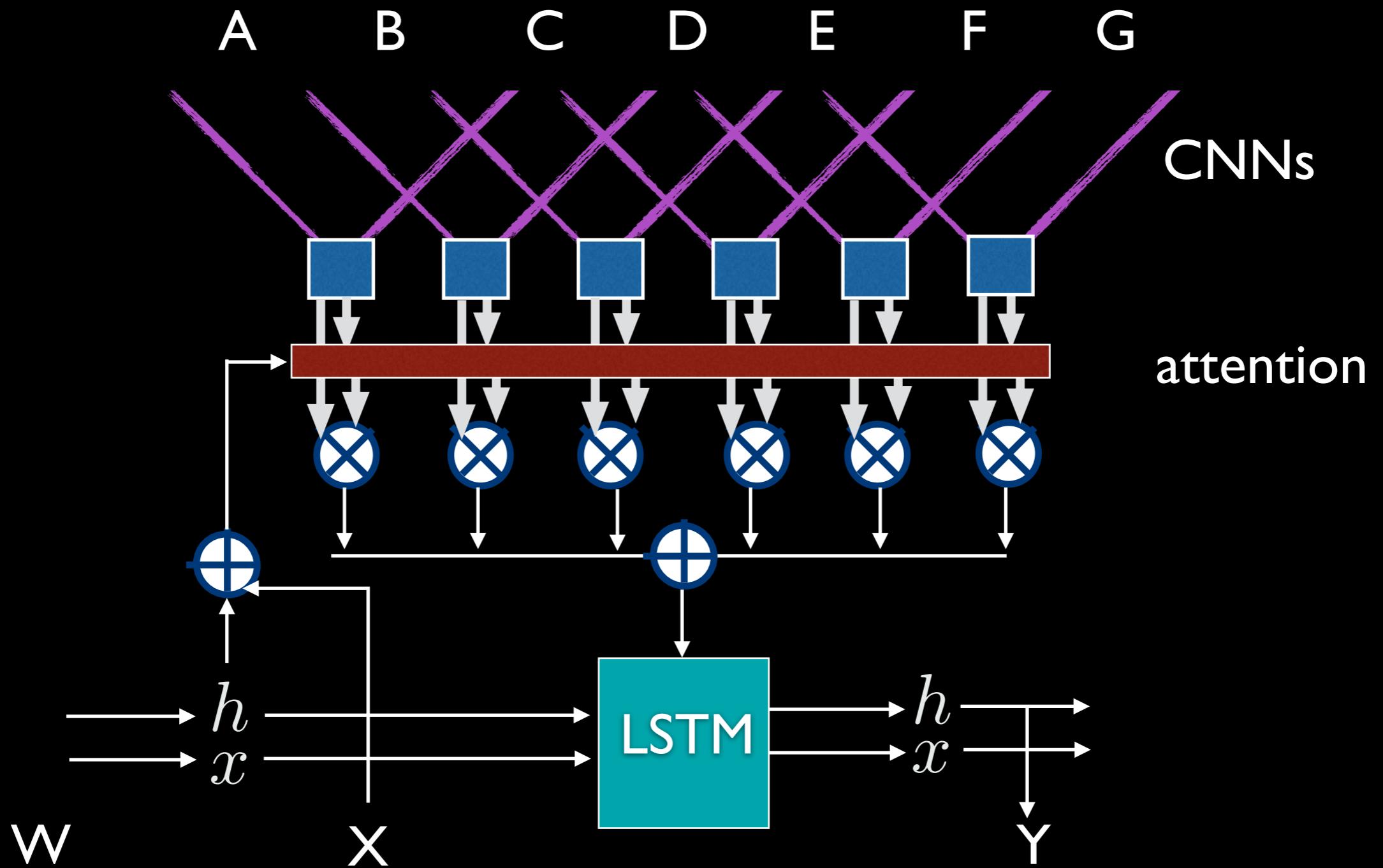
NMT III: multiple models with CNN

[Gehring, I6]



NMT III: multiple models with CNN

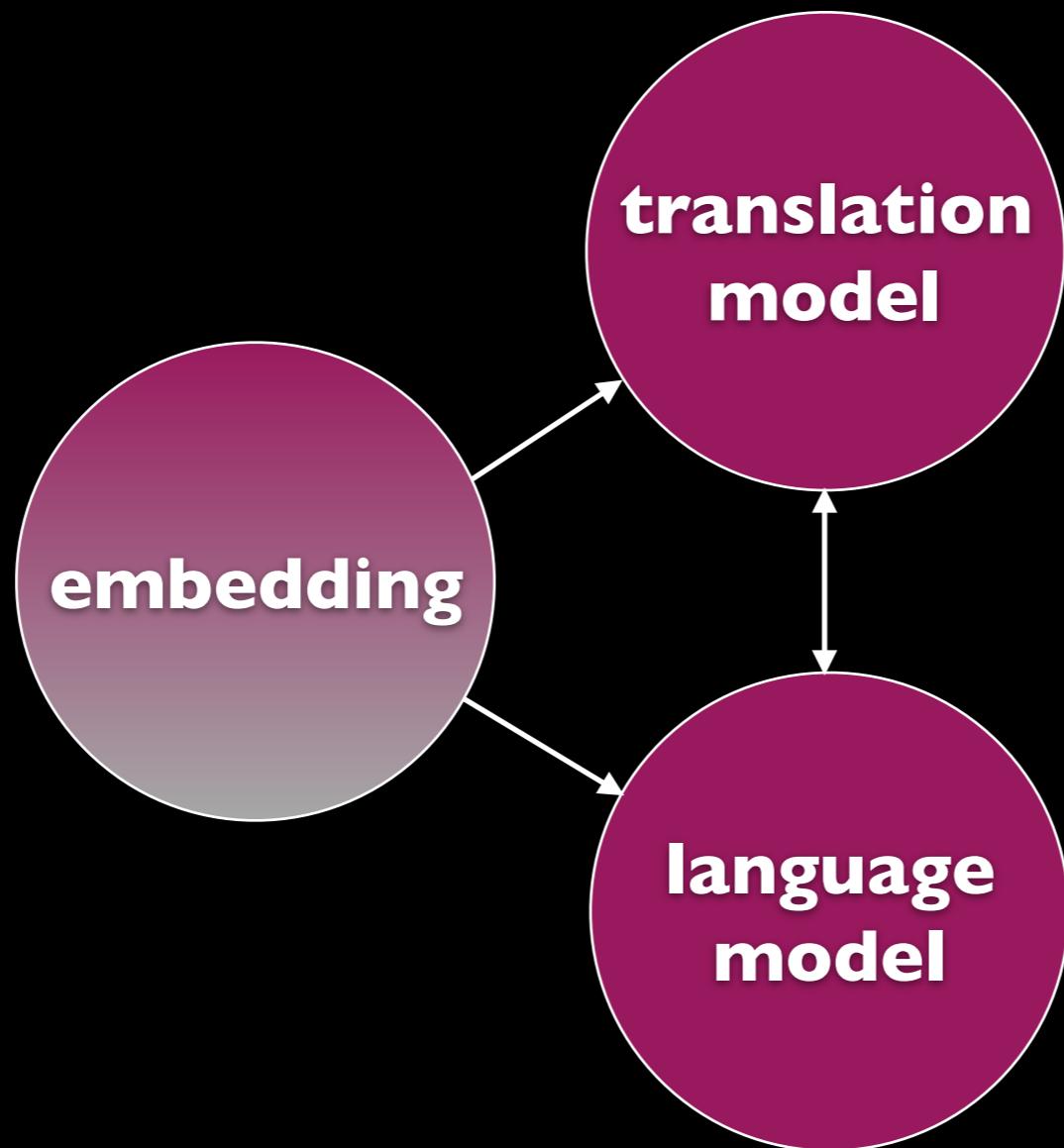
[Gehring, I6]



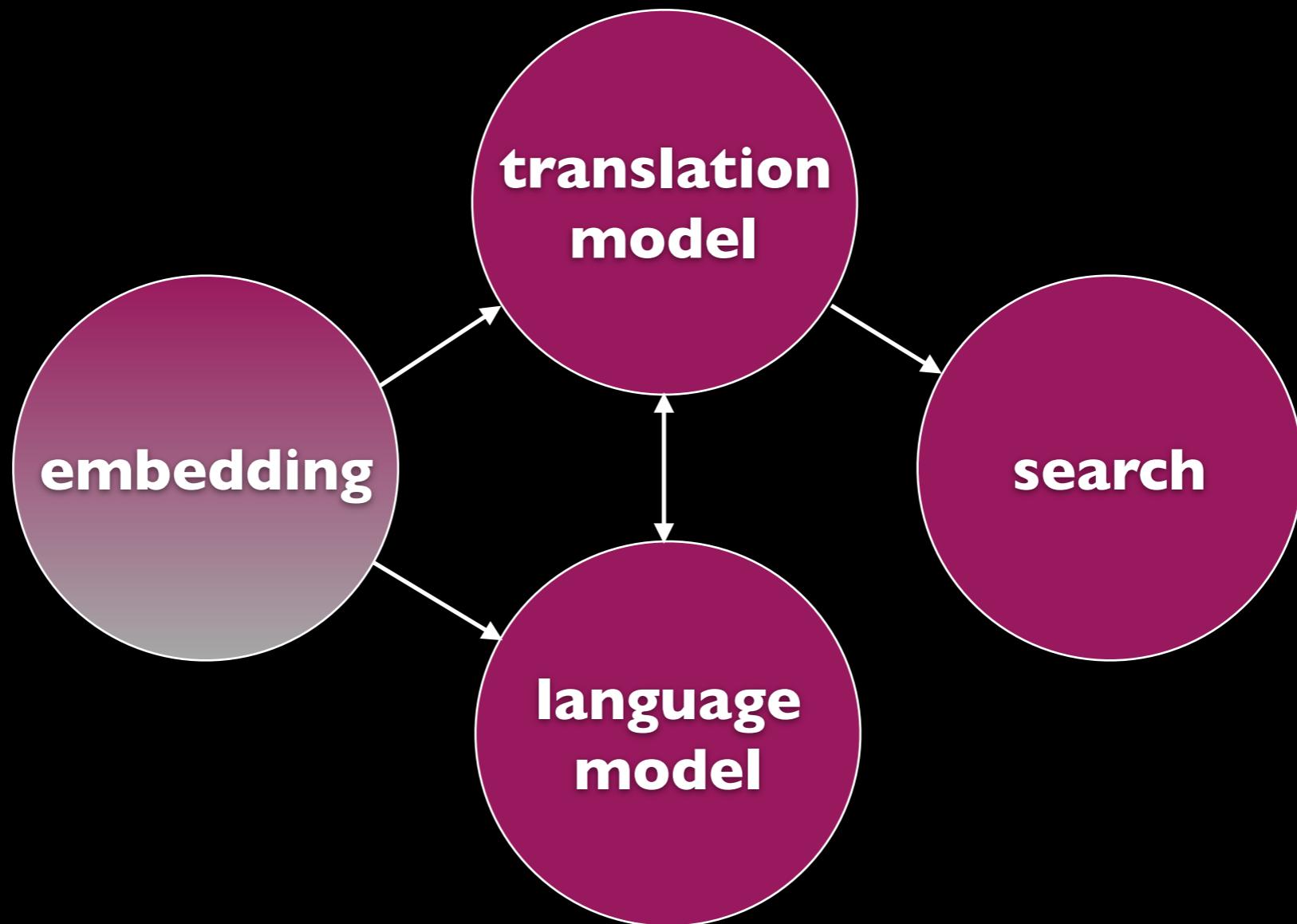
machine translation components



machine translation components



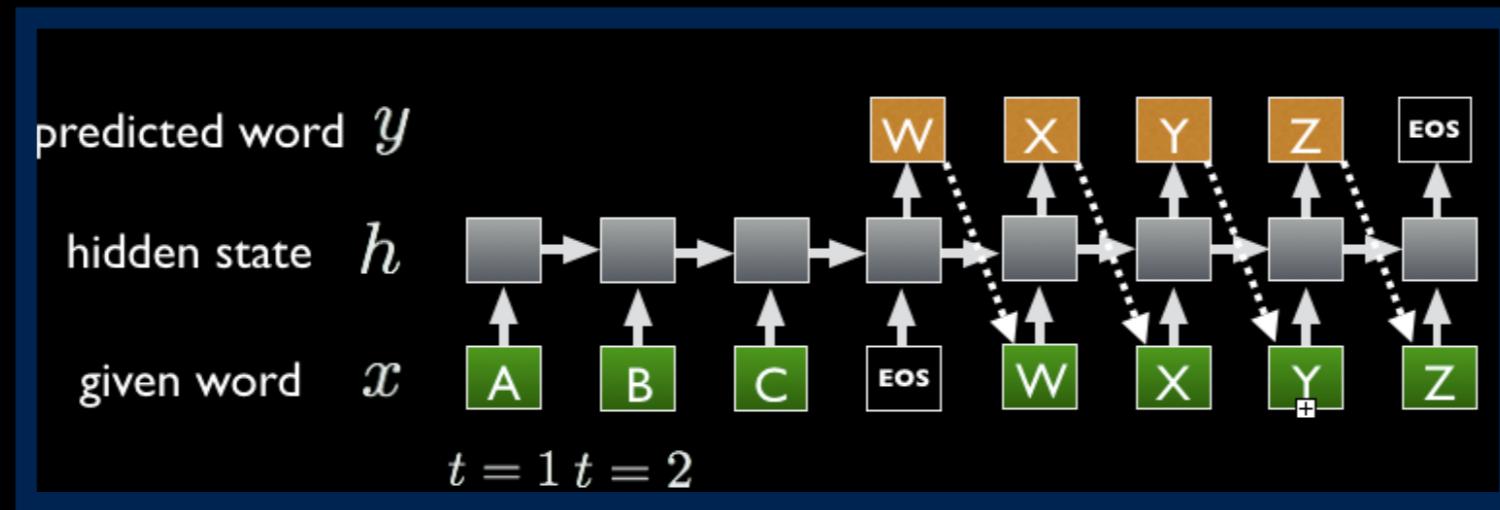
machine translation components



machine translation components



greedy search

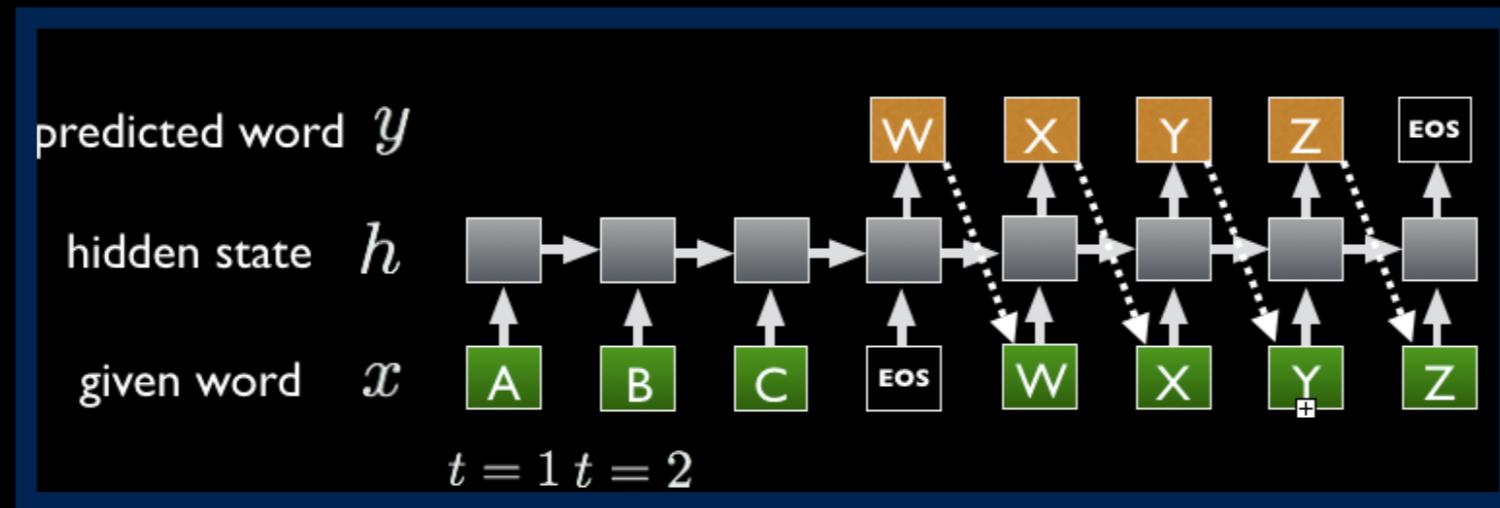


method: take most probable word in each step

problem: no way to undo decisions

- W __
- W X __
- W X Z __ (no way back!)

exhaustive search



ideally: find a translation that maximize

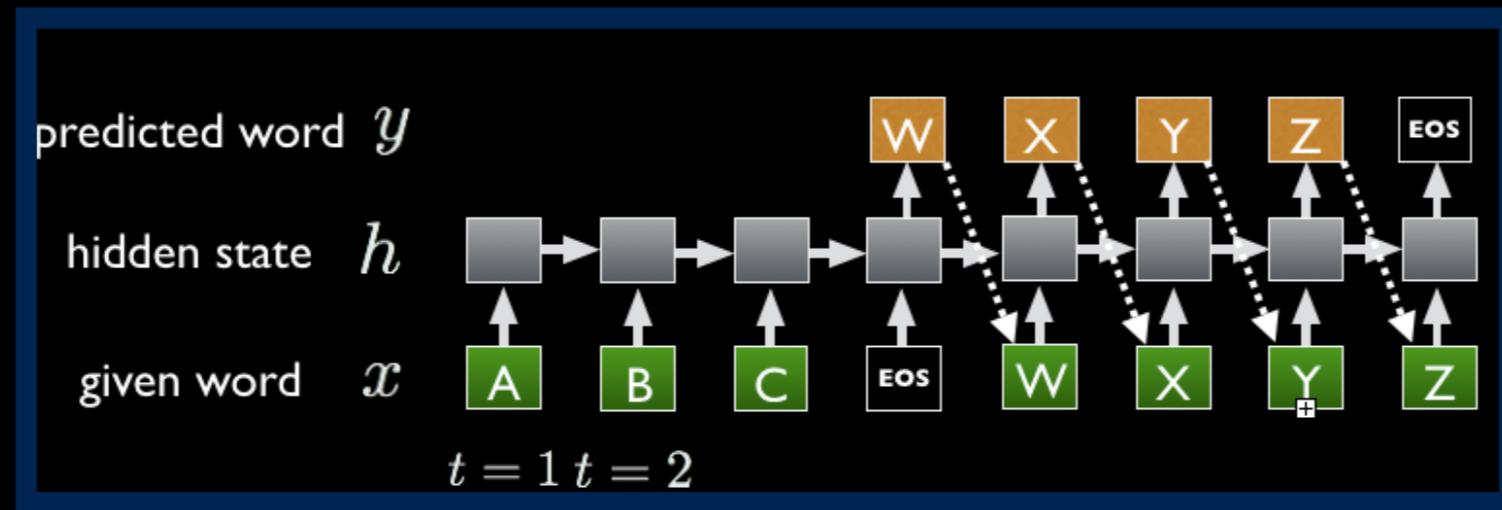
$$P(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} P(y_t | x, y_1, \dots, y_{t-1})$$

method: compute all possible sequences y

problem: expensive

each step tracking V (vocabulary) words
complexity $O(V^T)$

beam search



$$\text{score}(y_1, \dots, y_t) = \log P_{LM}(y_1, \dots, y_t | x)$$

$$= \sum_{i=1}^t \log P_{LM}(y_i | y_1, \dots, y_{i-1}, x)$$

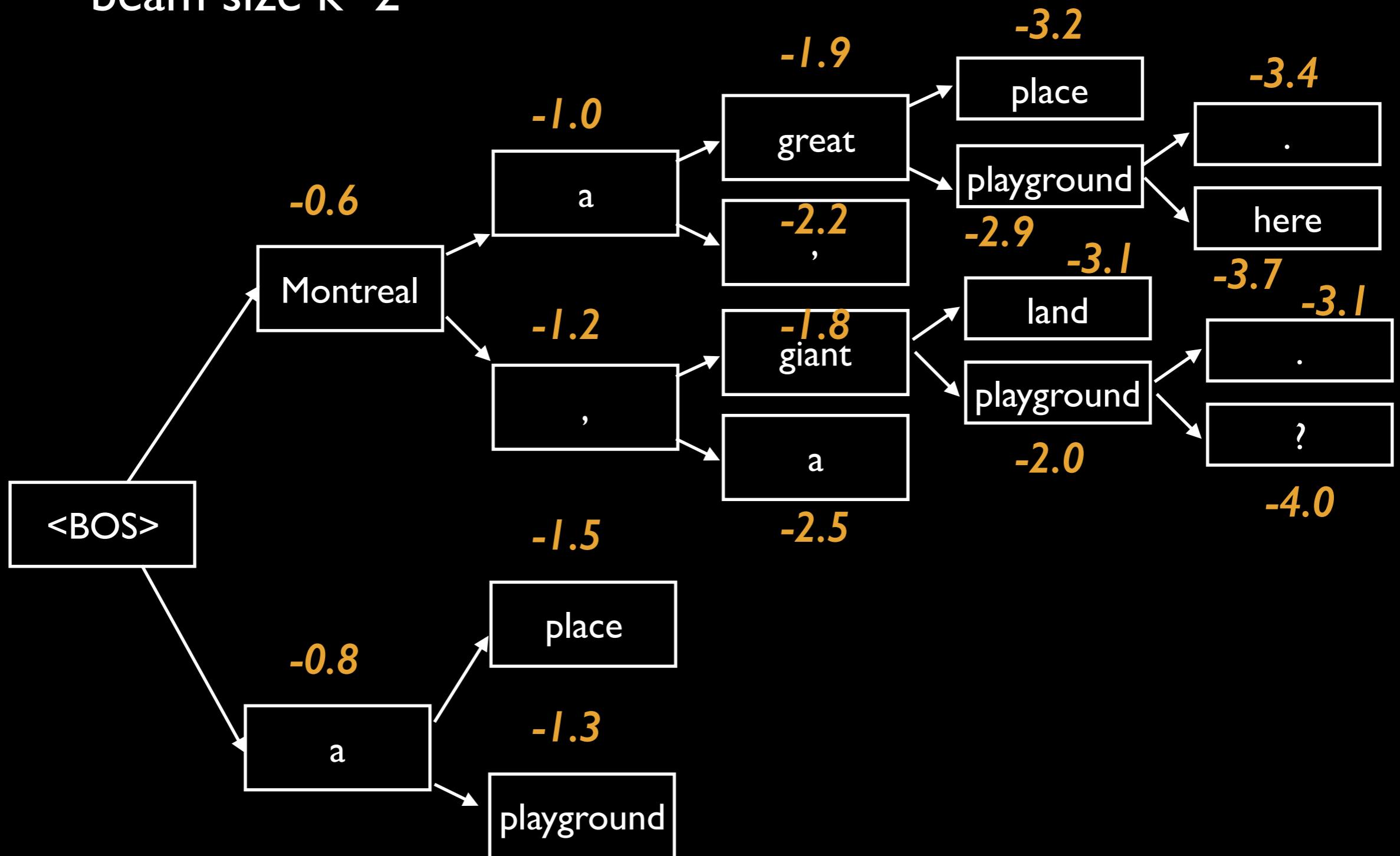
method: on each search step, keep track of the k most probable (higher score) partial translations

problem: no guarantee for optimal solution

efficient!

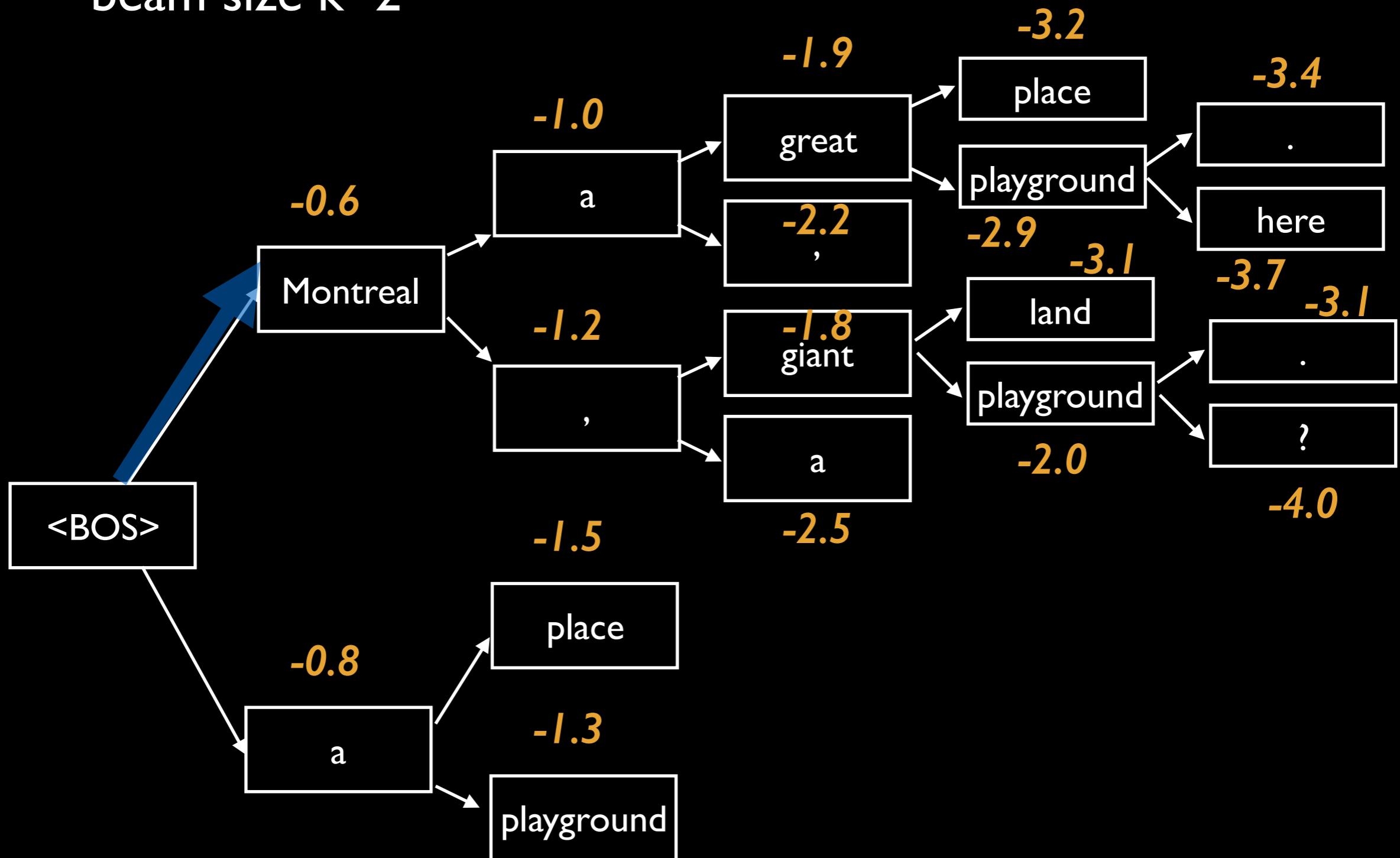
beam search example

beam size k=2



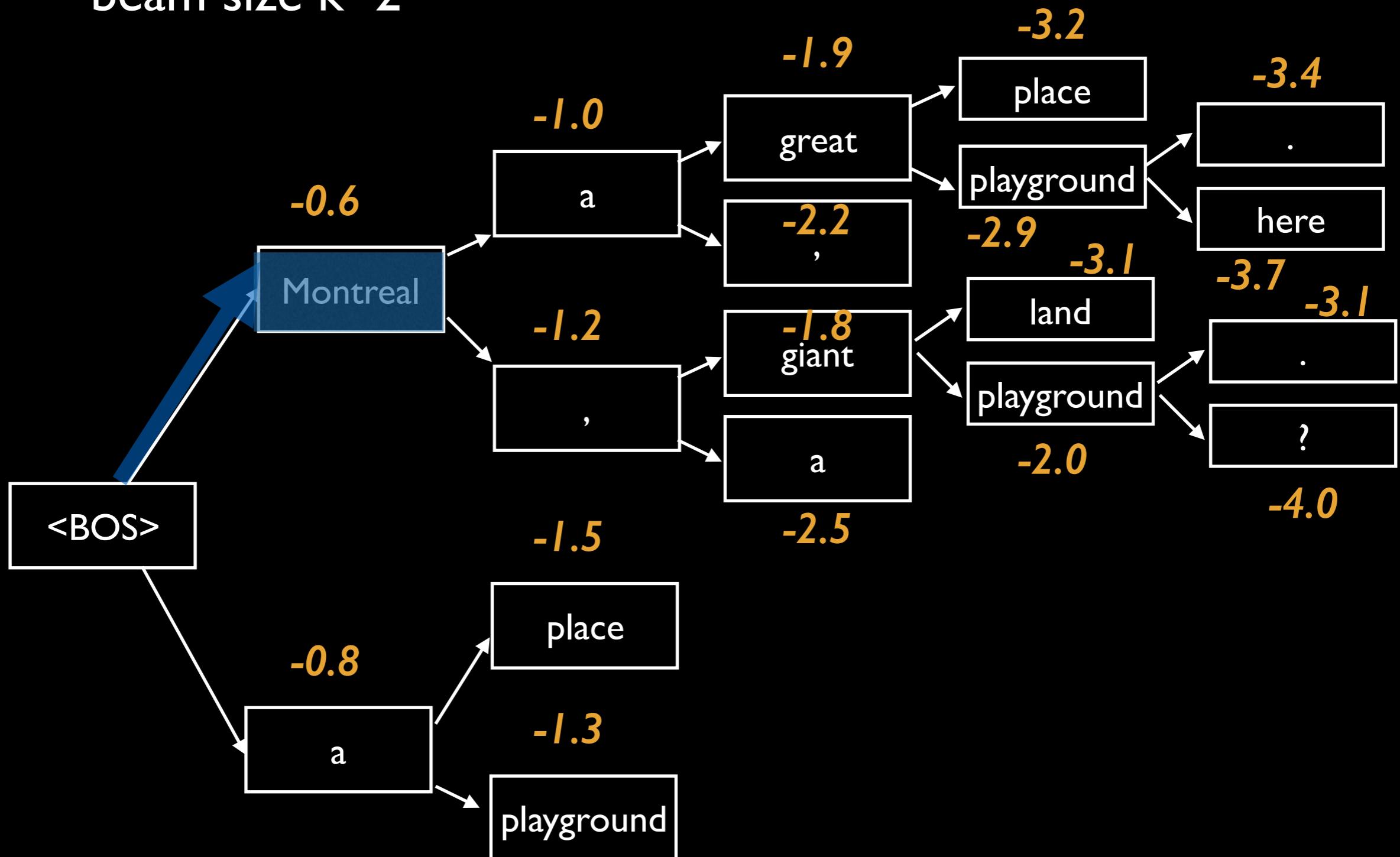
beam search example

beam size k=2



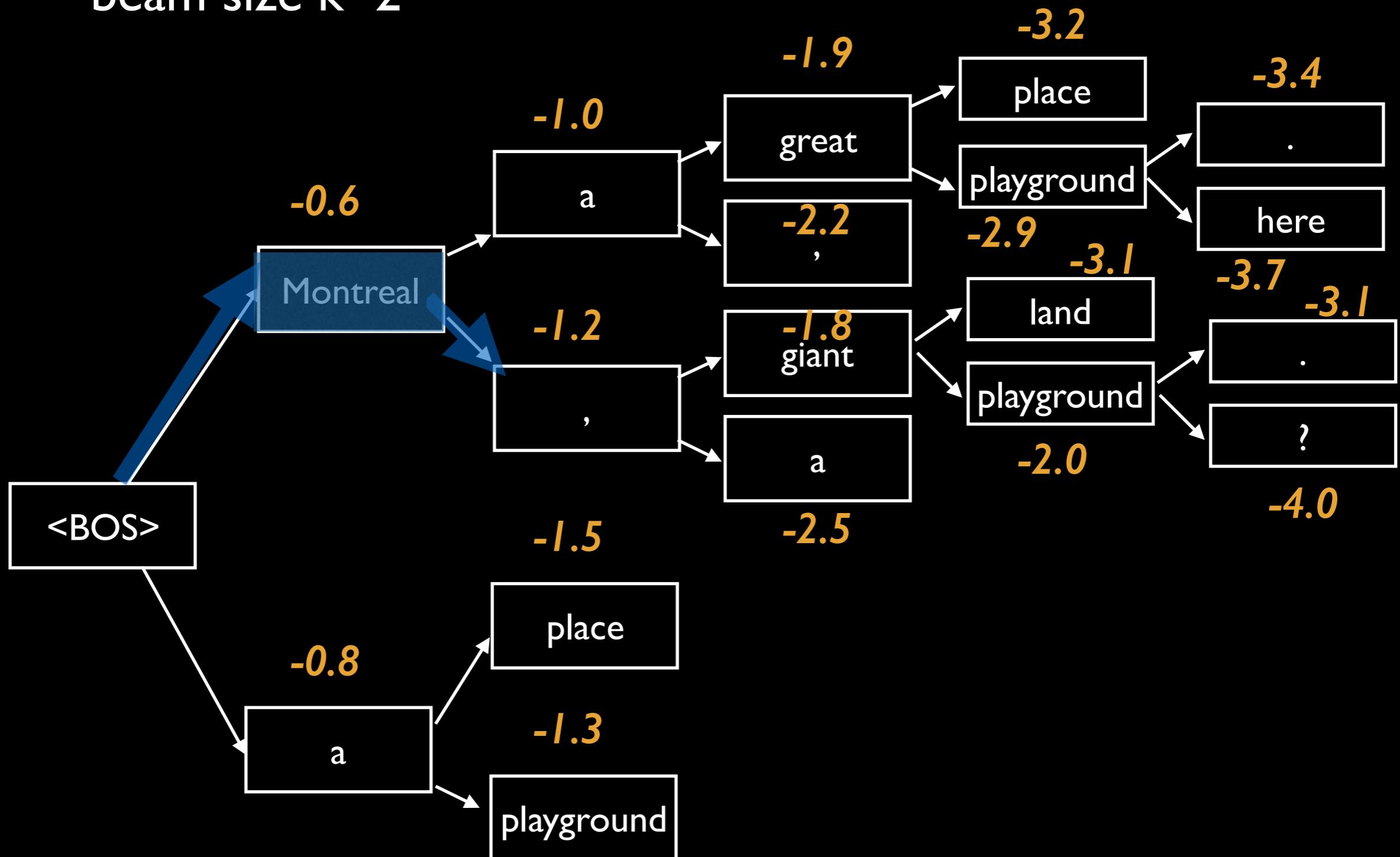
beam search example

beam size k=2



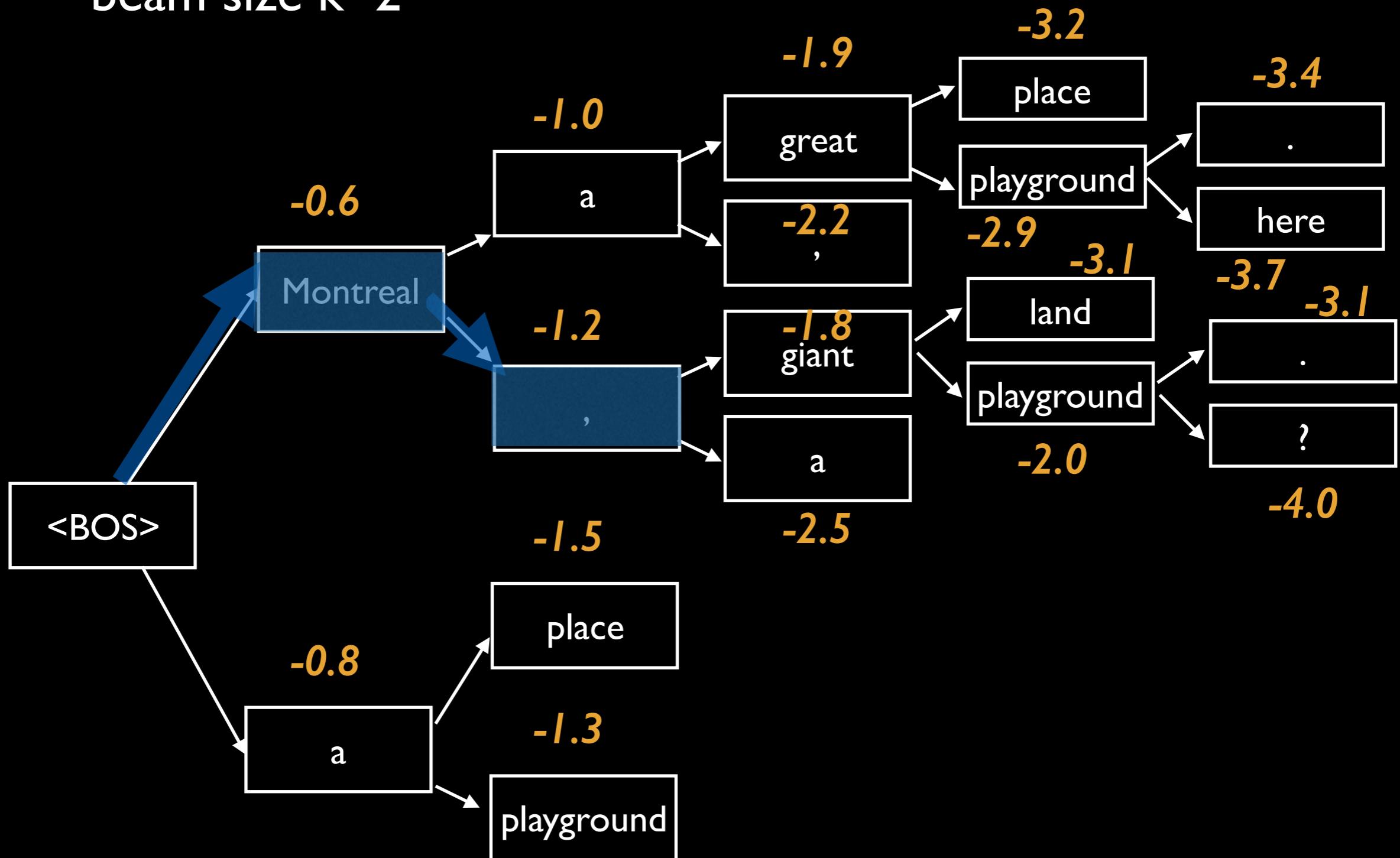
beam search example

beam size k=2



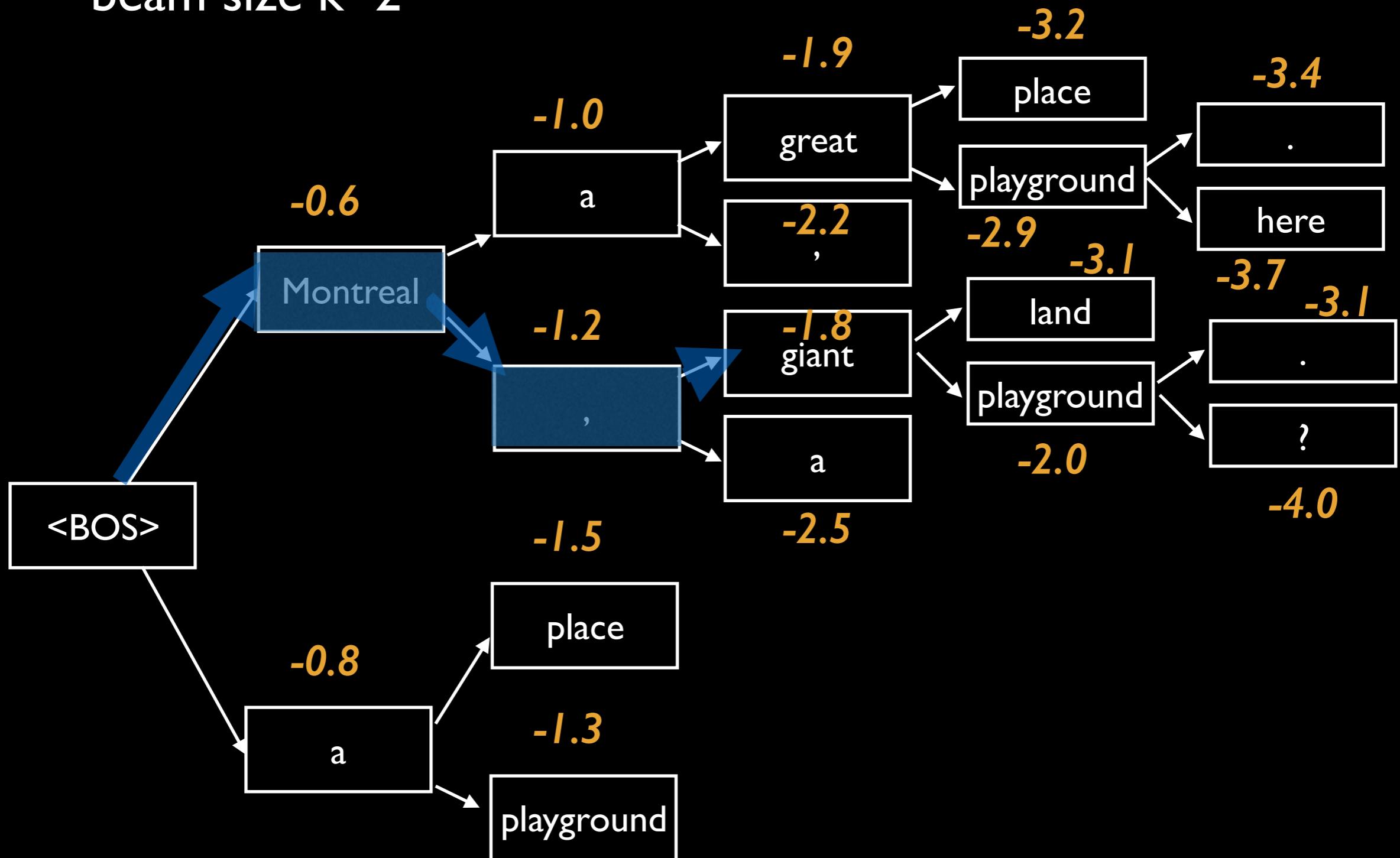
beam search example

beam size k=2



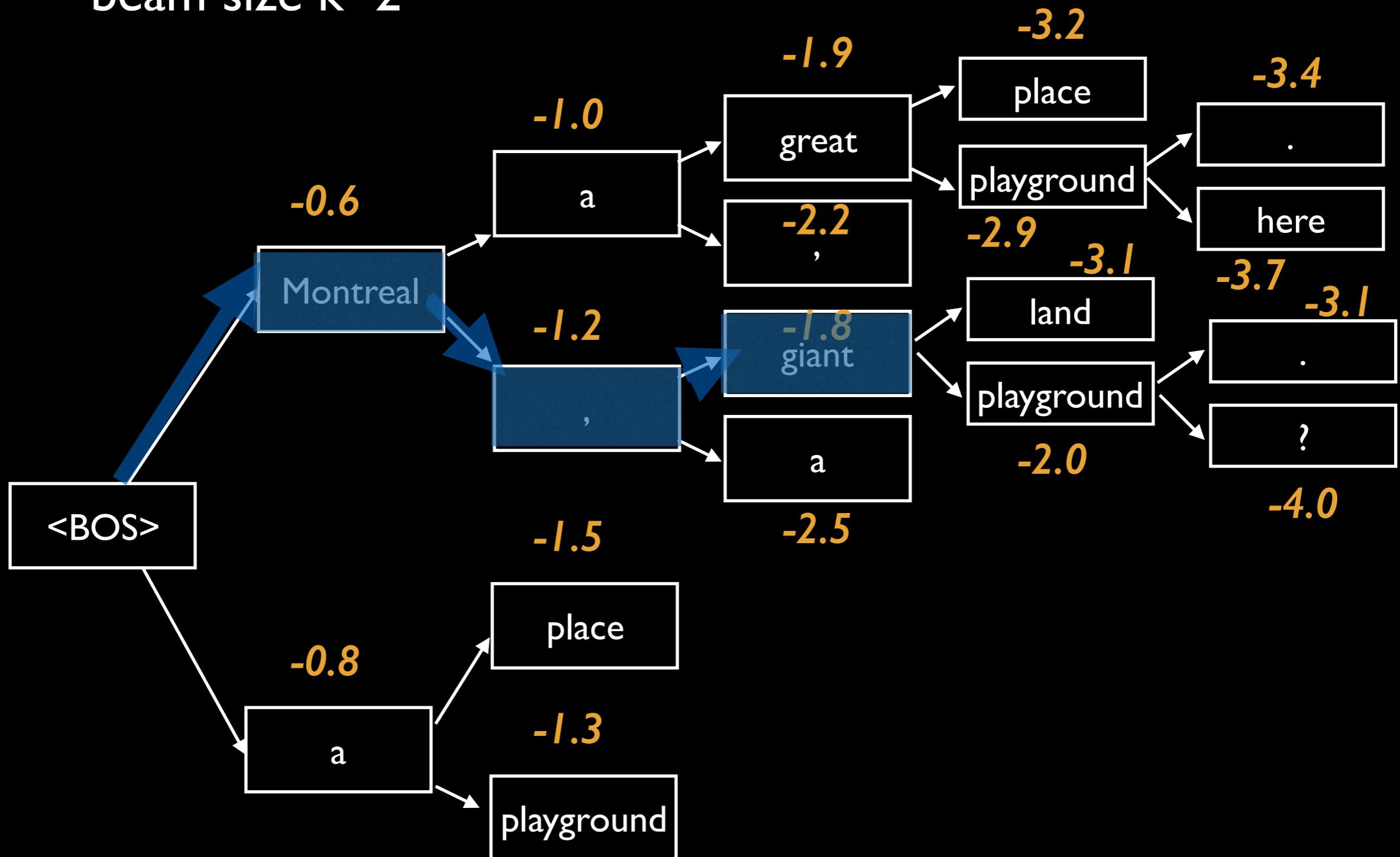
beam search example

beam size k=2



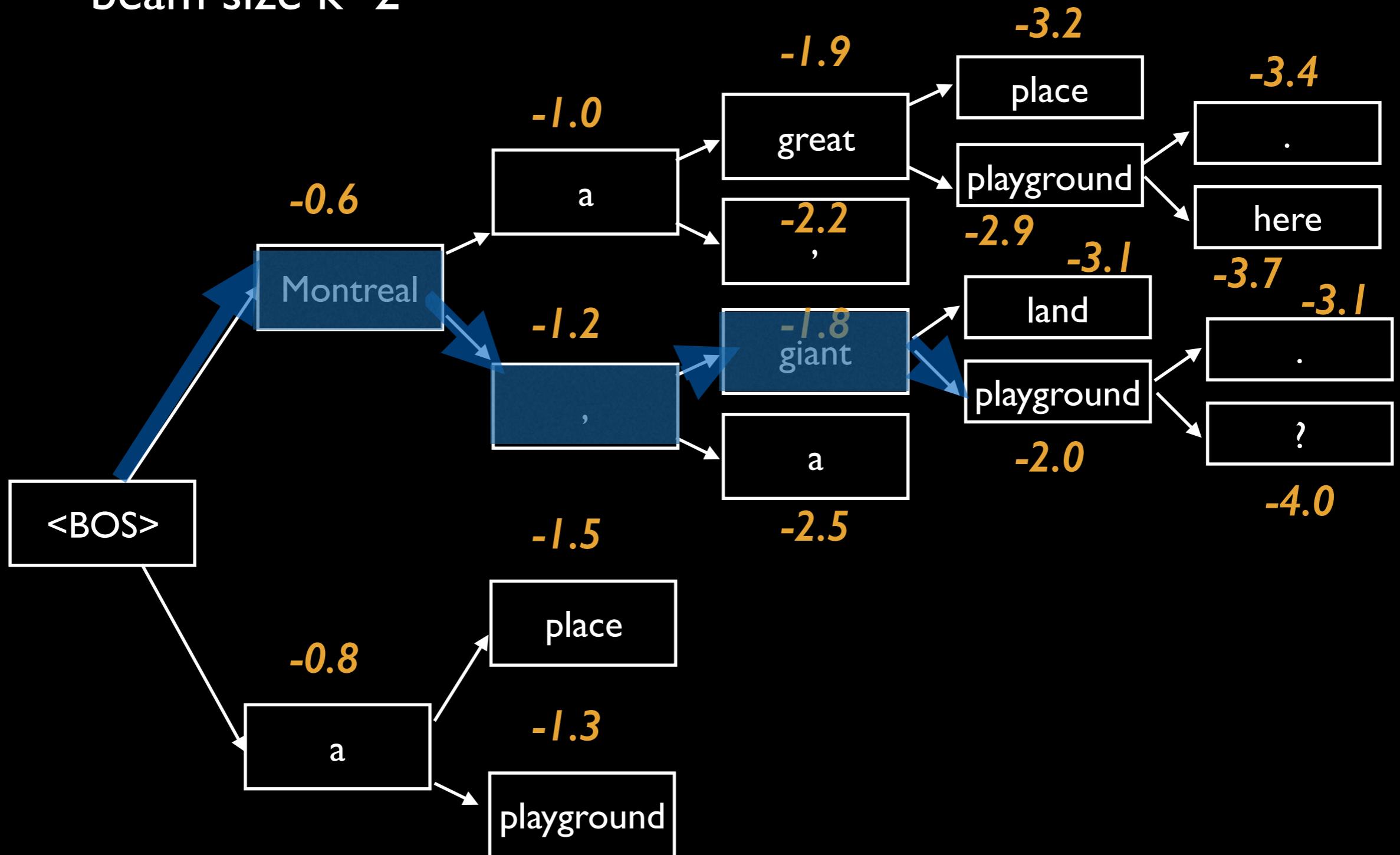
beam search example

beam size k=2



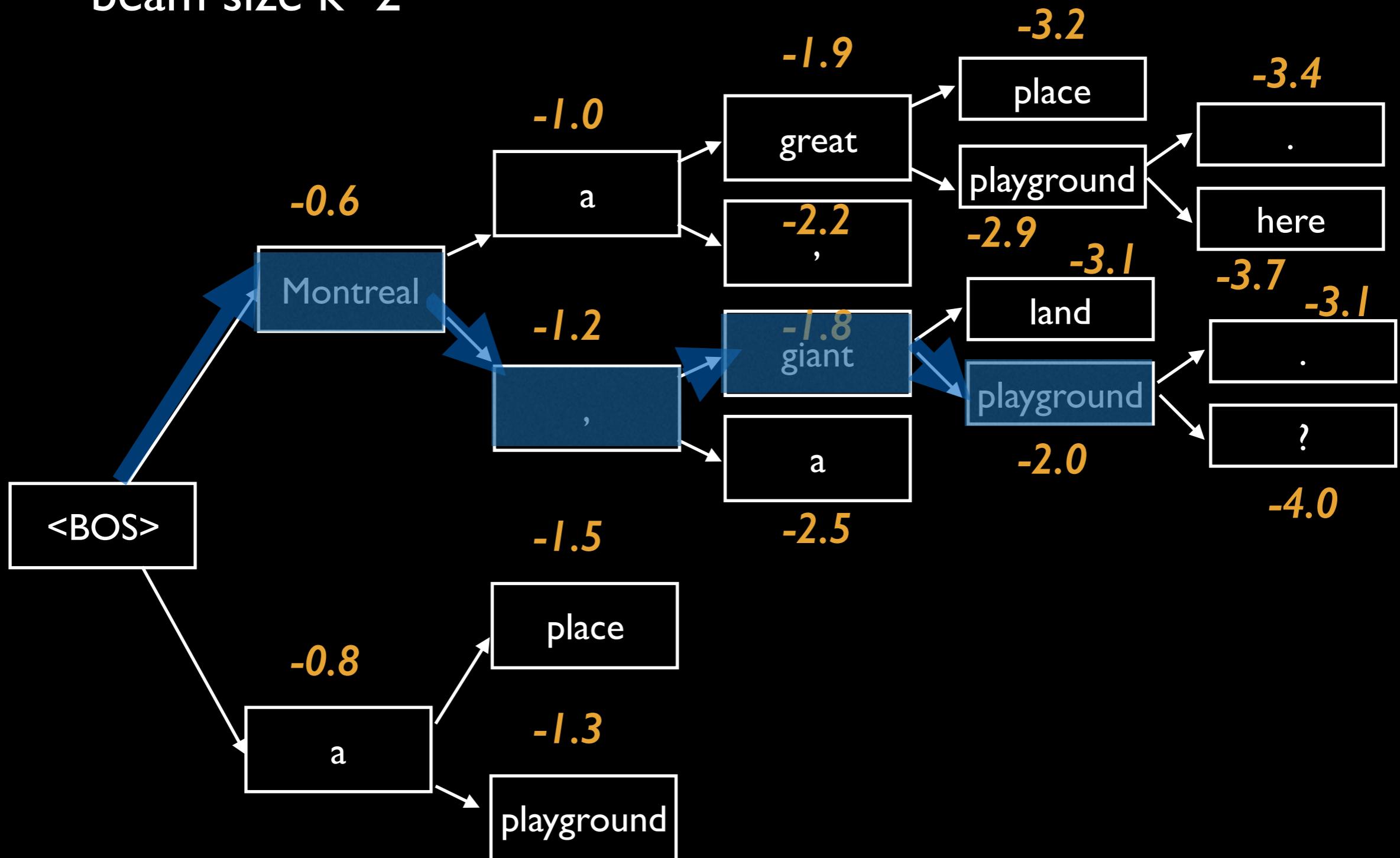
beam search example

beam size k=2



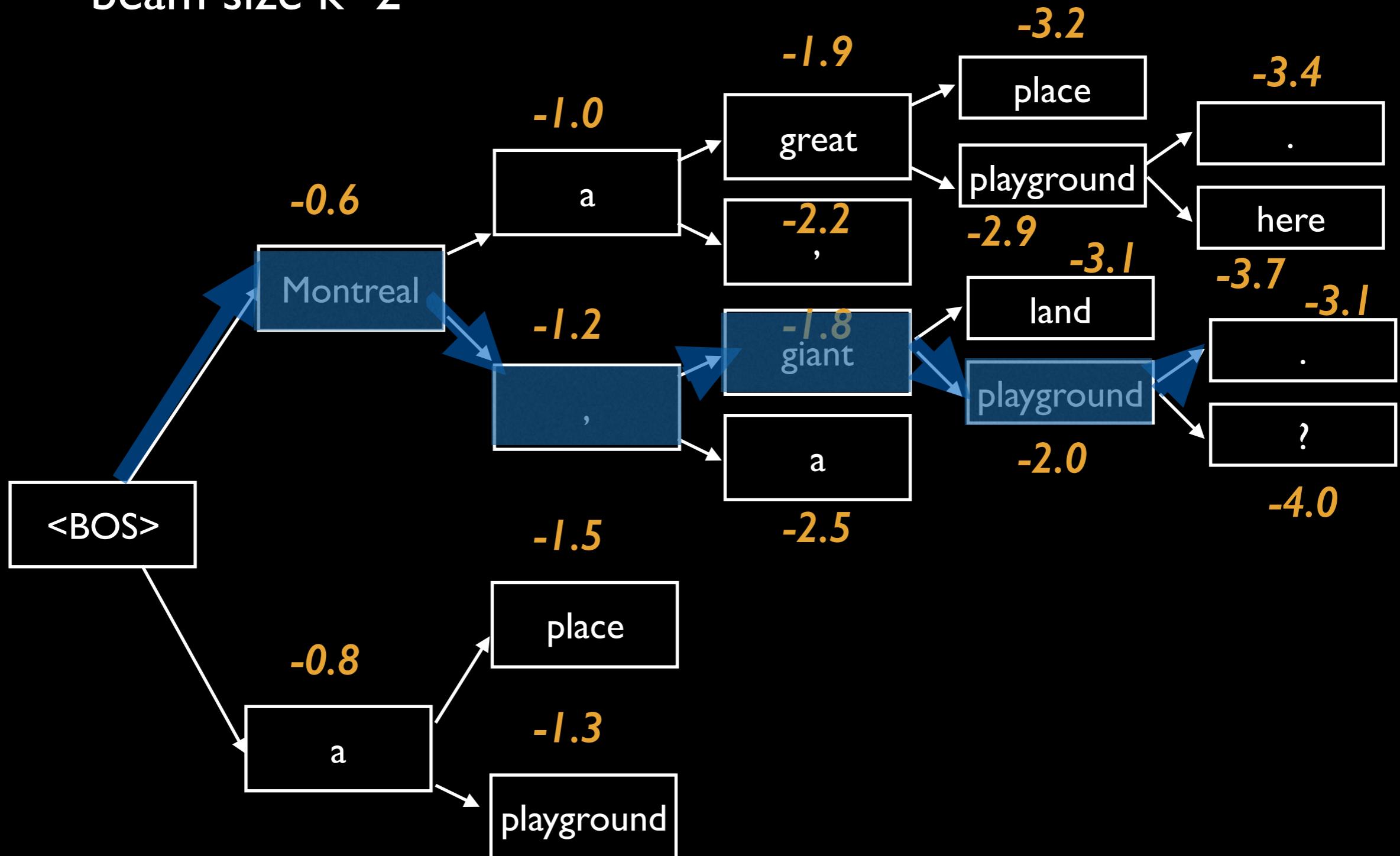
beam search example

beam size k=2



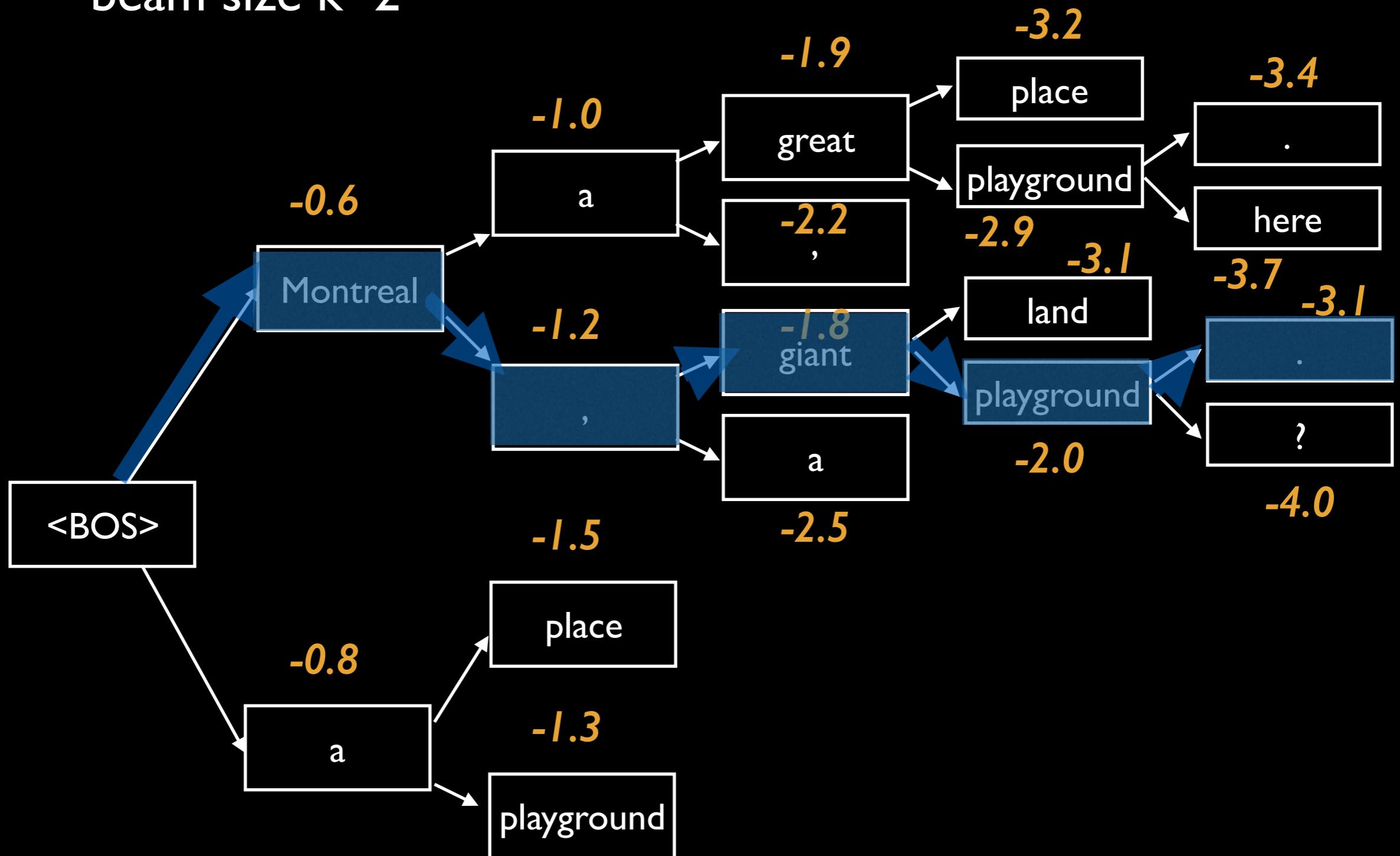
beam search example

beam size k=2



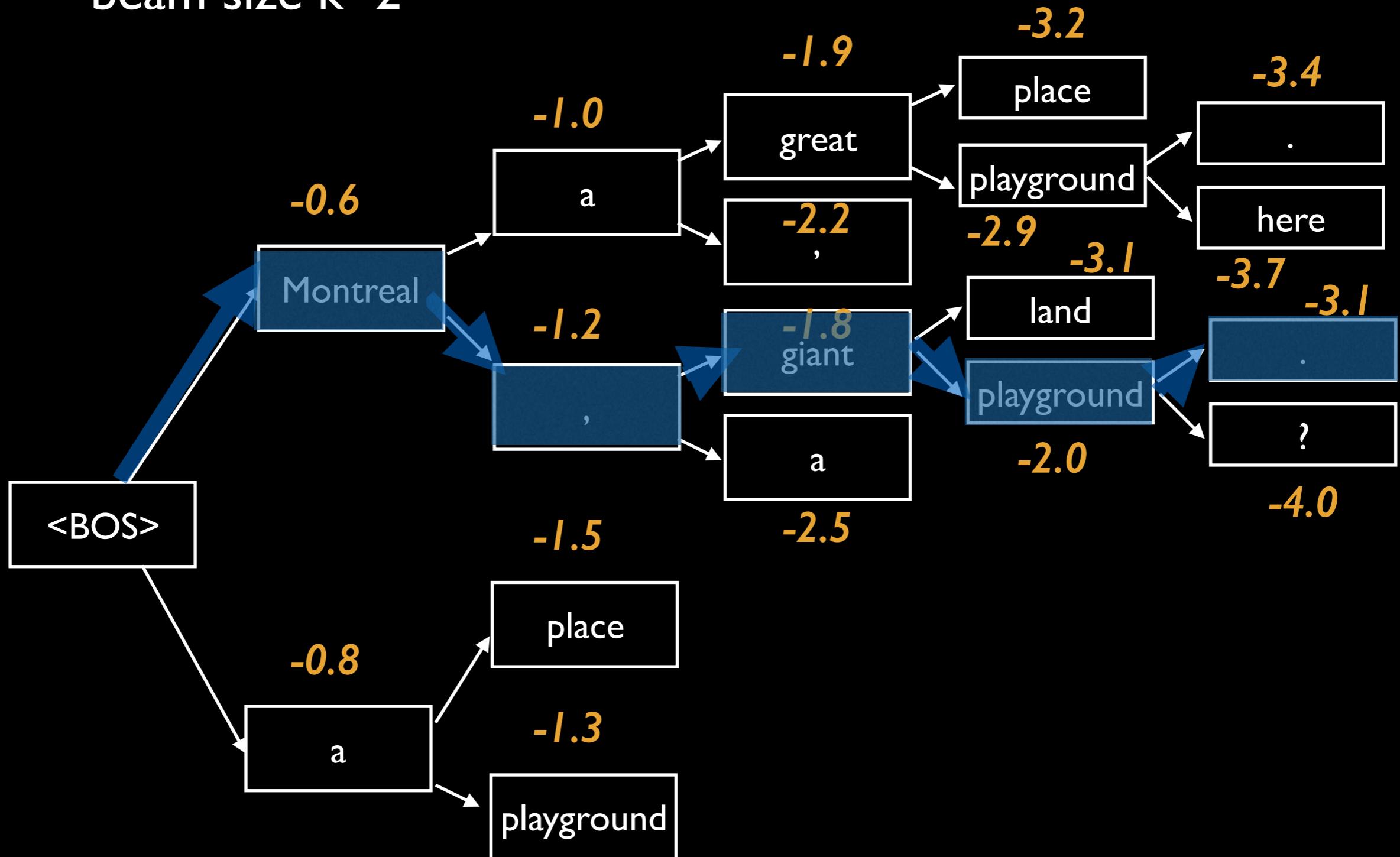
beam search example

beam size k=2



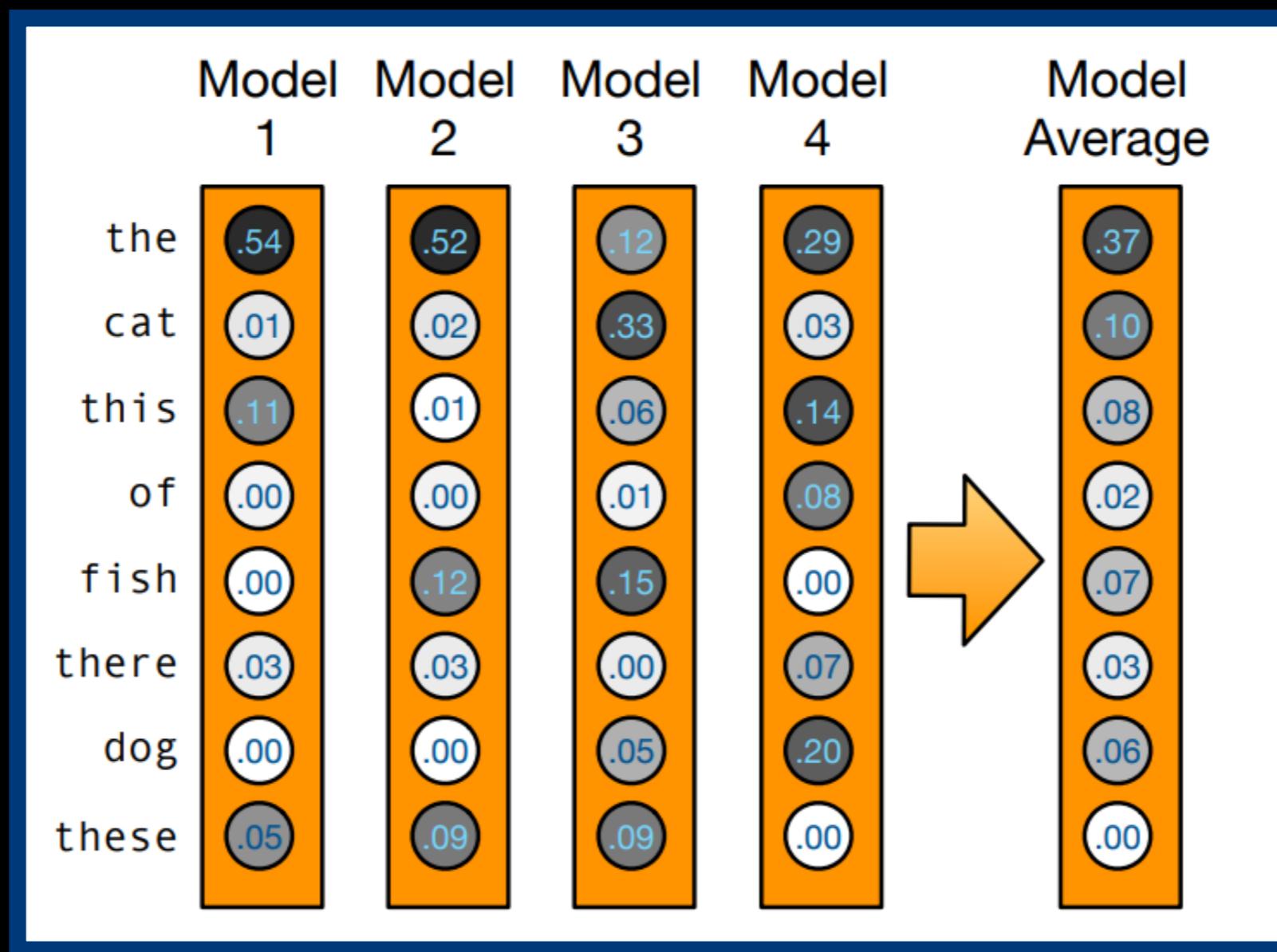
beam search example

beam size k=2



Question #8: more efficient or controlled search? binary NMT, constraint

ensemble

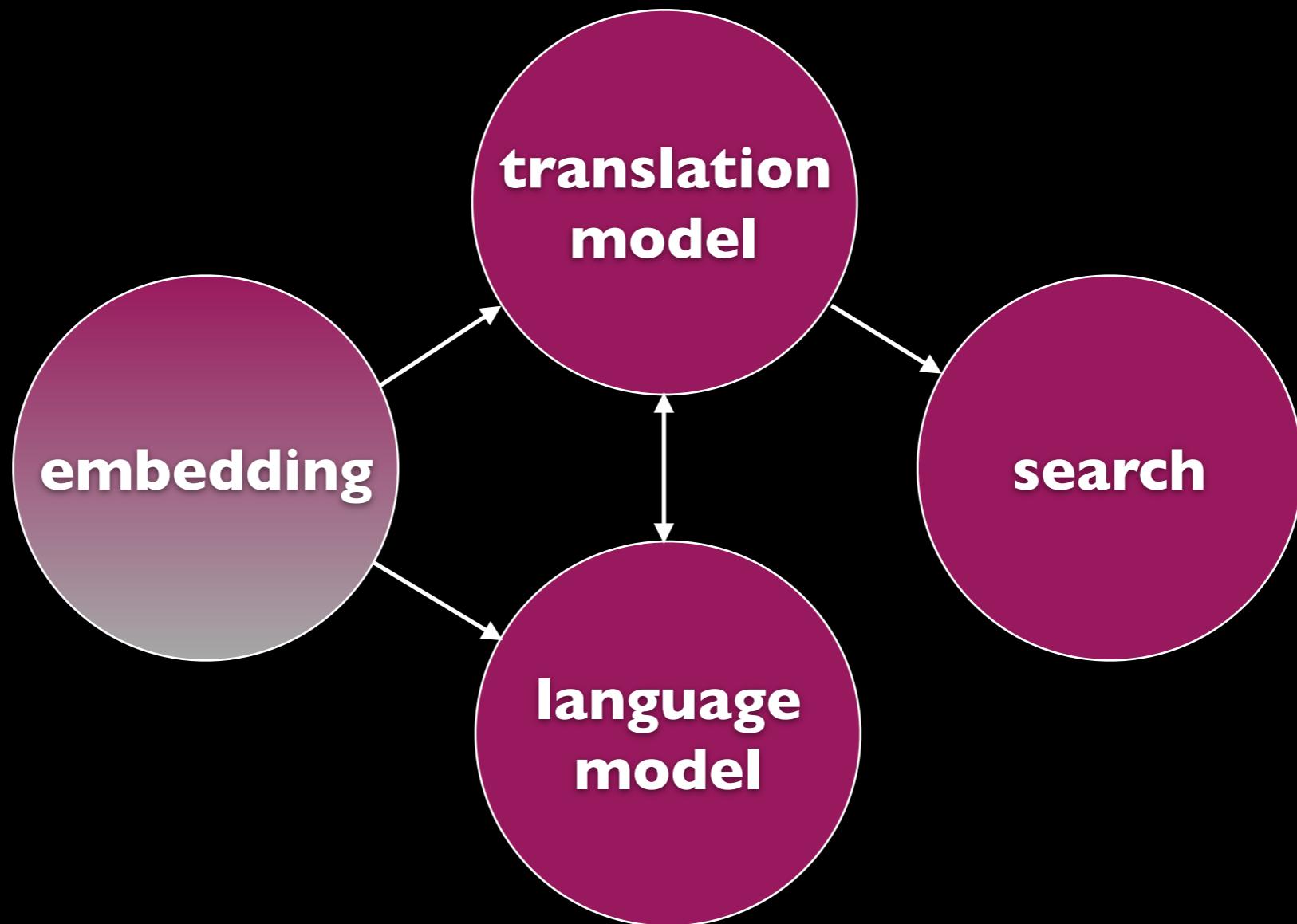


random initialization or
outputs from different iterations

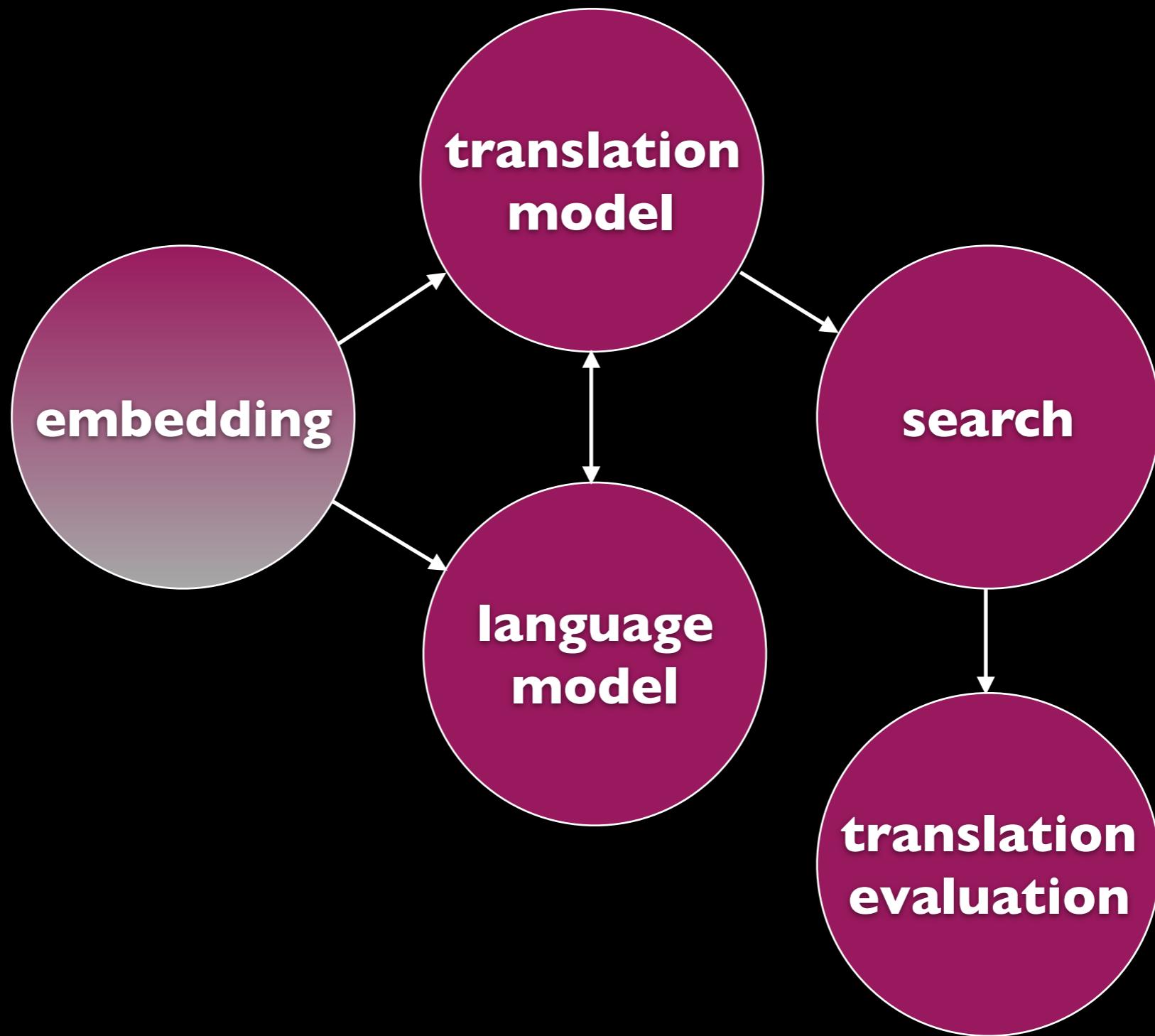
machine translation components



machine translation components



machine translation components



machine translation components

translation
evaluation

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure error rates:
 - edit distance: insertion, deletion, substitution
 - word error rate: normalized edit distance
 - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal ,
/ a giant playground .''

substitution#=1

- method to measure error rates:
 - edit distance: insertion, deletion, substitution
 - word error rate: normalized edit distance
 - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal ,[/]a giant playground .''

substitution#=|+|

- method to measure error rates:
 - edit distance: insertion, deletion, substitution
 - word error rate: normalized edit distance
 - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal ,a giant playground .''

substitution#=1+1 deletion#=1

- method to measure error rates:
 - edit distance: insertion, deletion, substitution
 - word error rate: normalized edit distance
 - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal ,a giant playground .''

substitution#=1+1 deletion#=1 insertion#=0

- method to measure error rates:
 - edit distance: insertion, deletion, substitution
 - word error rate: normalized edit distance
 - HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: ``Montreal is a city .''

reference: ``Montreal ,a giant playground .''

substitution#=|+| deletion#=| insertion#=0

edit distance#=|+|+|=3

- method to measure error rates:

- edit distance: insertion, deletion, substitution

- word error rate: normalized edit distance

- HTER: Human Targeted Translation Error Rate

machine translation evaluation

human evaluation is expensive, develop automatic evaluation criteria

hypothesis: "Montreal is a city."

reference: "Montreal, a giant playground."

substitution#=1+1 deletion#=1 insertion#=0

edit distance#=1+1+1=3

- method to measure error rates:

- edit distance: insertion, deletion, substitution
- word error rate: normalized edit distance $3/6=0.5$
- HTER: Human Targeted Translation Error Rate

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
- mainly relies on n-gram coverage

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
- mainly relies on n-gram coverage

1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far
 - precision of 1-gram through 4-gram
 - brevity penalty
 - mainly relies on n-gram coverage

1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

Question #9: higher correlation with human judgement? rich literature

machine translation evaluation: BLEU

BLEU (Bilingual Evaluation Understudy)

hypothesis: ``Montreal is a city .''

reference: ``Montreal , a giant playground .''

- method to measure accuracy
- most well-cited evaluation criterion so far

- precision of 1-gram through 4-gram

1-gram#=3
2-gram#=0
3-gram#=0
4-gram#=0

- brevity penalty

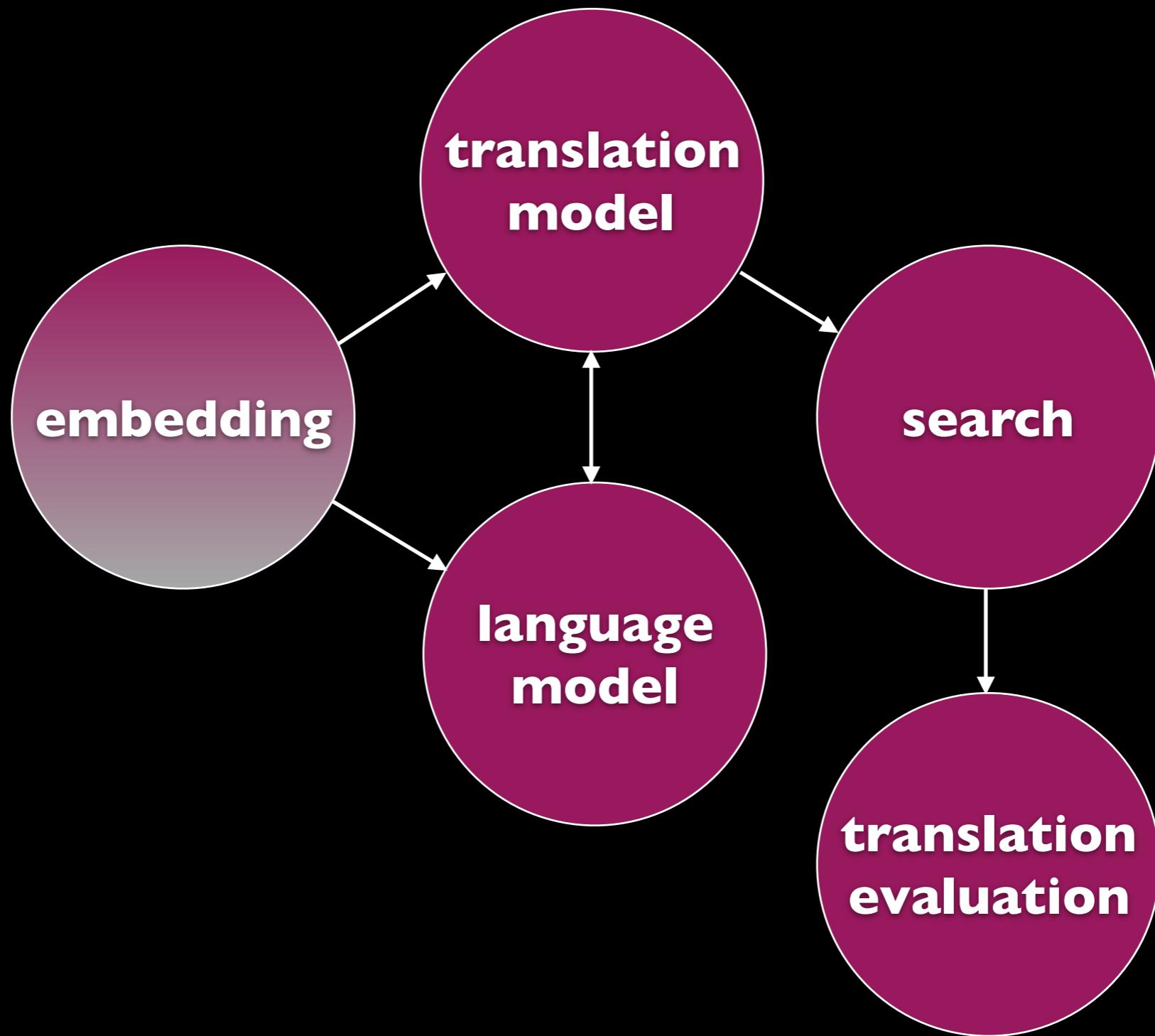
Question #10: better quality estimation?

Question #9: higher correlation with human judgement? rich literature

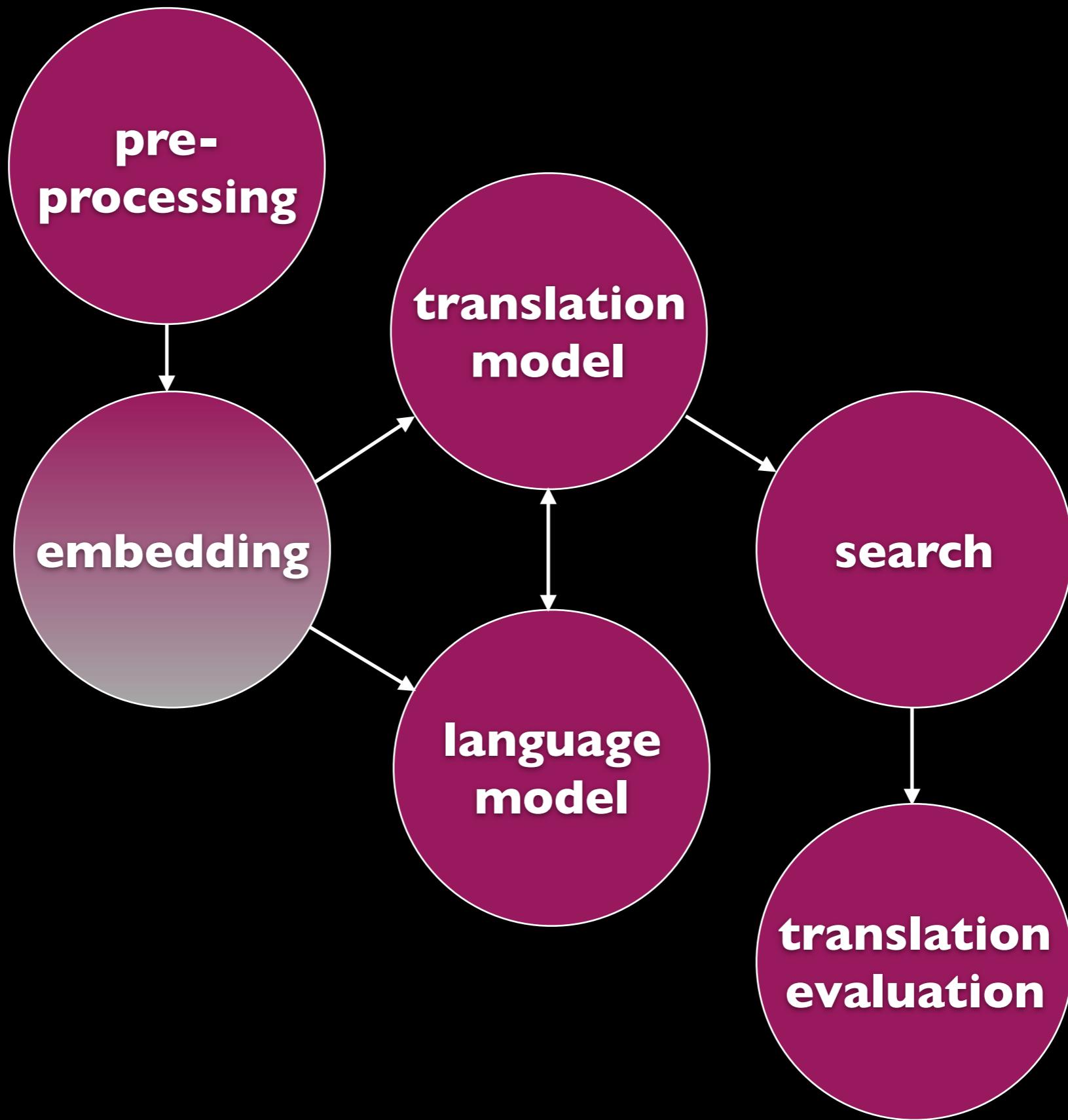
machine translation components

translation
evaluation

machine translation components



machine translation components



machine translation components



pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization
- word segmentation
- sentence segmentation
- domain classification

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization Question #11: text normalization
- word segmentation
- sentence segmentation
- domain classification

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization **Question #11: text normalization**
- word segmentation **Question #12: better subword?**
- sentence segmentation
- domain classification

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization **Question #11: text normalization**
- word segmentation **Question #12: better subword?**
- sentence segmentation **Question #13: monolingual and bilingual sentence segmentation**
- domain classification

pre-processing

- tokenization: separate words from punctuation marks, typically based on rules
- text normalization **Question #11: text normalization**
- word segmentation **Question #12: better subword?**
- sentence segmentation **Question #13: monolingual and bilingual sentence segmentation**
- domain classification **Question #14: domain adaptation**

text normalization

formal text: ``are you coming to the class tomorrow?''

informal text: ``r u cuming 2 class tomr?''

- bad translation: style, domain change, noise e.g. mis-spelling
- goal: translate different lexical variations
 - add noise to training: [Michell, et.al., 19]
 - word clustering [Khan et. al., 19]

بہت
bohot, bht, buhat [*very*]

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}
 - k: number of characters
- n-gram approach:

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

word segmentation

明天来上课吗?

- possible segmentation boundaries: 2^{k-1}

- k: number of characters
- n-gram approach:

← **statistical**

n-gram approach

$$\hat{f}_1^J(c_1^K) = \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}}^{k_j} | c_{k_{j-2-n}+1}^{k_{j-1-n}}, \dots, c_{k_{j-2}+1}^{k_{j-1}})$$

unigram

$$= \operatorname{argmax}_{k_1^J, J} \prod_{j=1}^J Pr(c_{k_{j-1}+1}^{k_j})$$

character based statistical MT

明天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\operatorname{argmax}_{f_1^J} \{ P(f_1^{J\lambda_1} P(e_1^J, b_1 I | f_1^J)^{\lambda_2} P(f_1^J, a_1^J | e_1^I)^{\lambda_3}$$

Transition from a boundary(+) to a no-boundary (-)

- C-to-E:** $P_{fe}(cba_k^- | dh_{sk}^+)$: $\frac{1}{I+1} P_G(f|e'_* \cdot)$ or $\frac{1}{I+1} P_G(f|e''_* \cdot)$, $I = 3$

$e_3 \cdot \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \blacksquare \cdot$ $e'_*(e_1) \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$	\gg	$e_3 \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \cdot$ $e'_*(e_1) \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$	or	$e_3 \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \blacksquare \cdot$ $e'_*(e_1) \cdot \cdot \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$
---	-------	---	-----------	---

Transition from boundary(+) to no-boundary (-)

- E-to-C:** $P_{ef}(cba_k | dh_{sk}^-)$: $\frac{1}{J} |e| P_G(e'|f) P_G(e''|f)$, here $|e| = 2, J = 3$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \blacksquare \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	\gg	$e_4 \cdot \cdot \blacksquare$ $e'' \cdot \blacksquare \cdot \cdot$ $e' \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f \quad f_2$
---	-------	--

Transition from no-boundary (-) to boundary (+)

- E-to-C:** $P_{ef}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{J+2} |e| P(e_1|f') P(e_2|f'')$ or ..., $J = 2$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \blacksquare \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f_2$	\gg	$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \blacksquare \cdot \cdot$ $e' \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	or	$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \cdot \blacksquare \cdot \cdot$ $e' \cdot \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$
---	-------	---	-----------	---

$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \cdot \blacksquare \cdot \cdot$ $e' \cdot \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	\gg	$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \cdot \cdot \blacksquare$ $e' \cdot \cdot \cdot \blacksquare$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$
---	-------	---

Transition from no-boundary(-) to boundary(+)

- C-to-E:** $P_{fe}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{I+1}^2 P_G(f'|e_*) P_G(f''|e_*)$, $I = 2$

$e_2 \cdot \cdot \blacksquare$ $e_* \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$	\gg	$e_2 \cdot \cdot \cdot \blacksquare$ $e_* \cdot \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$
---	-------	--

character based statistical MT

明天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\operatorname{argmax}_{f_1^J} \{ P(f_1^{J\lambda_1} P(e_1^J, b_1 I | f_1^J)^{\lambda_2} P(f_1^J, a_1^J | e_1^I)^{\lambda_3}$$

statistical

Transition from a boundary(+) to a no-boundary (-)

- C-to-E:** $P_{fe}(cba_k^- | dh_{sk}^+)$: $\frac{1}{I+1} P_G(f|e''_*)$ or $\frac{1}{I+1} P_G(f|e'_*)$, $I = 3$

$e_3 \cdot \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \blacksquare \cdot$ $e'_*(e_1) \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$	$e_3 \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \cdot$ $e'_*(e_1) \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$	$e_3 \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \cdot$ $e'_*(e_1) \cdot \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$
---	---	--

>> or

Transition from no-boundary (-) to boundary (+)

- E-to-C:** $P_{ef}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{J+2}^{|e|} P(e_1|f') P(e_2|f'')$ or ..., $J = 2$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f \quad f_2$	$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$
---	--	--

>> or

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$
--	--

Transition from boundary(+) to no-boundary (-)

- E-to-C:** $P_{ef}(cba_k | dh_{sk}^-)$: $\frac{1}{J}^{|e|} P_G(e'|f) P_G(e''|f)$, here $|e| = 2, J = 3$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f \quad f_2$
--	---

>>

Transition from no-boundary(-) to boundary(+)

- C-to-E:** $P_{fe}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{I+1}^2 P_G(f'|e_*) P_G(f''|e_*)$, $I = 2$

$e_2 \cdot \cdot \cdot \blacksquare$ $e_*(e_1) \cdot \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f \quad f_3$	$e_2 \cdot \cdot \cdot \blacksquare$ $e_*(e_1) \cdot \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$
---	--

>>

character based statistical MT

明天|来|上 课|吗?

- Gibbs sampling: joint model word alignment and word segmentation

$$\operatorname{argmax}_{f_1^J} \{ P(f_1^{J\lambda_1} P(e_1^J, b_1 I | f_1^J)^{\lambda_2} P(f_1^J, a_1^J | e_1^I)^{\lambda_3}$$

statistical

Transition from a boundary(+) to a no-boundary (-)

- C-to-E:** $P_{fe}(cba_k^- | dh_{sk}^+)$: $\frac{1}{I+1} P_G(f|e'_*)$ or $\frac{1}{I+1} P_G(f|e''_*)$, $I = 3$

$e_3 \cdot \cdot \cdot \blacksquare$ $e''_*(e_2) \cdot \cdot \blacksquare \cdot$ $e'_*(e_1) \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$	$e_3 \cdot \cdot \blacksquare$ $e''_* \cdot \cdot \cdot$ $e'_* \cdot \blacksquare \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f f_3$	$e_3 \cdot \cdot \blacksquare$ $e''_* \cdot \cdot \cdot$ $e'_* \cdot \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f f_3$
---	---	--

>> or

Transition from no-boundary (-) to boundary (+)

- E-to-C:** $P_{ef}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{J+2}^{|e|} P(e_1|f')P(e_2|f'')$ or ..., $J = 2$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f f_2$	$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \cdot \blacksquare \cdot$ $e' \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	$e_4 \cdot \cdot \cdot \blacksquare$ $e'' \cdot \cdot \cdot$ $e' \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$
---	---	--

>> or

Transition from boundary(+) to no-boundary (-)

- E-to-C:** $P_{ef}(cba_k | dh_{sk}^-)$: $\frac{1}{J}^{|e|} P_G(e'|f)P_G(e''|f)$, here $|e| = 2, J = 3$

$e_4 \cdot \cdot \cdot \blacksquare$ $e''(e_3) \cdot \cdot \cdot$ $e'(e_2) \cdot \blacksquare \cdot \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f' f'' f_3$	$e_4 \cdot \cdot \blacksquare$ $e'' \cdot \blacksquare \cdot$ $e' \cdot \blacksquare \cdot$ $e_1 \blacksquare \cdot \cdot \cdot$ $f_0 f f_2$
--	--

Transition from no-boundary(-) to boundary(+)

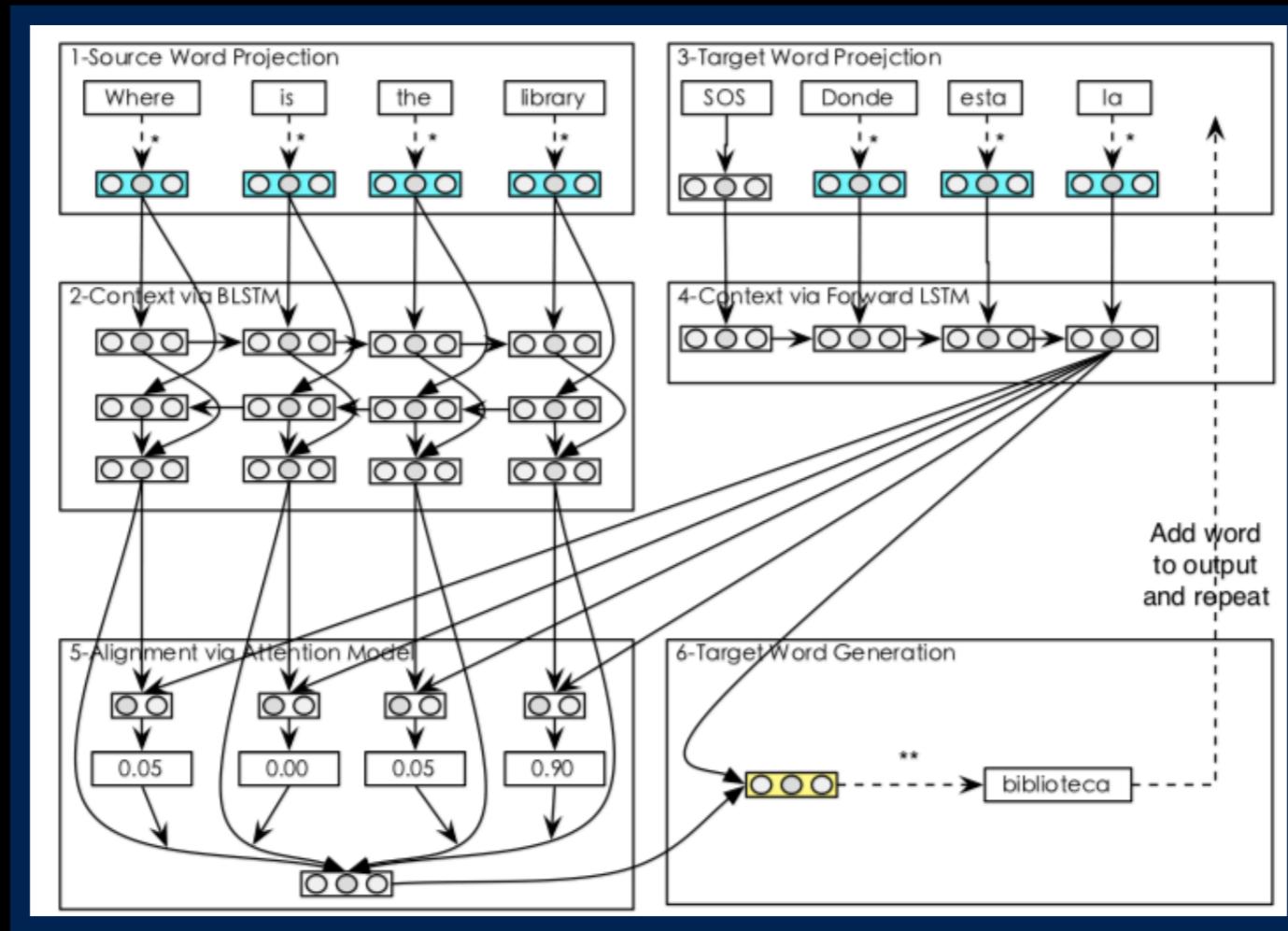
- C-to-E:** $P_{fe}(cba_k^+ | dh_{sk}^-)$: $\frac{1}{I+1}^2 P_G(f'|e_*)P_G(f''|e_*)$, $I = 2$

$e_2 \cdot \cdot \blacksquare$ $e_* \cdot \blacksquare \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f f_3$	$e_2 \cdot \cdot \cdot \blacksquare$ $e_* \cdot \blacksquare \cdot \cdot$ $e_0 \blacksquare \cdot \cdot \cdot$ $f_1 f' f'' f_4$
---	--

Question #15: what can we borrow from statistical MT?

character based neural MT

- integrate with neural network framework
- [Ling, et.al., 15], [Cherry, et.al. 18], [Lee, et.al., 18]...



Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
 - aaabdaaaabac
 - ZabdZbac ($Z=aa$)
 - ZYdZYac ($Y=ab$; $Z=aa$)
 - XdXac ($X=ZY$; $Y=ab$; $Z=aa$)

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
 - aaabdaaaabac
 - ZabdZbac ($Z=aa$)
 - ZYdZYac ($Y=ab$; $Z=aa$)
 - XdXac ($X=ZY$; $Y=ab$; $Z=aa$)

Question #16: better subword
e.g. with morphological knowledge?

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
 - aaabdaaaabac
 - ZabdZbac ($Z=aa$)
 - ZYdZYac ($Y=ab$; $Z=aa$)
 - XdXac ($X=ZY$; $Y=ab$; $Z=aa$)

Question #16: better subword
e.g. with morphological knowledge?

Question #17: unseen words?

Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
 - aaabdaaaabac
 - ZabdZbac ($Z=aa$)
 - ZYdZYac ($Y=ab$; $Z=aa$)
 - XdXac ($X=ZY$; $Y=ab$; $Z=aa$)

Question #16: better subword
e.g. with morphological knowledge?

Question #17: unseen words?

Question #18: named entities?

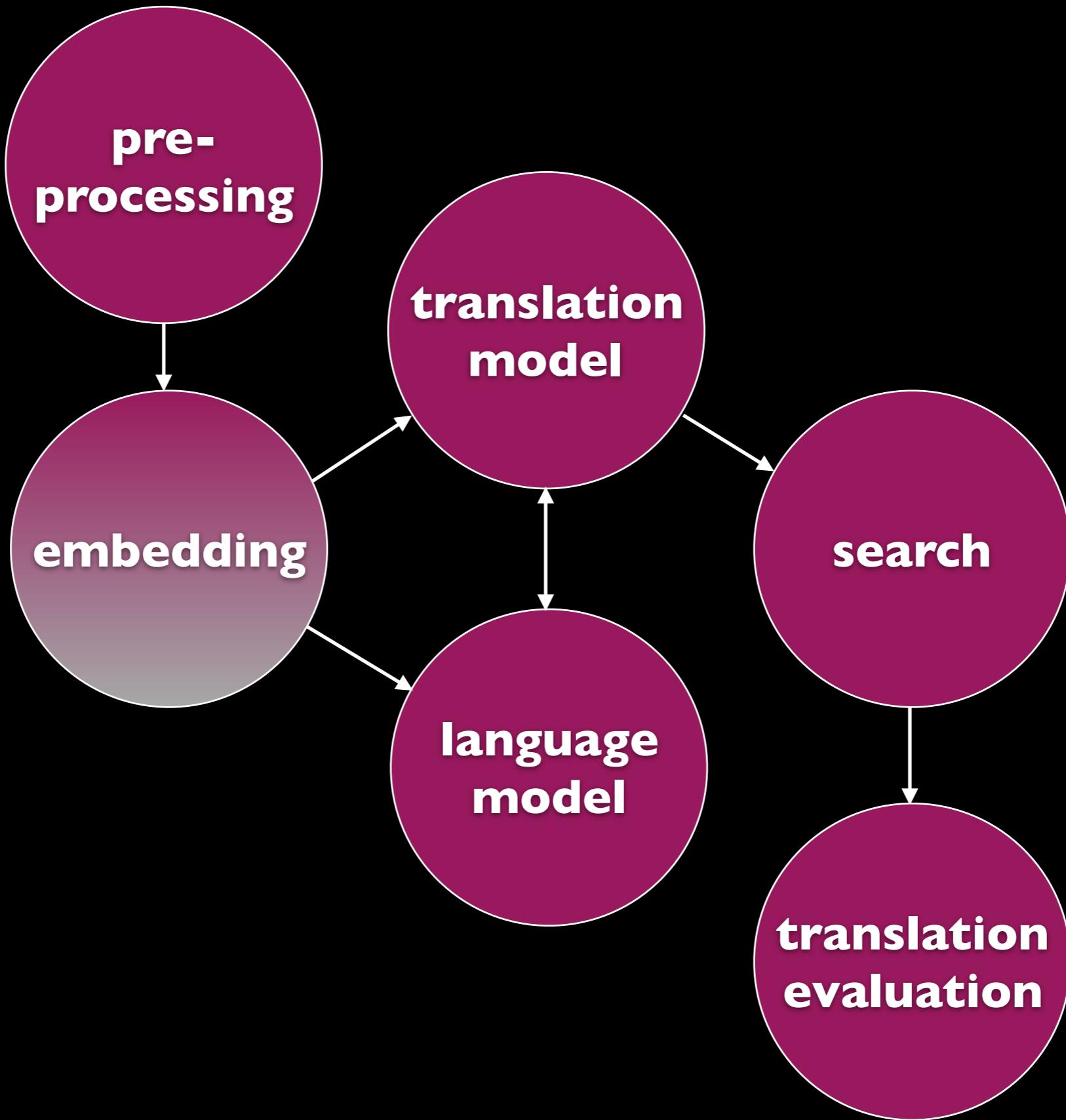
Byte-Pair-Encoding (BPE)

- efficient and affective, NMT independent
- [Gage, 1994], [Sennrich, et.al., 16]
 - aaabdaaaabac
 - ZabdZbac ($Z=aa$)
 - ZYdZYac ($Y=ab$; $Z=aa$)
 - XdXac ($X=ZY$; $Y=ab$; $Z=aa$)

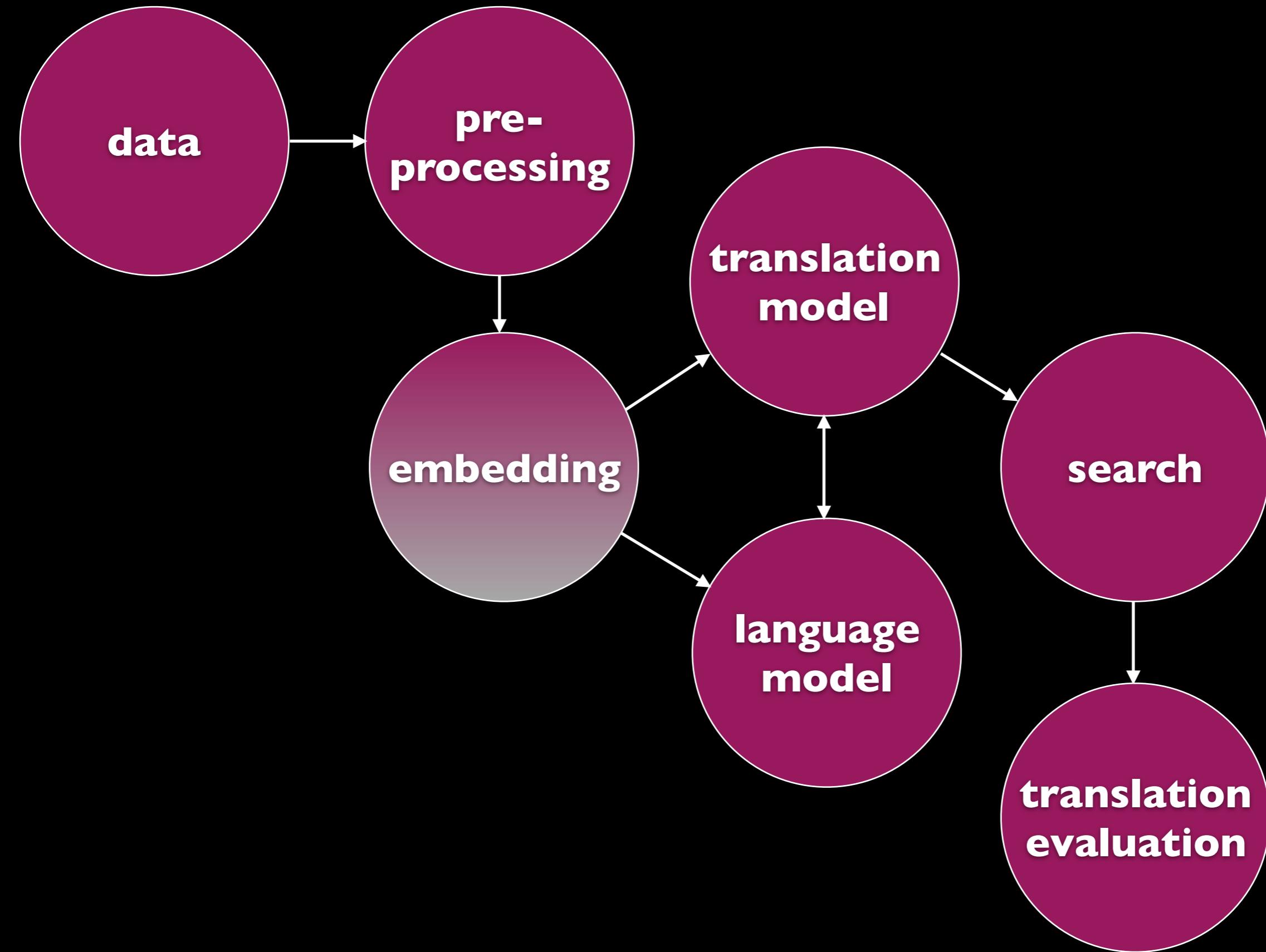
machine translation components



machine translation components



machine translation components



machine translation components

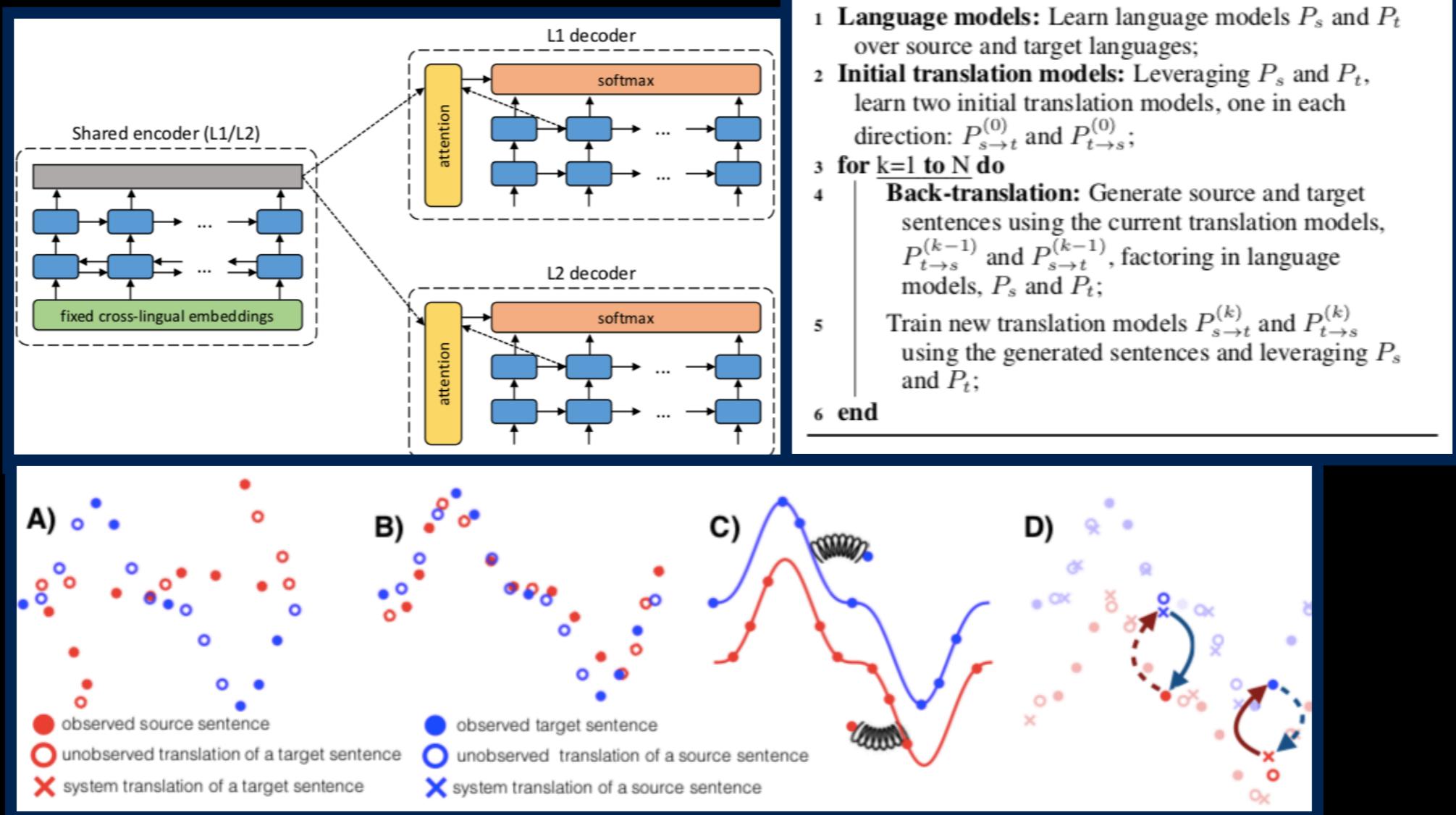


data

unsupervised NMT

- NO parallel training data possible?

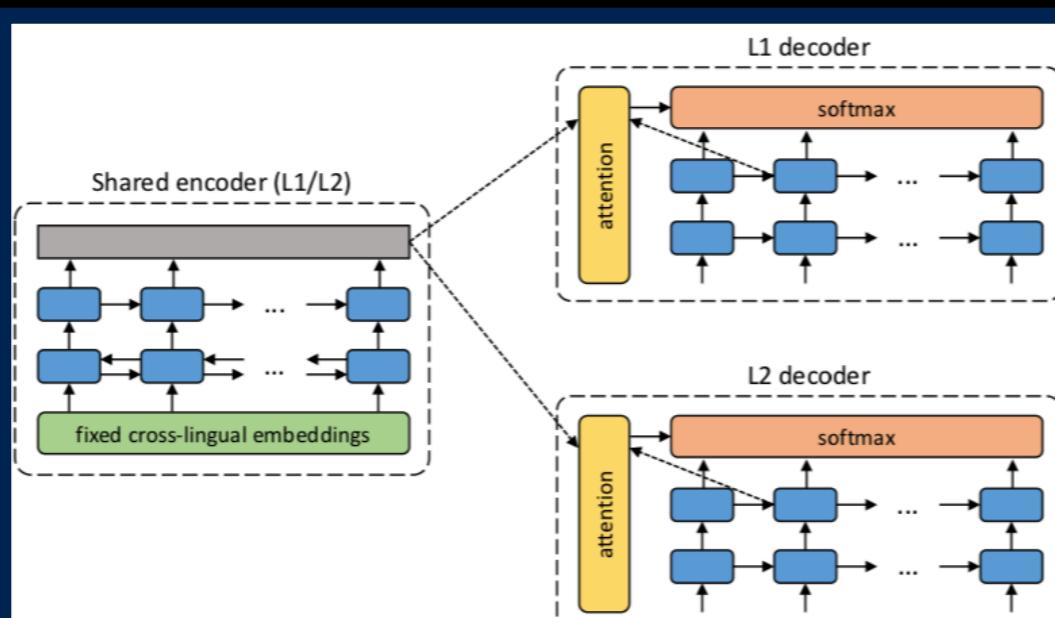
- [Artetxe, et.al., 17],[Lample, et.al., 16]



unsupervised NMT

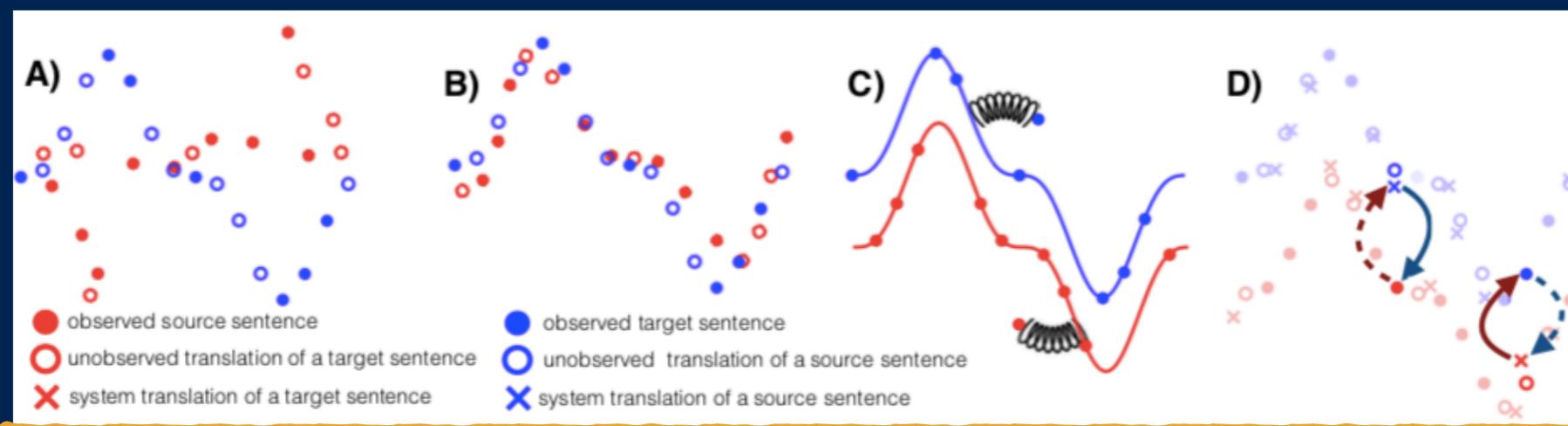
- NO parallel training data possible?

- [Artetxe, et.al., 17],[Lample, et.al., 16]



Algorithm 1: Unsupervised MT

```
1 Language models: Learn language models  $P_s$  and  $P_t$  over source and target languages;  
2 Initial translation models: Leveraging  $P_s$  and  $P_t$ , learn two initial translation models, one in each direction:  $P_{s \rightarrow t}^{(0)}$  and  $P_{t \rightarrow s}^{(0)}$ ;  
3 for  $k=1$  to  $N$  do  
4   Back-translation: Generate source and target sentences using the current translation models,  $P_{t \rightarrow s}^{(k-1)}$  and  $P_{s \rightarrow t}^{(k-1)}$ , factoring in language models,  $P_s$  and  $P_t$ ;  
5   Train new translation models  $P_{s \rightarrow t}^{(k)}$  and  $P_{t \rightarrow s}^{(k)}$  using the generated sentences and leveraging  $P_s$  and  $P_t$ ;  
6 end
```

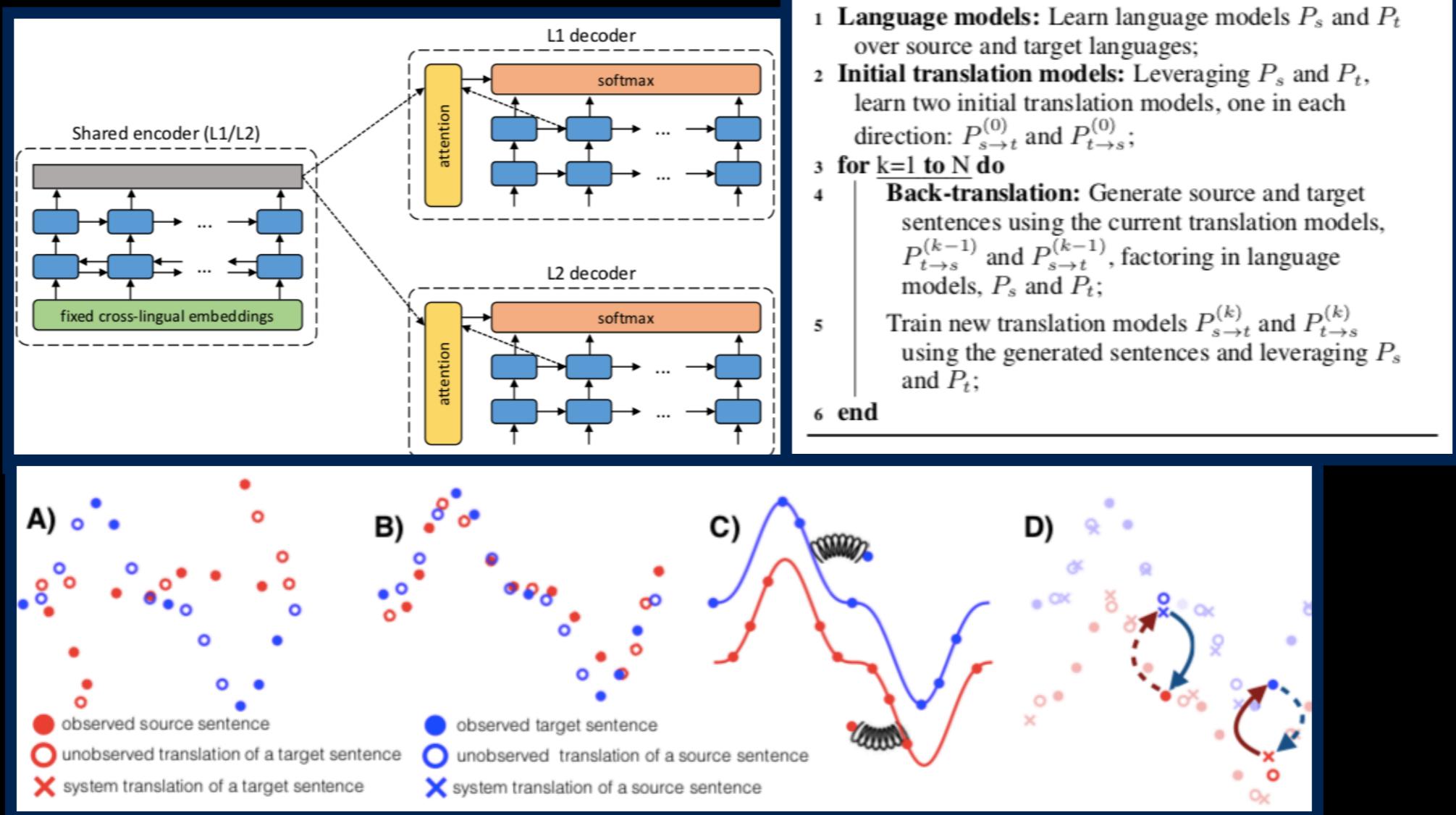


Question #19: higher quality in unsupervised MT?

unsupervised NMT

- NO parallel training data possible?

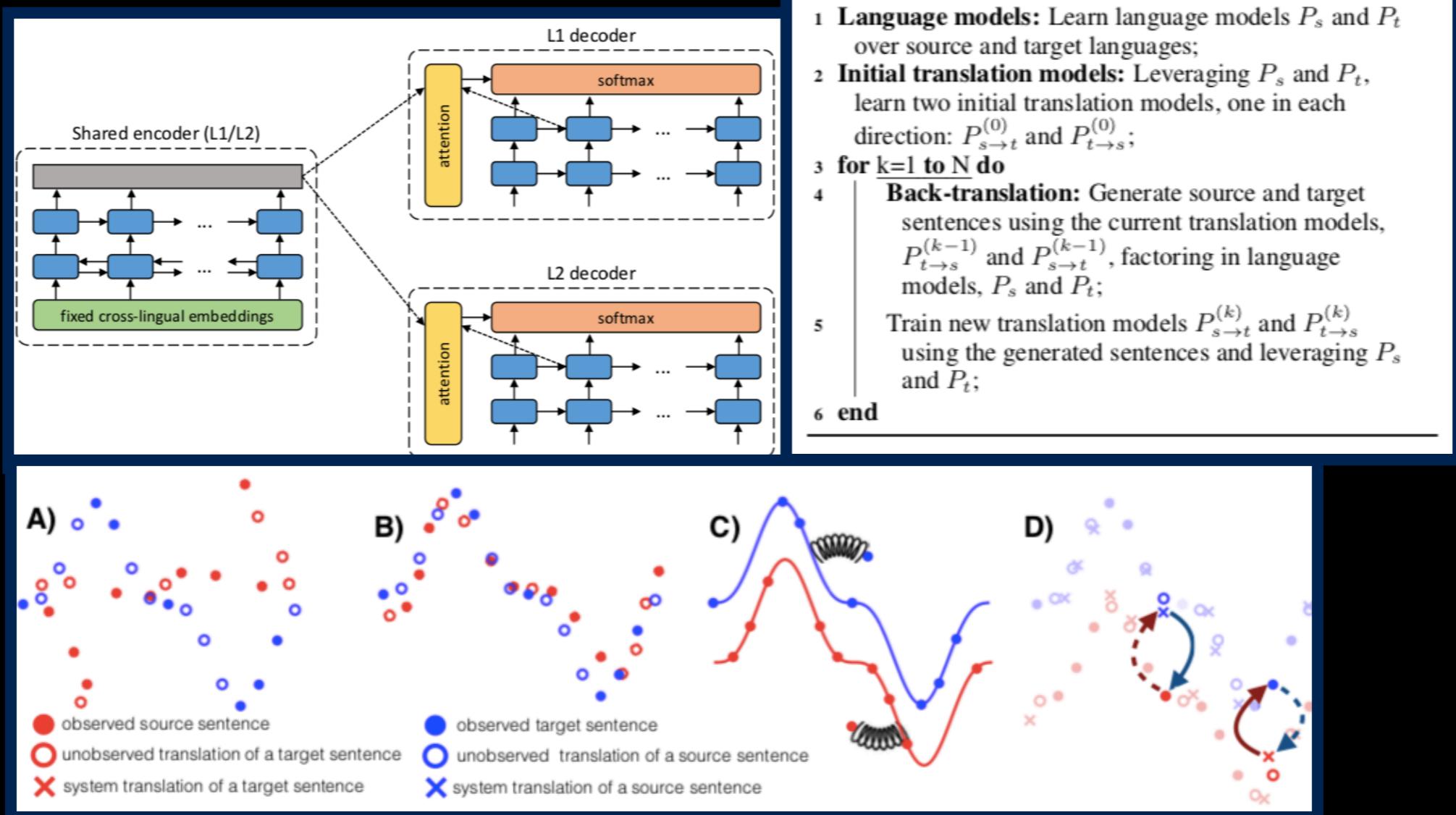
- [Artetxe, et.al., 17],[Lample, et.al., 16]



unsupervised NMT

- NO parallel training data possible?

- [Artetxe, et.al., 17],[Lample, et.al., 16]



back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality due to data augmentation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality due to data augmentation

Question #20: back translation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality due to data augmentation

back translation

- German to French translation is good
- French to German translation is bad
- use German to French MT system to translate German monolingual data e.g. [Sennrich, et.al.,16]
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality due to data augmentation

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality

Question #21: pivot translation

pivot translation

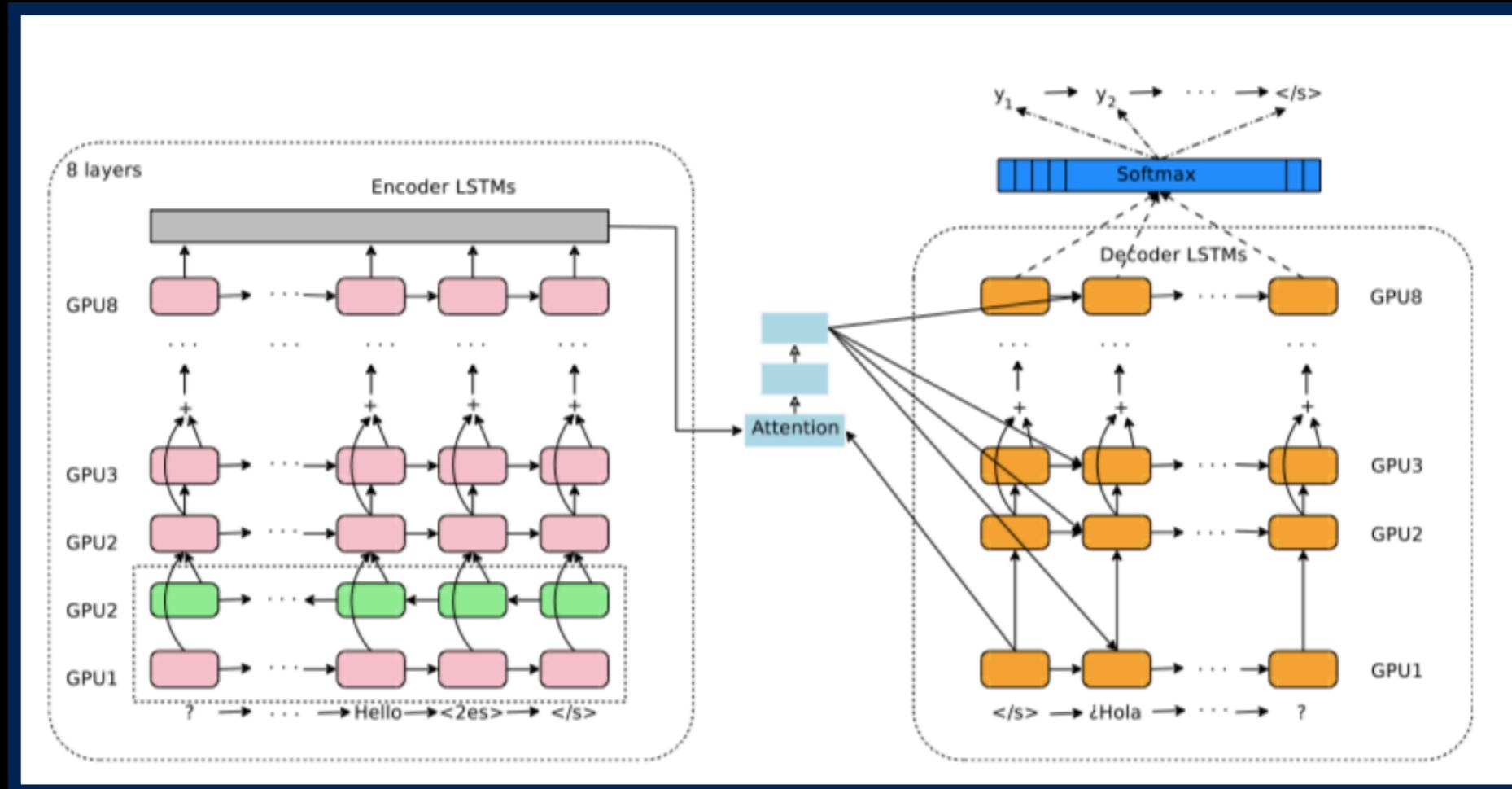
- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality

pivot translation

- lack of German - French parallel training data
- rich data of German - English, and French - English
- generate German - French parallel data using German - English and French - English MT systems
- add the synthetic data into French to German MT training
 - usually greatly improves translation quality

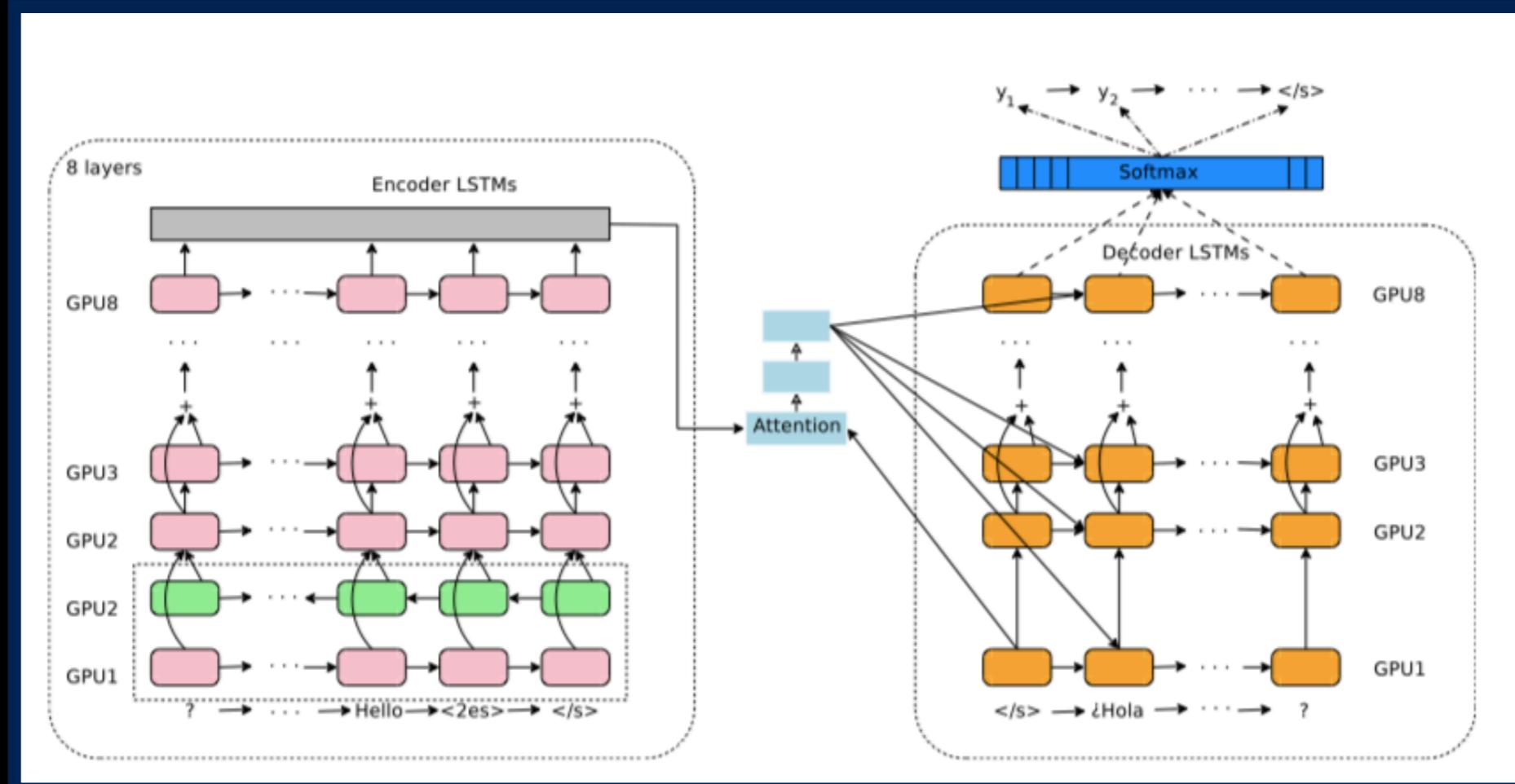
multiple languages

- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]



multiple languages

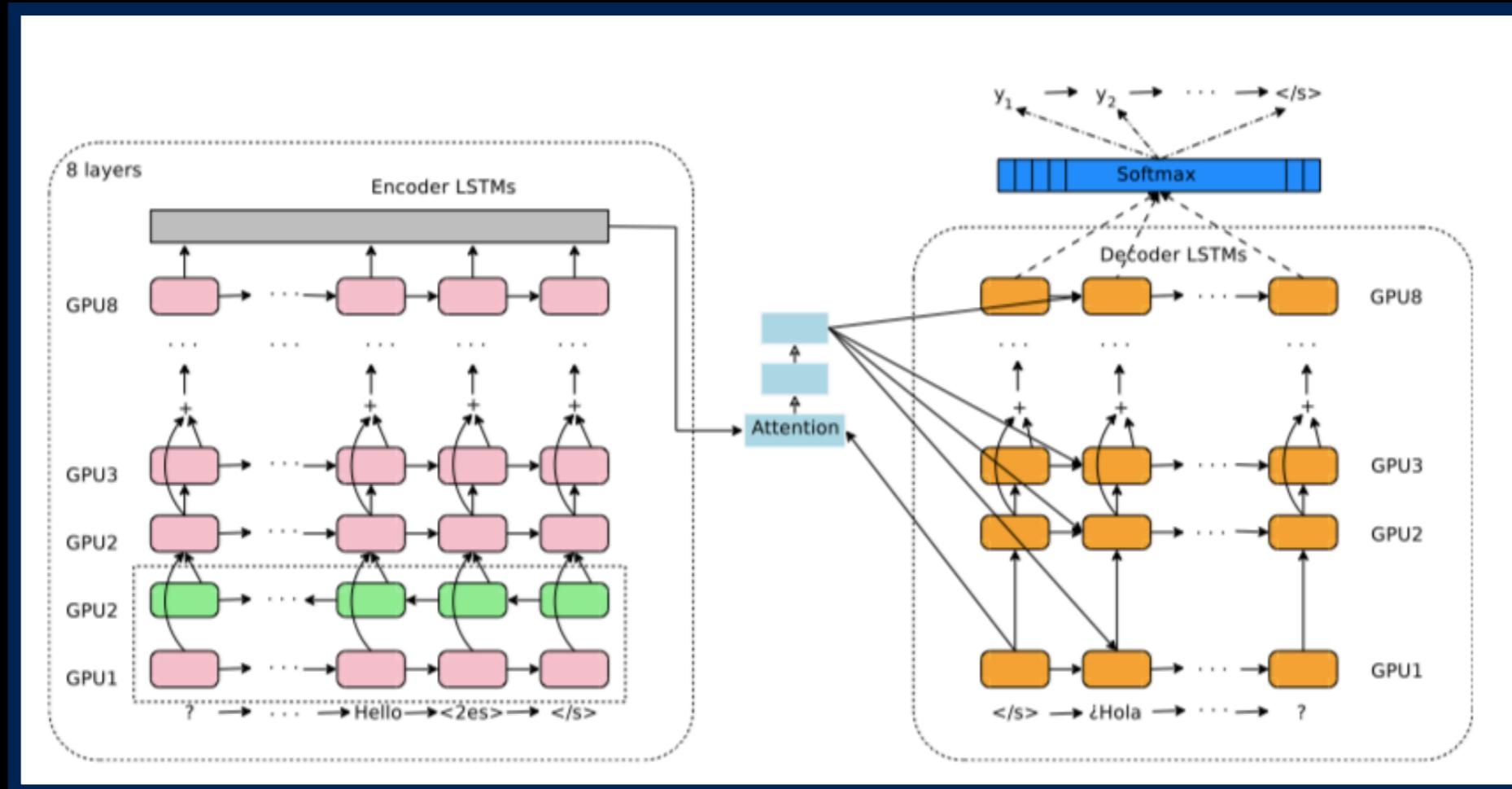
- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]



Question #22: multi-lingual and zero resource

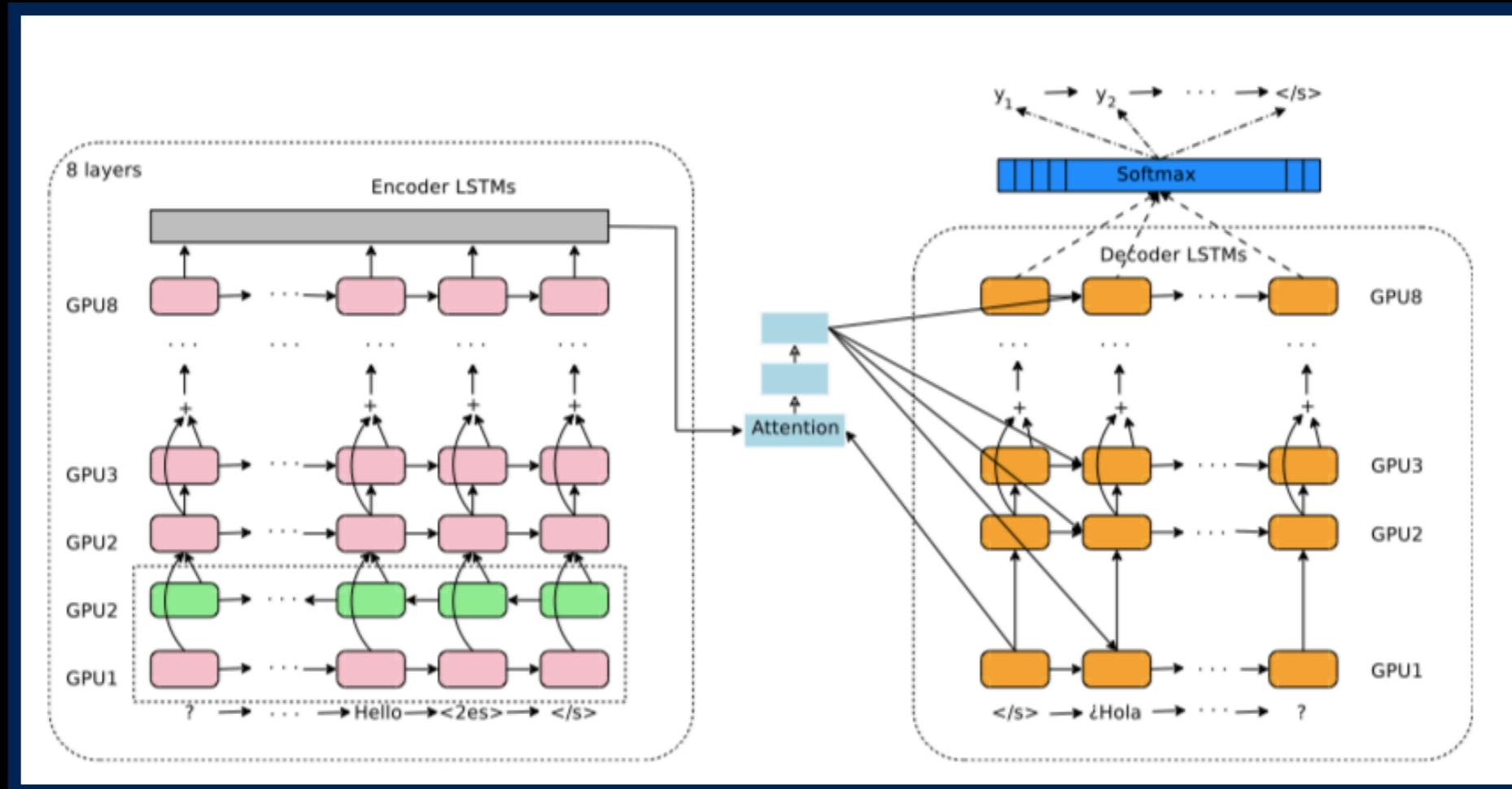
multiple languages

- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]

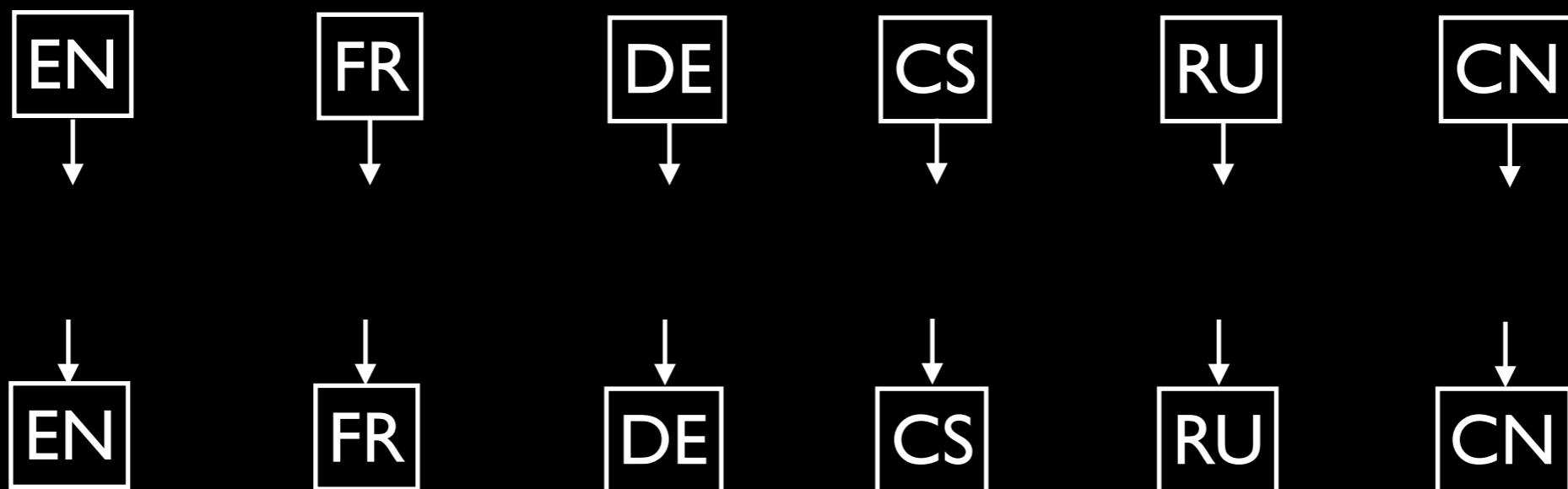
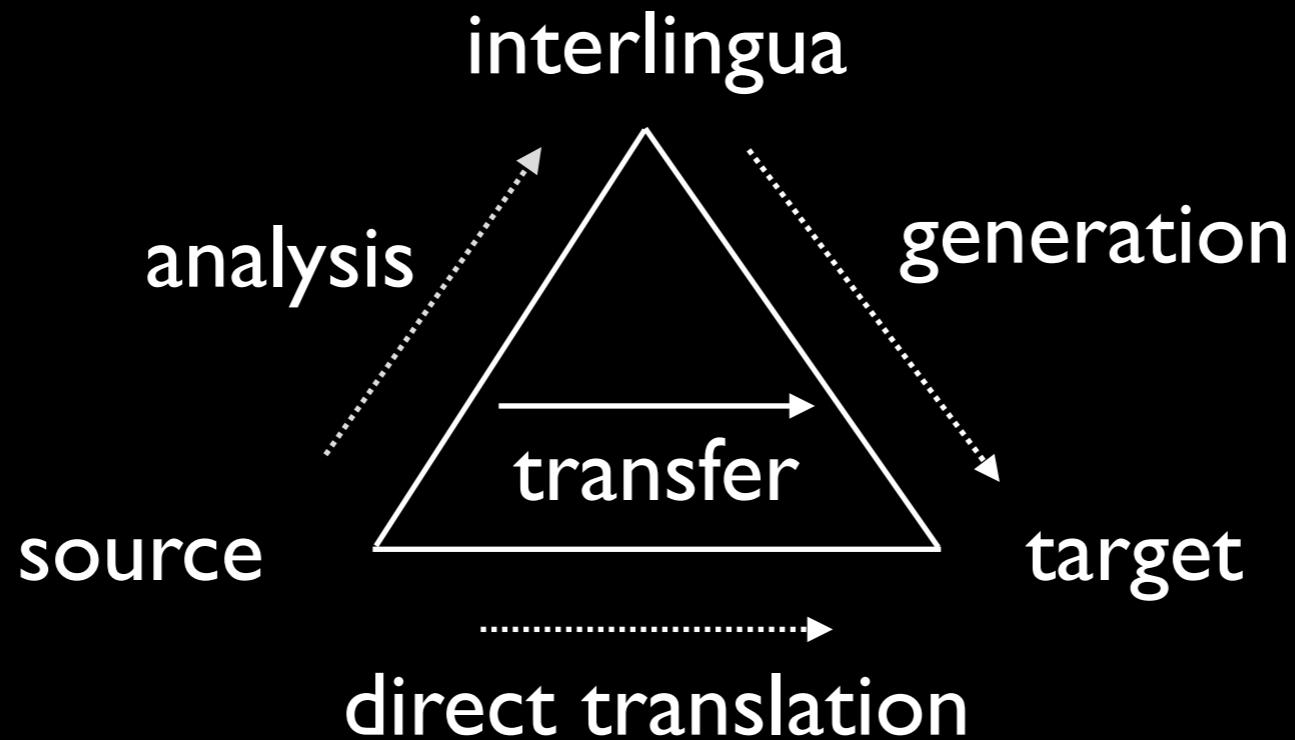


multiple languages

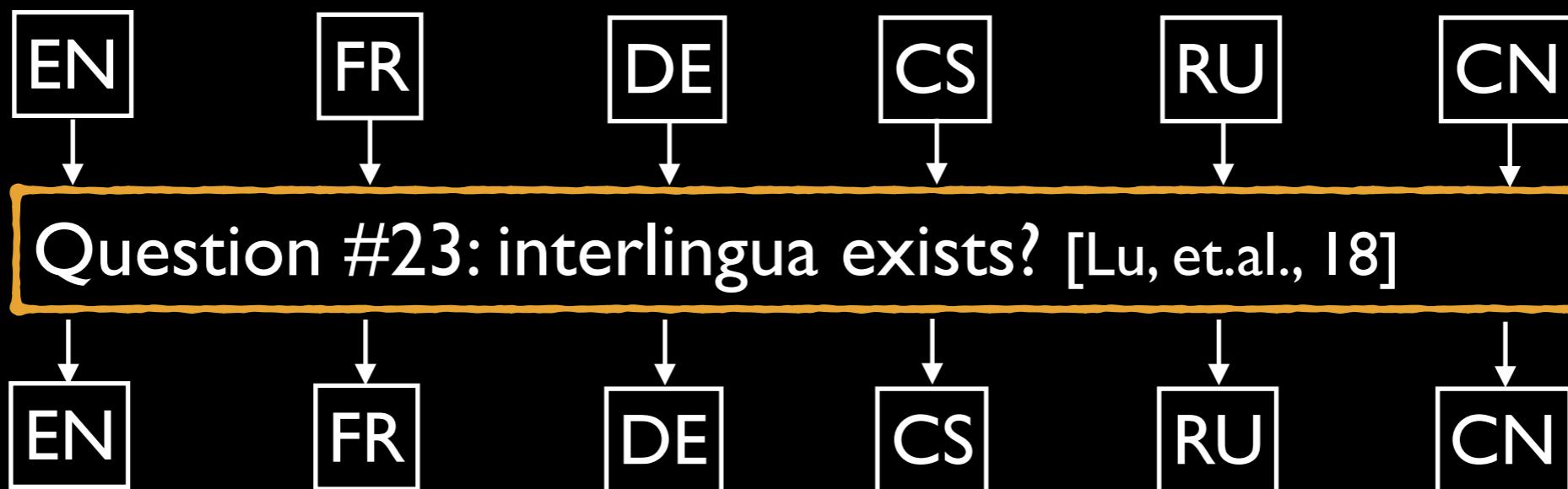
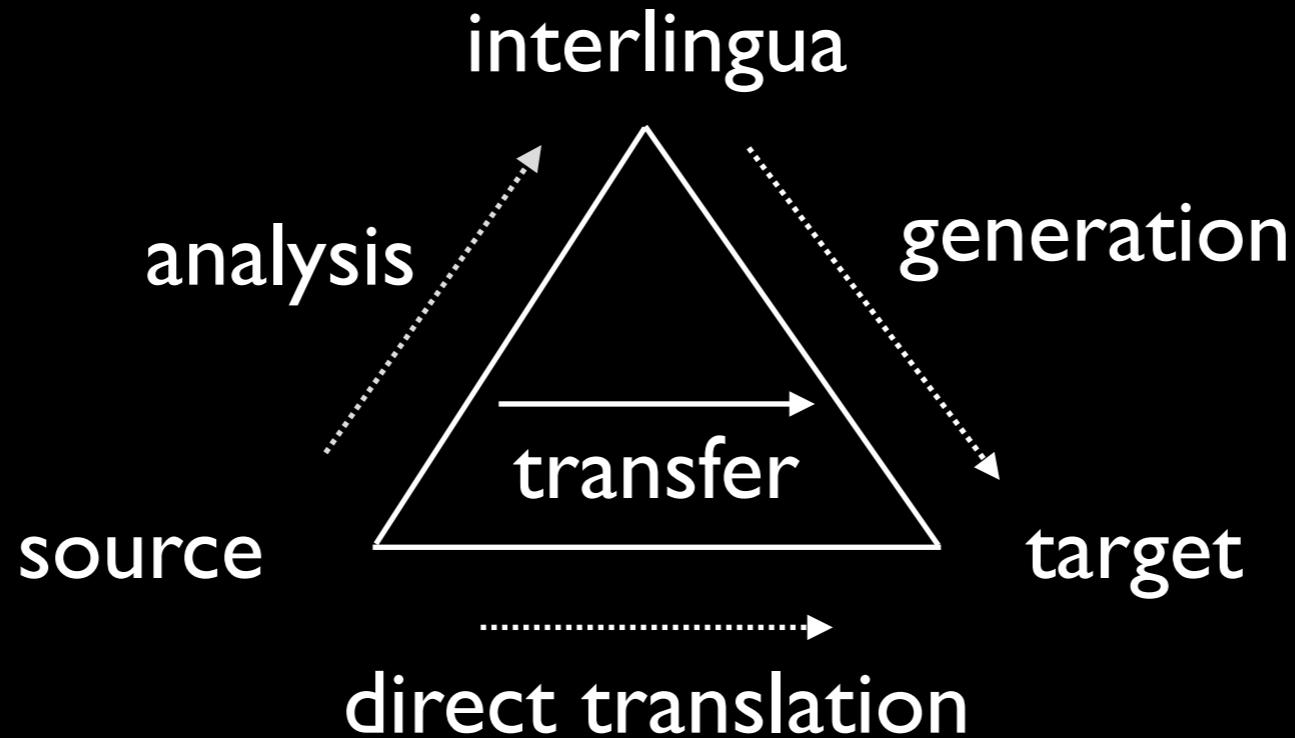
- observation in WMT'19: adding Hindi - English parallel data improves Gujarati - English translation
- how about many other languages? [Firat, et.al., 16], [Johnson, et.al., 17]



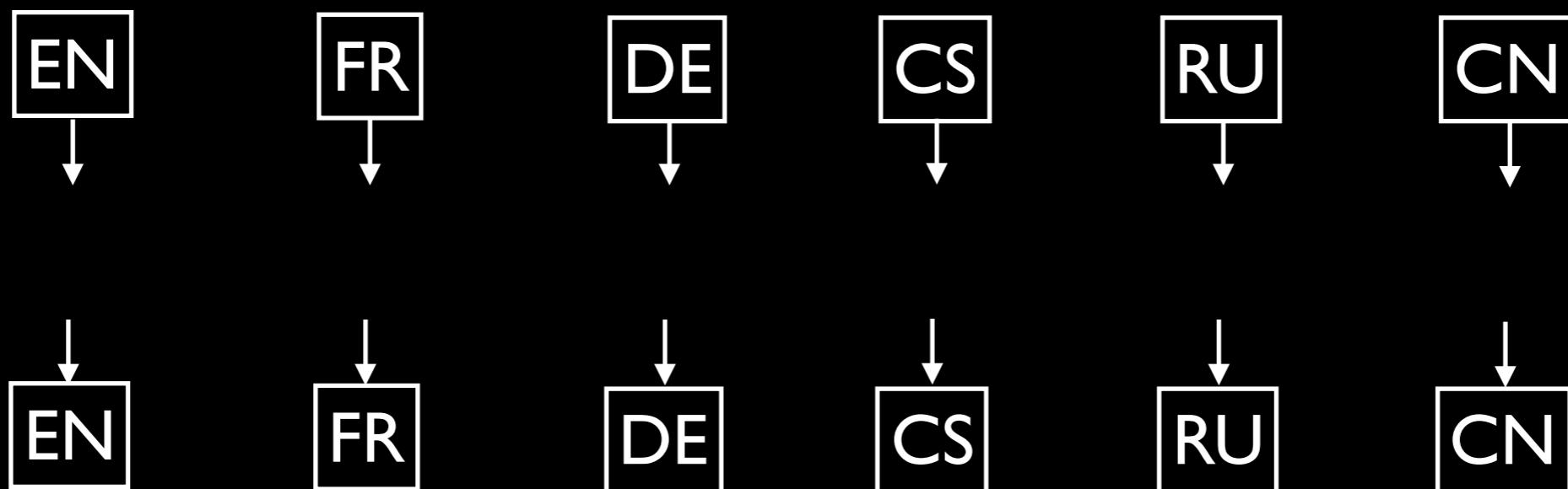
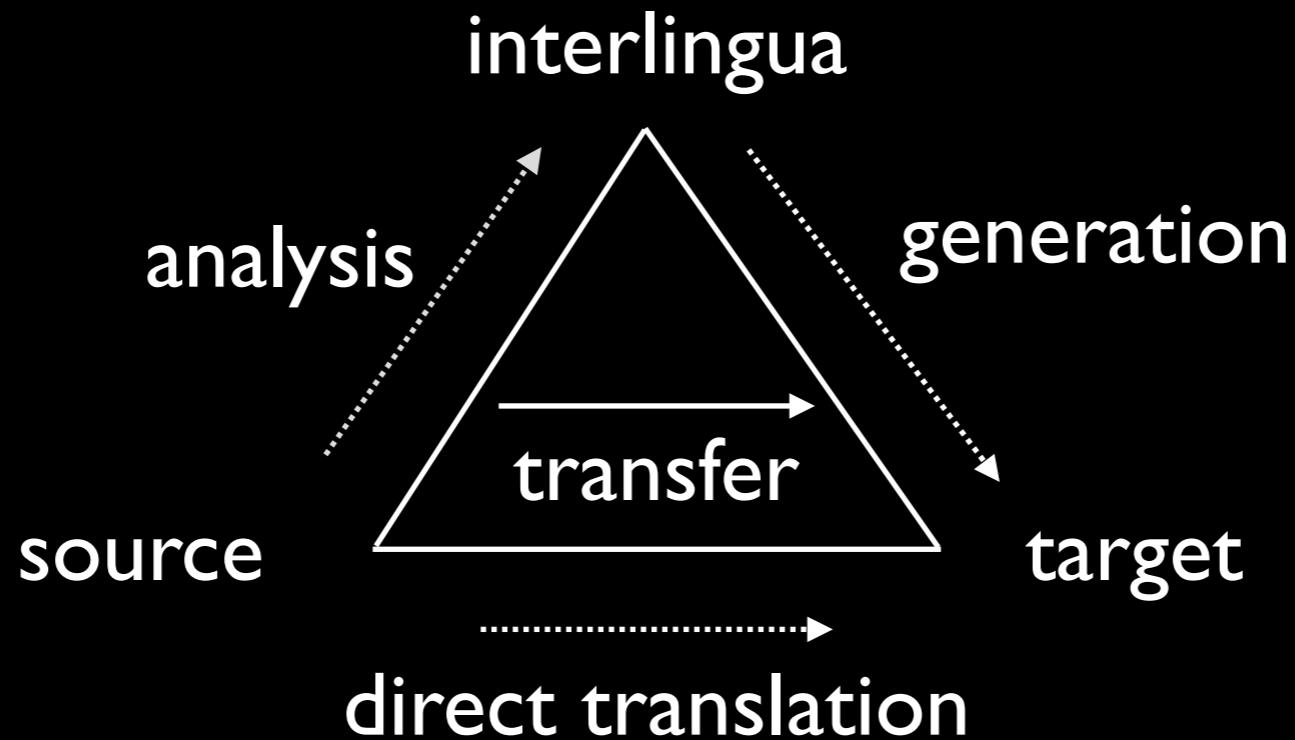
the concept of interlingua



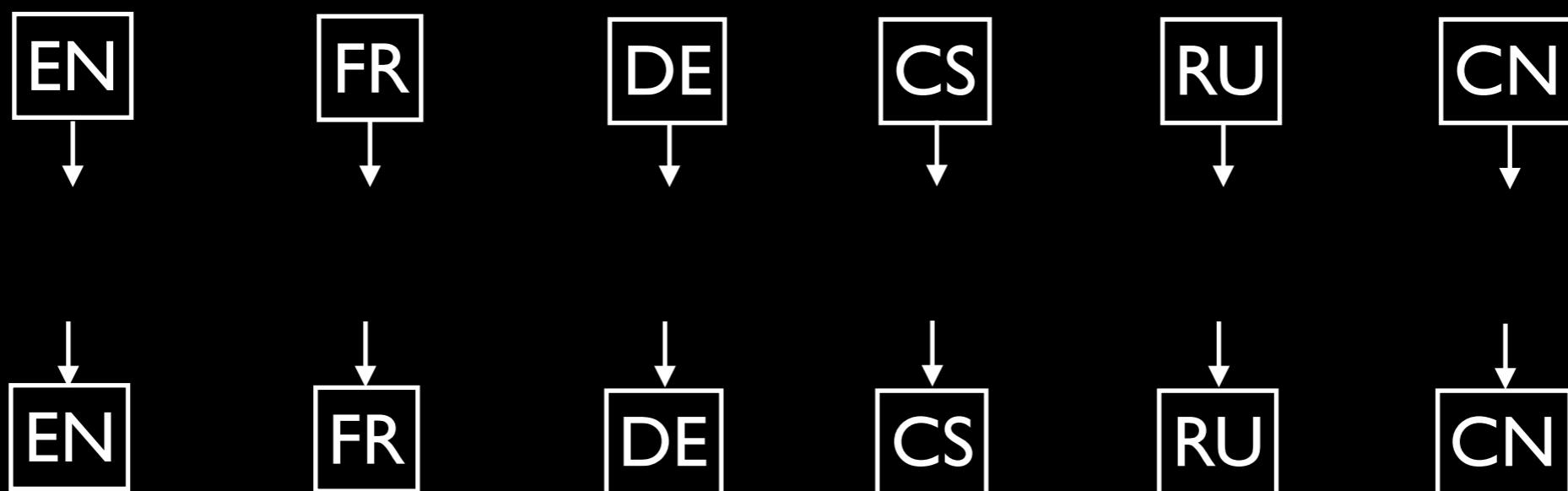
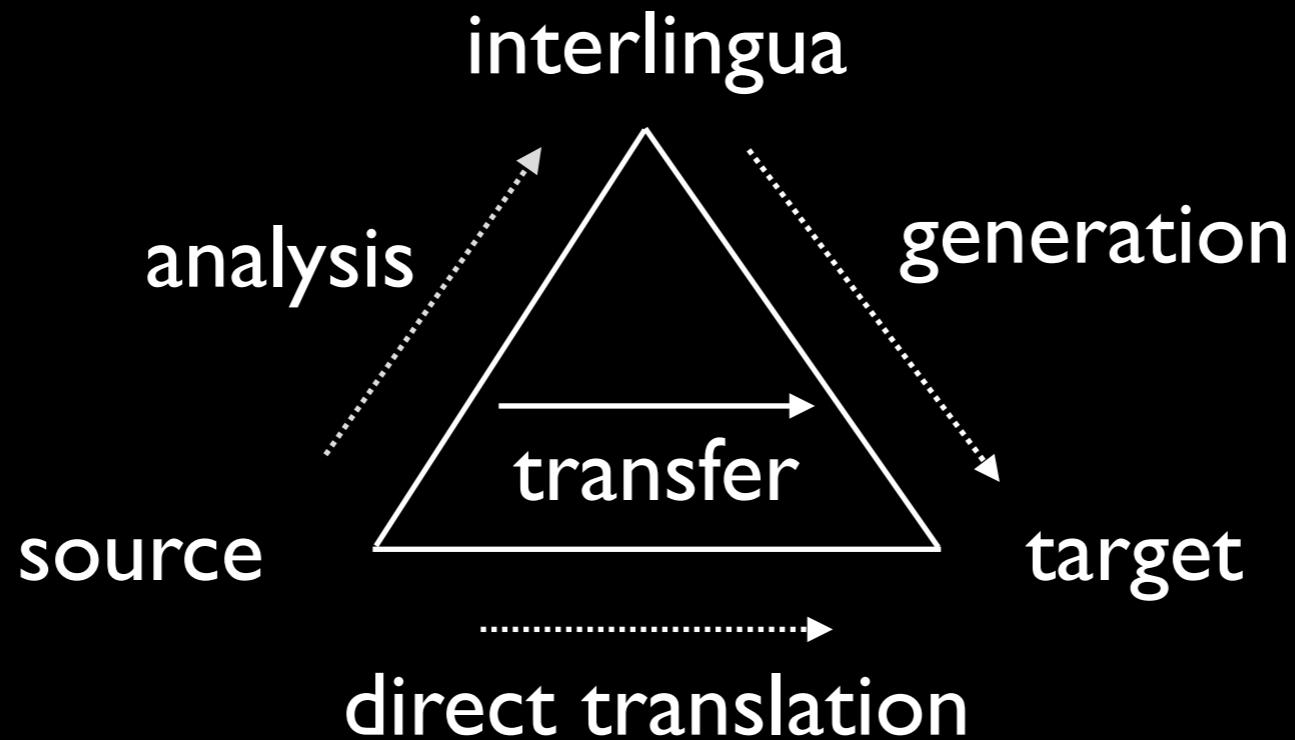
the concept of interlingua



the concept of interlingua



the concept of interlingua

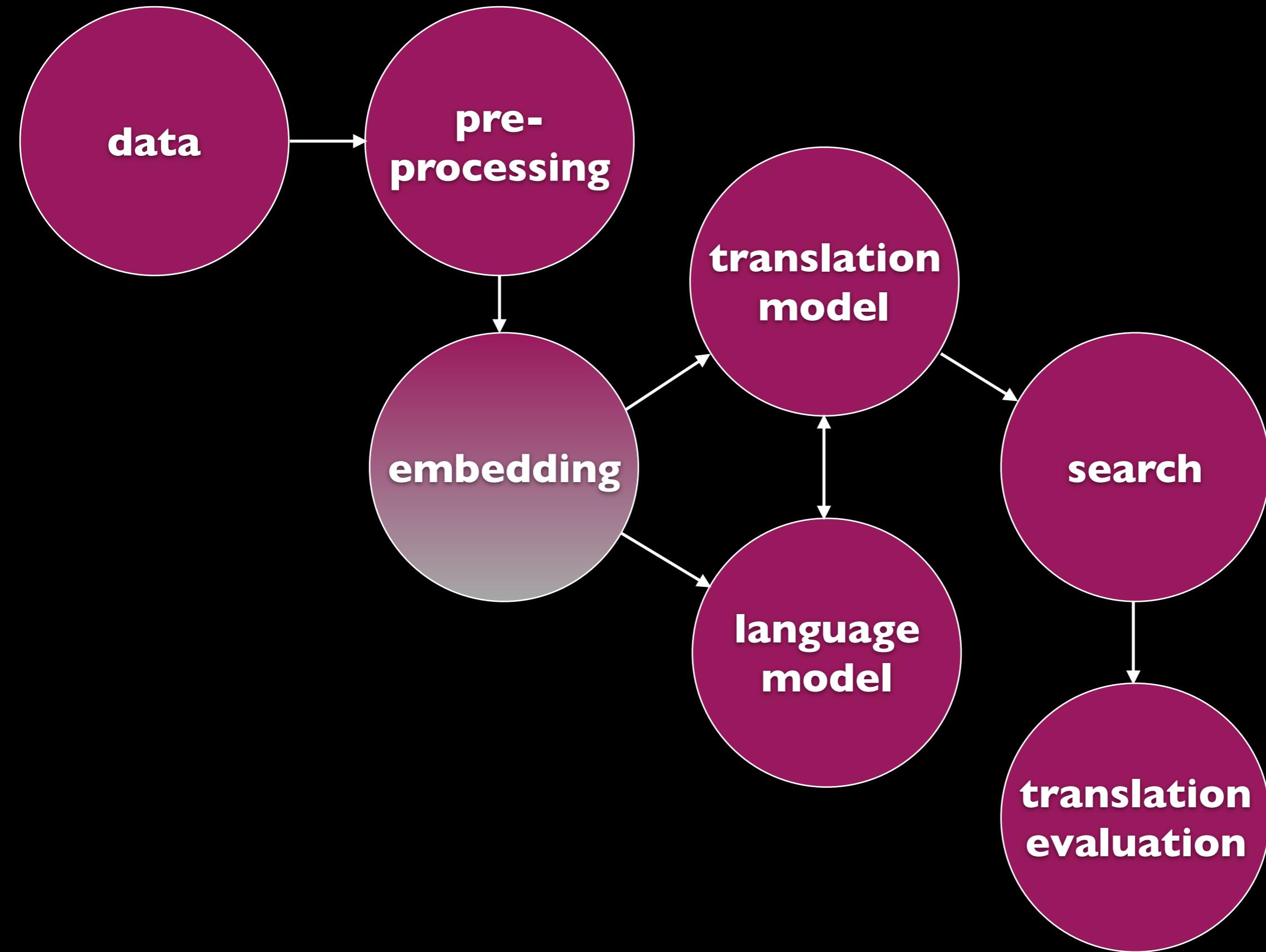


machine translation components

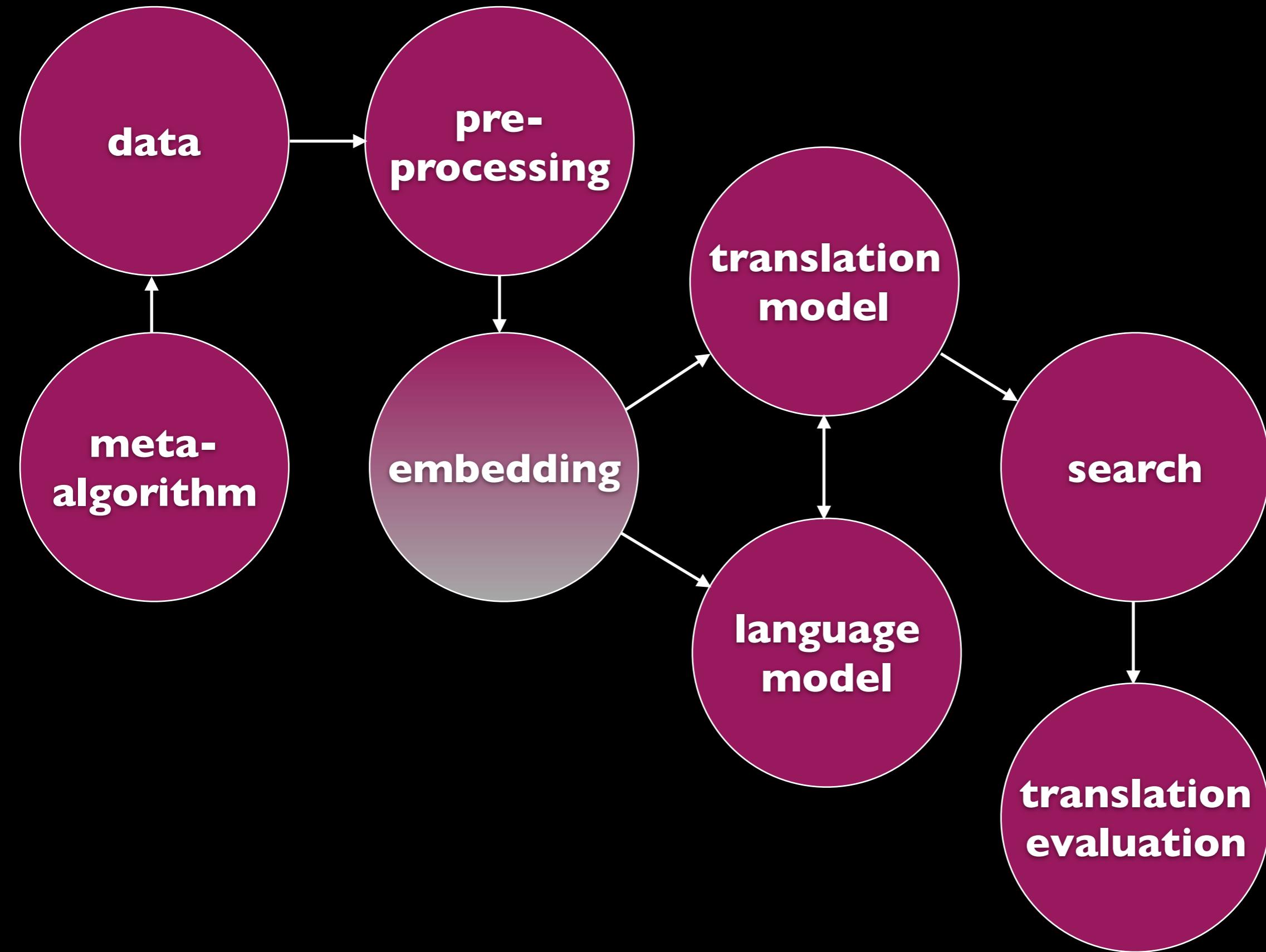


data

machine translation components



machine translation components

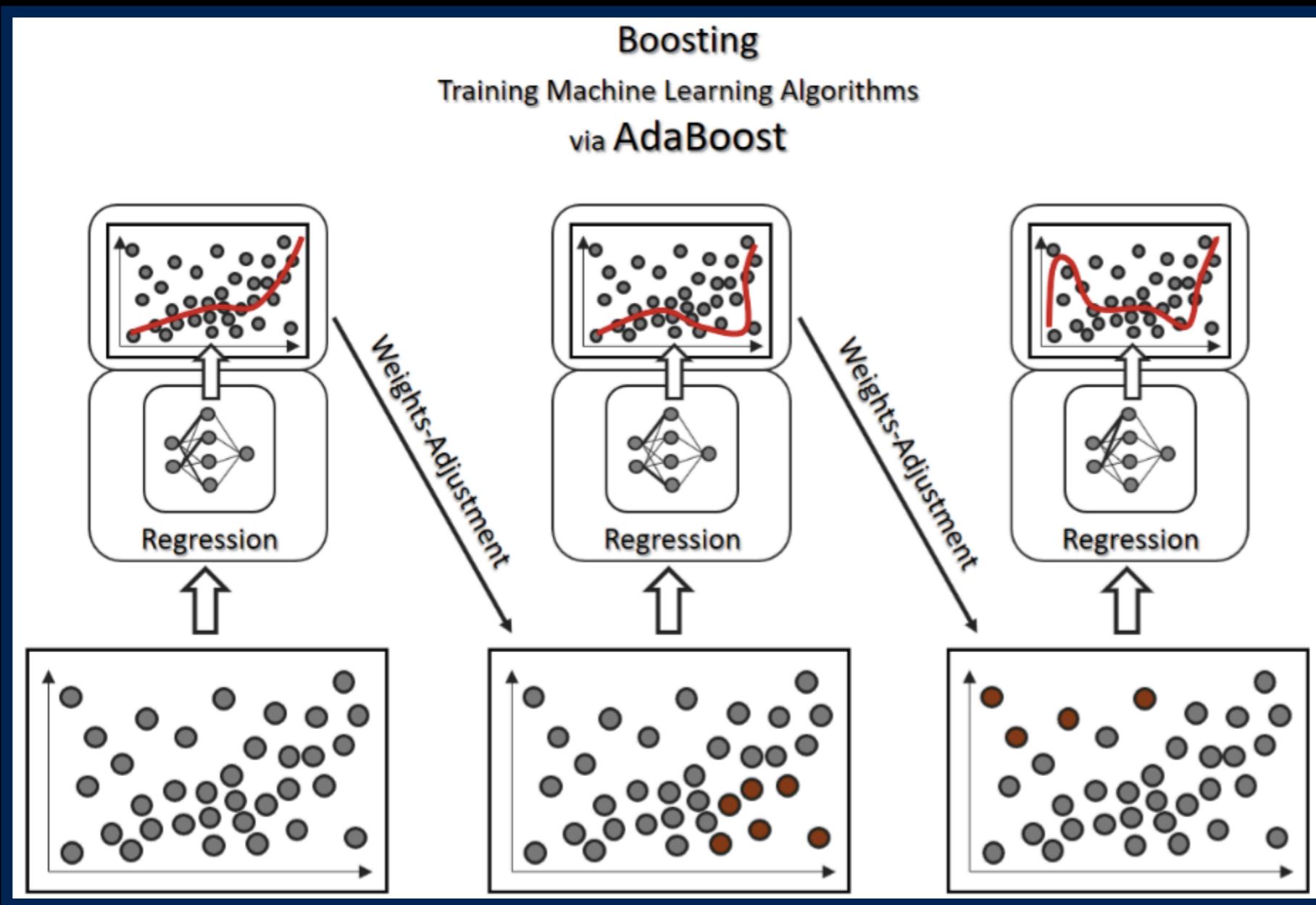


machine translation components

**meta-
algorithm**

boosting

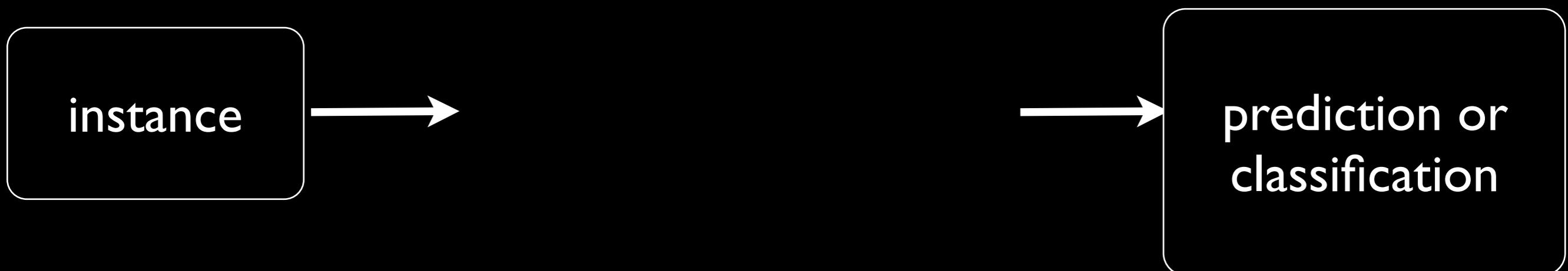
- improves prediction accuracy; non-parallelizable



subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.
Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.
That cannot continue.
That concludes the debate.



subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

I hat cannot continue.

That concludes the debate.

instance



prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

I hat cannot continue.

That concludes the debate.



classifier I

instance



prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.



classifier I

instance



prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

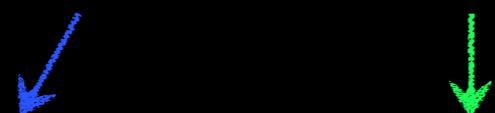
Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.



classifier 1

classifier 2

instance

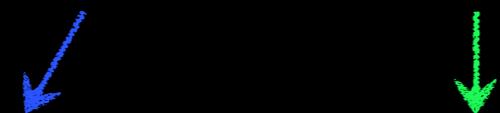


prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.
Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.
That cannot continue.
That concludes the debate.



classifier 1

classifier 2

instance



prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.
Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.
That cannot continue.
That concludes the debate.



classifier 1

classifier 2

classifier 3

instance



prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.
Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.
That cannot continue.
That concludes the debate.



classifier 1

classifier 2

classifier 3

instance

vote
or averaging

prediction or
classification

subsamples generate multiple classifiers

Natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.



classifier 1

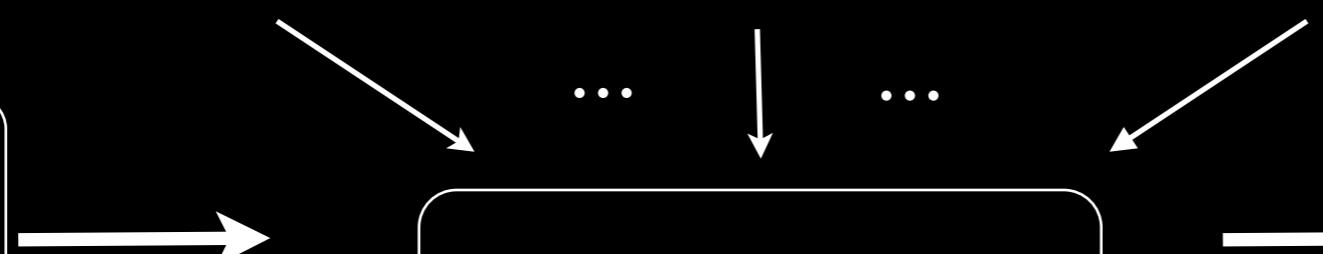
classifier 2

classifier 3

instance

vote
or averaging

prediction or
classification



subsamples generate multiple classifiers

In naturale Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.

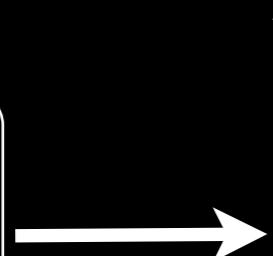


classifier 1

classifier 2

classifier 3

instance



vote
or averaging

prediction or
classification

subsamples generate multiple classifiers

In naturale Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.



instance

vote
or averaging

prediction or
classification

subsamples generate multiple classifiers

In naturale Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

Natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.

classifier 1

classifier 2

classifier 3

different ways to
subsample

instance

vote
or averaging

prediction or
classification

better subsampling, better translation?

analyze with abstract formulation for all applications,
algorithms, datasets and domains

subsamples generate multiple classifiers

In naturale Lebensräume wurden zerstört.
Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.
Die Aussprache ist geschlossen.

Natural habitats were destroyed.
This is a major task.

That cannot continue.
That concludes the debate.



classifier 1

classifier 2

classifier 3

instance

vote
or averaging

prediction or
classification

subsamples generate multiple classifiers

each sample ● is
a *parallel sentence*

In natürliche Lebensräume wurden zerstört.

Dies ist eine ihrer Hauptaufgaben.

Das kann so nicht weitergehen.

Die Aussprache ist geschlossen.

In natural habitats were destroyed.

This is a major task.

That cannot continue.

That concludes the debate.



classifier 1

classifier 2

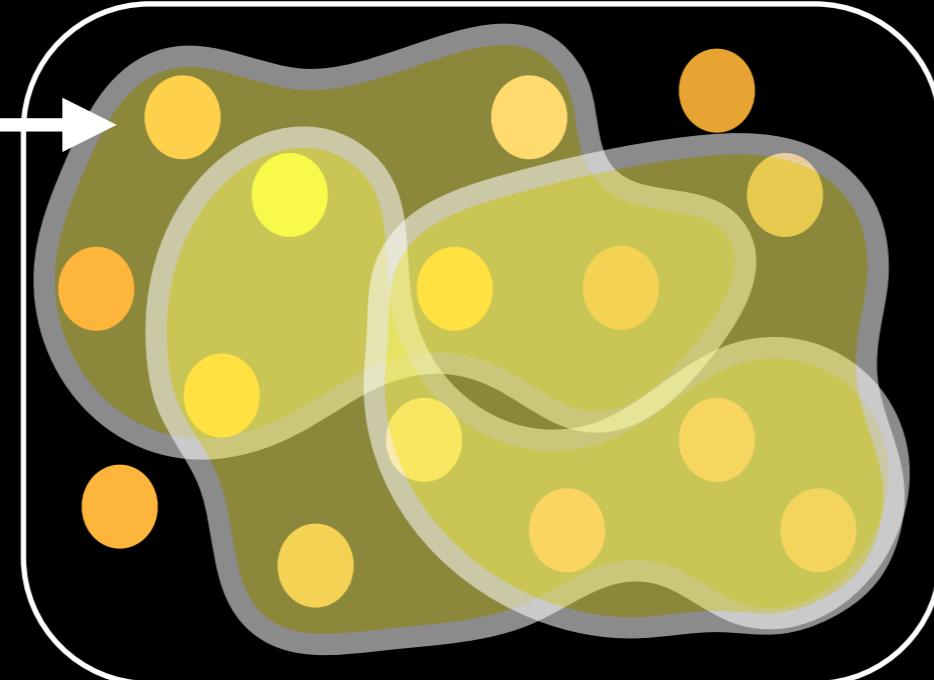
classifier 3

instance

vote
or averaging

prediction or
classification

a sample:
instance+class



classifier 1

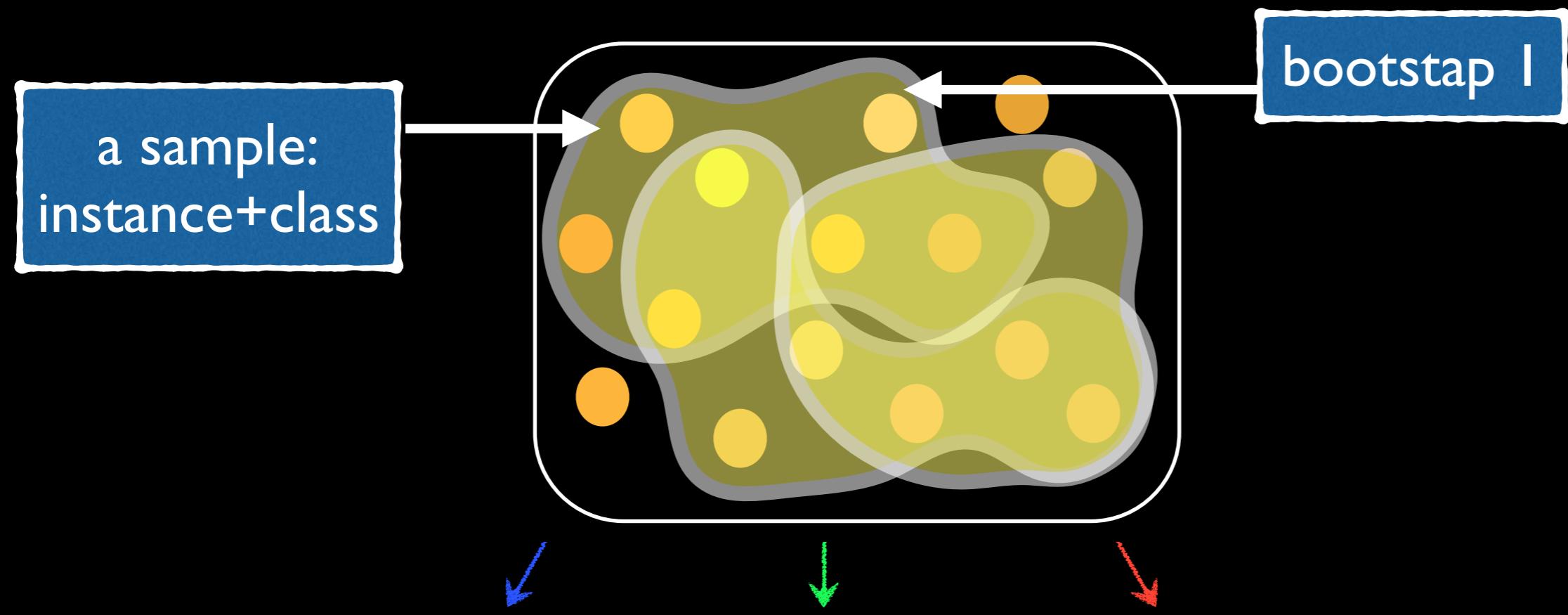
classifier 2

classifier 3

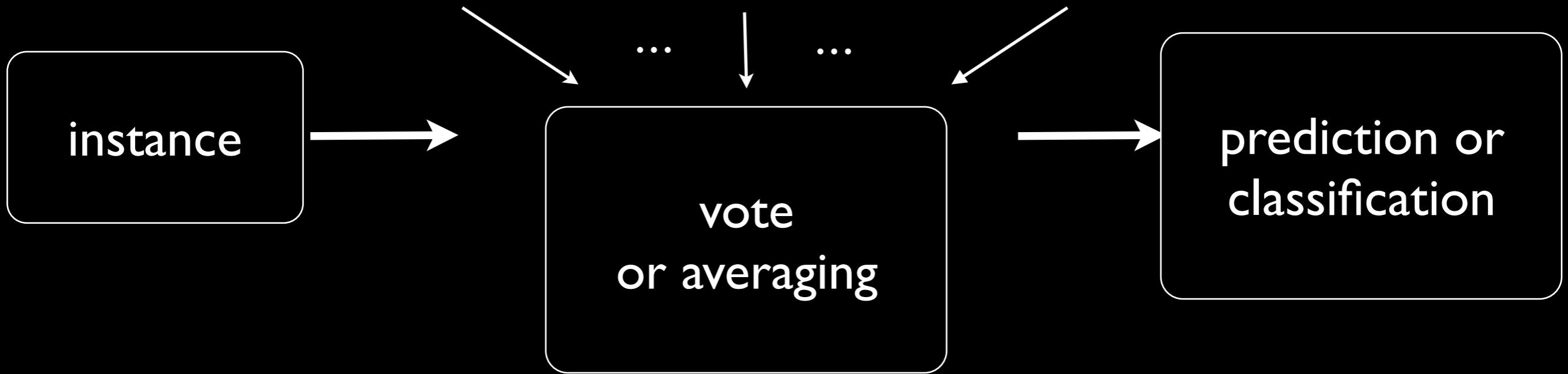
instance

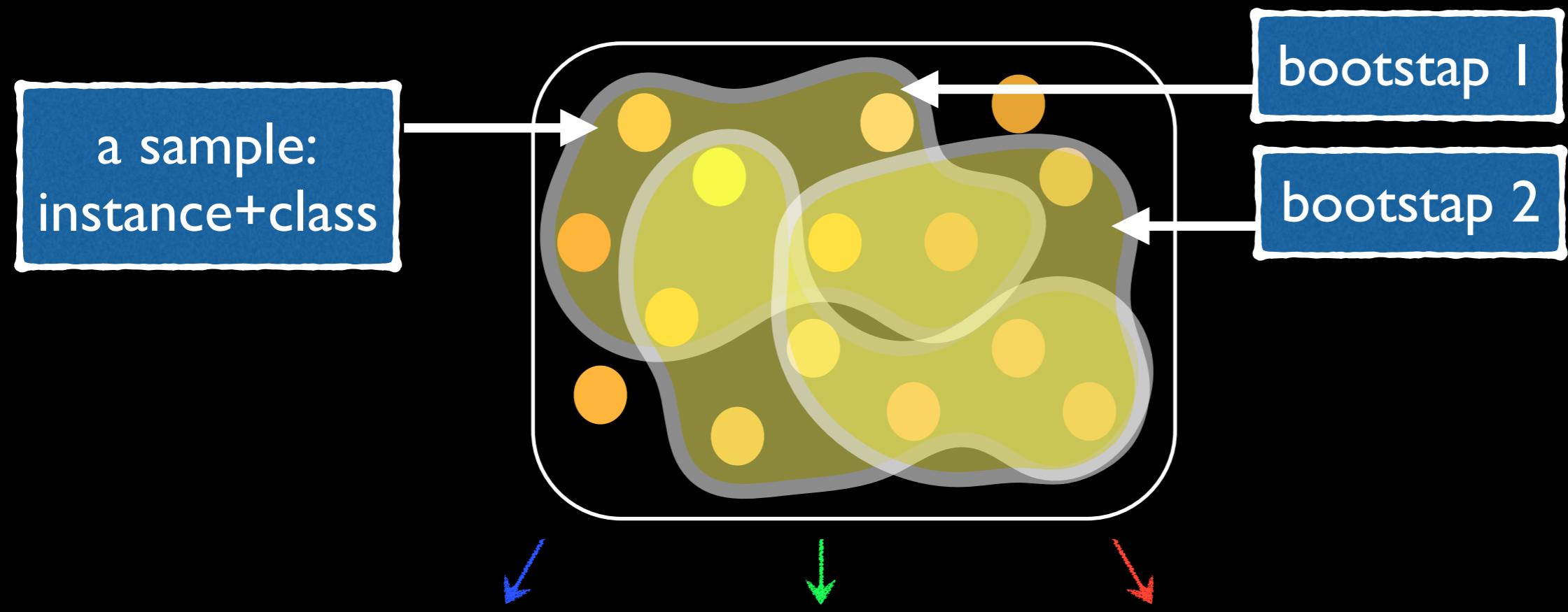
vote
or averaging

prediction or
classification

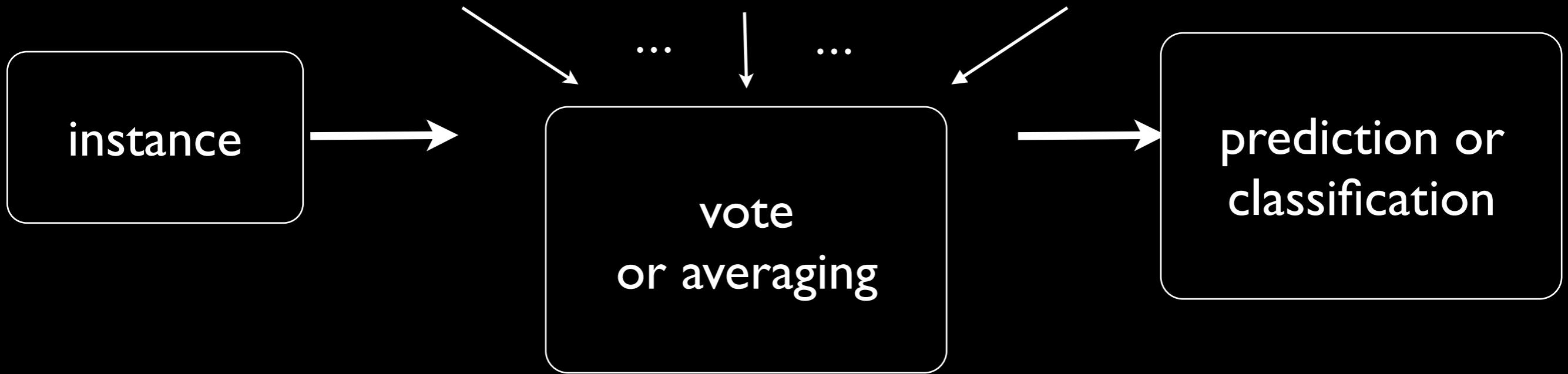


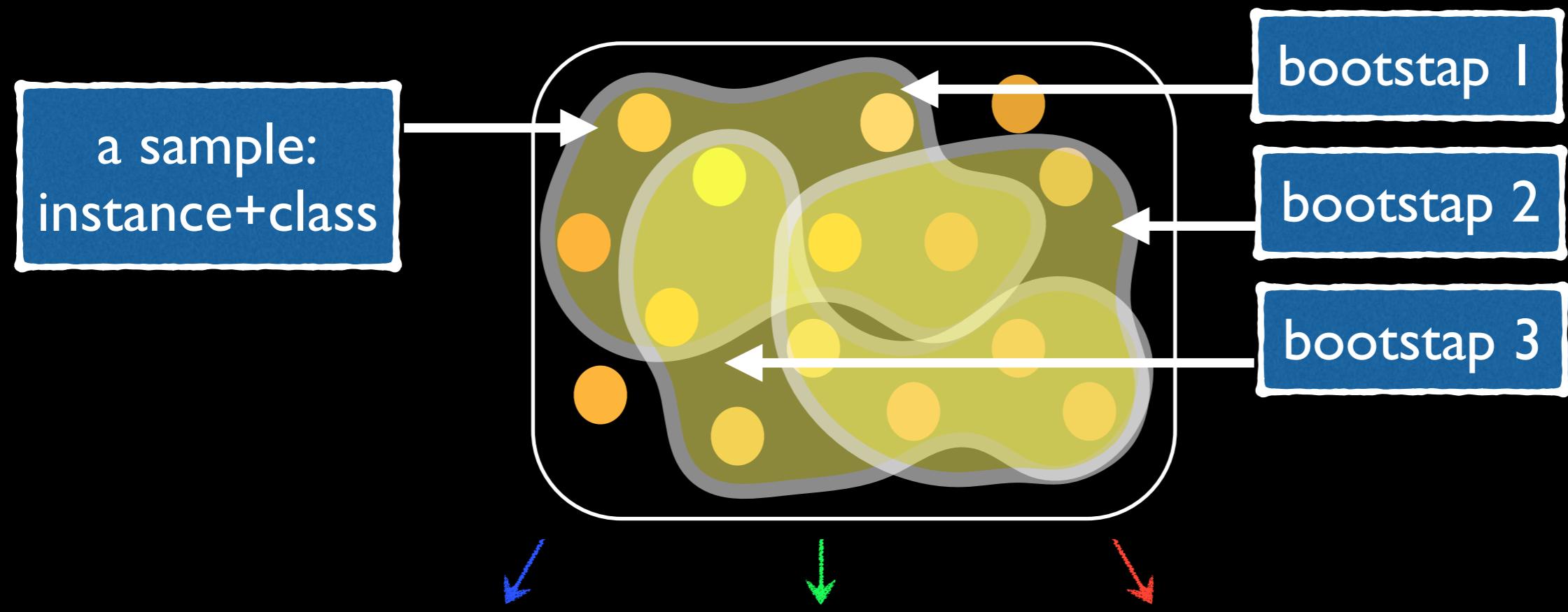
classifier 1 classifier 2 classifier 3



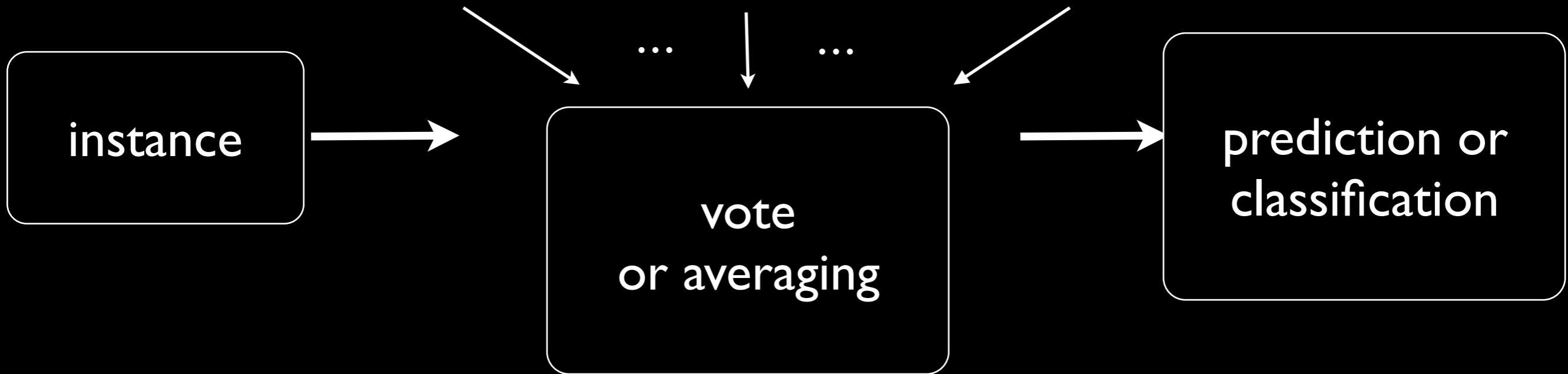


classifier 1 classifier 2 classifier 3



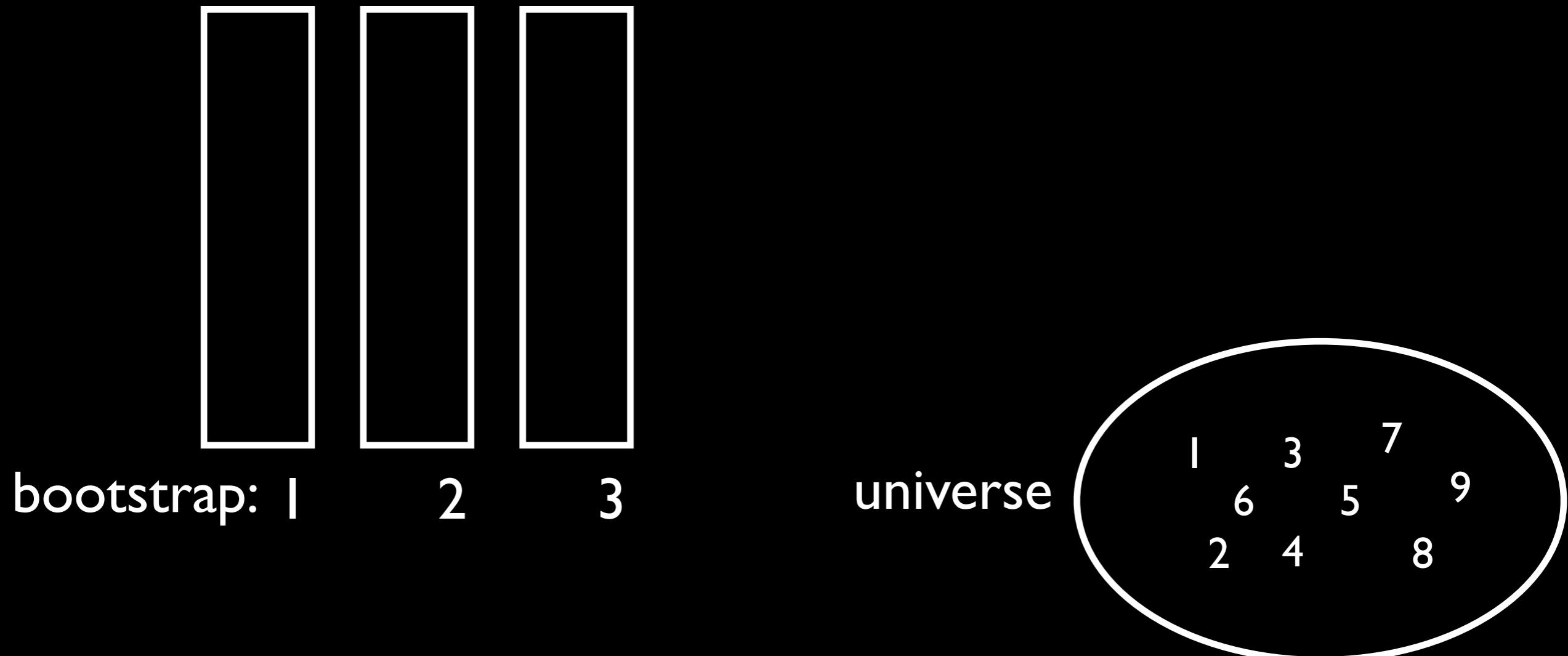


classifier 1 classifier 2 classifier 3



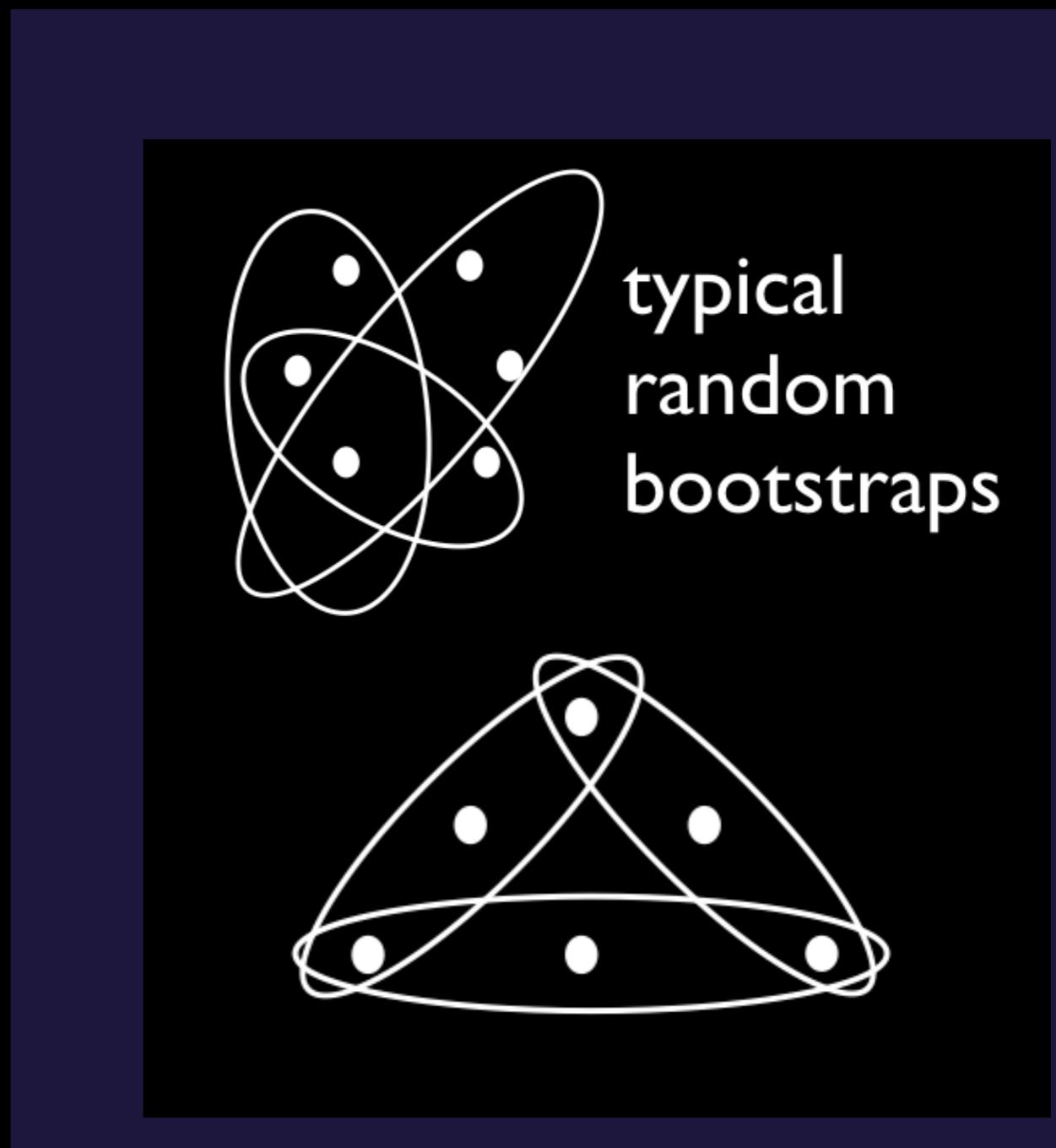
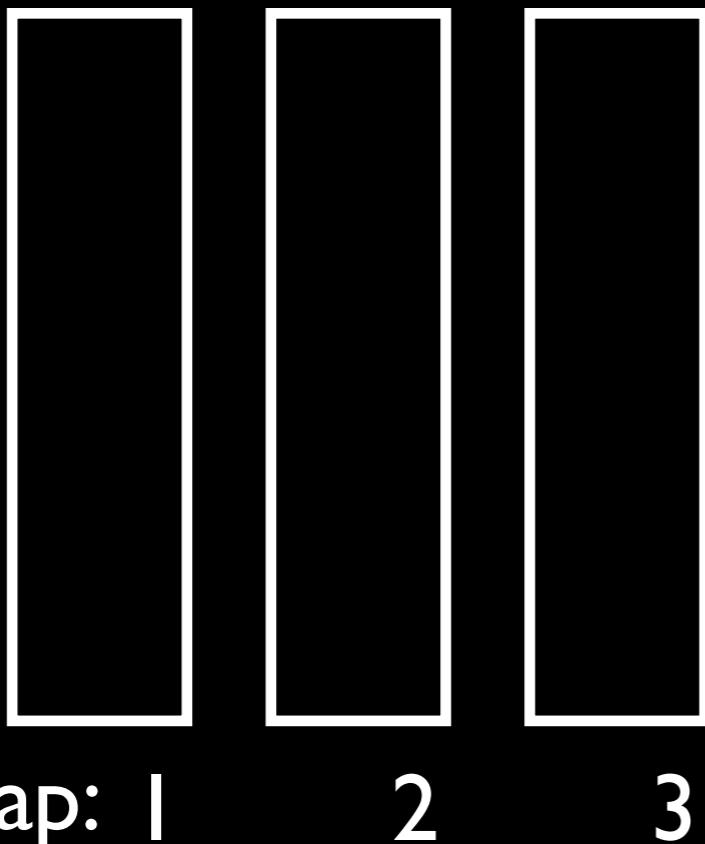
bootstrapping with combinatorial design

N=9, m=3, b=6



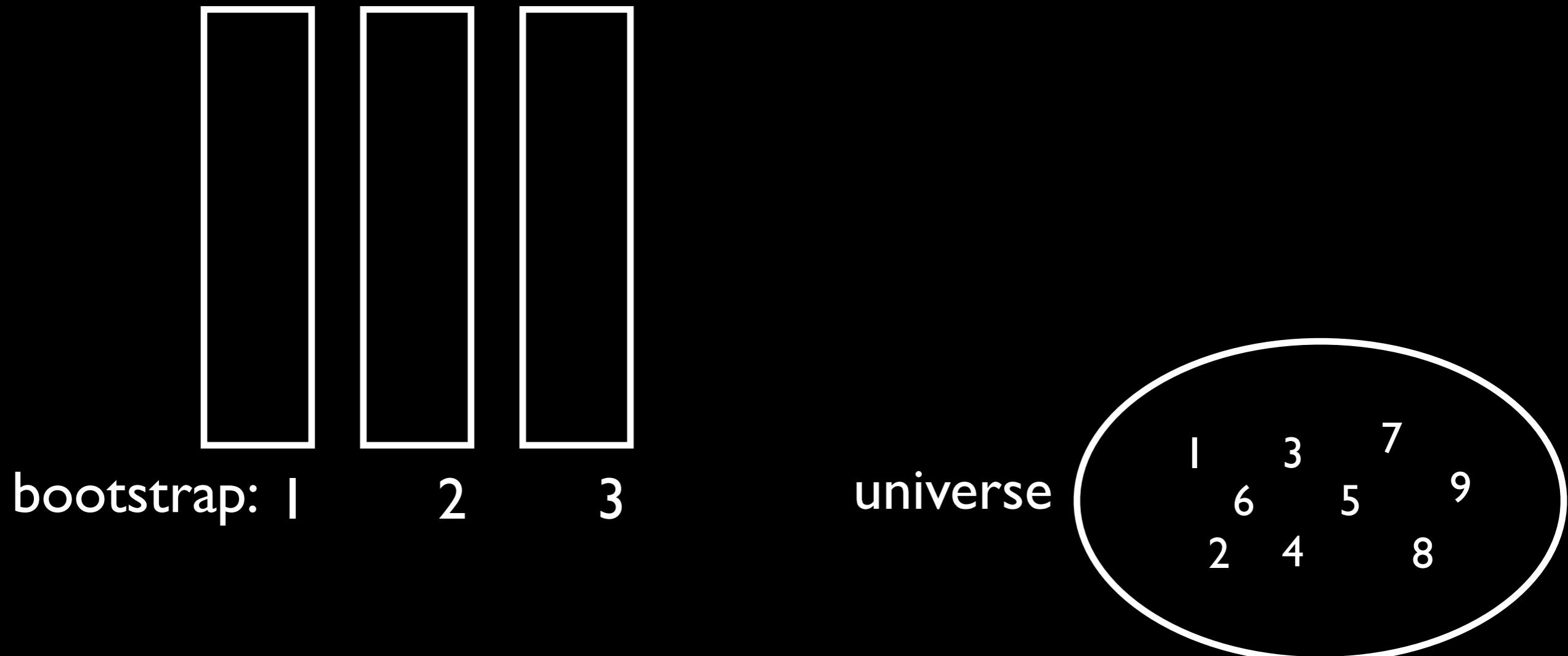
bootstrapping with combinatorial design

$N=9, m=3, b=6$



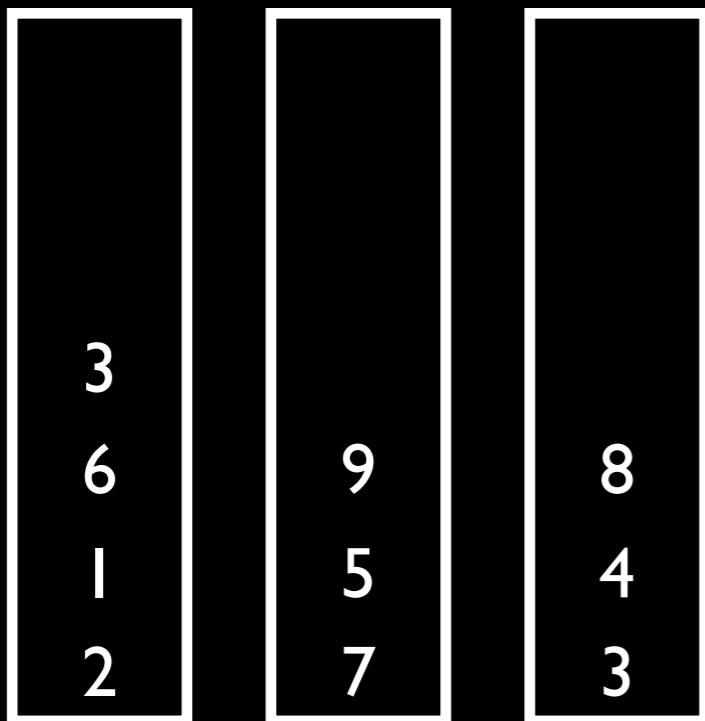
bootstrapping with combinatorial design

N=9, m=3, b=6



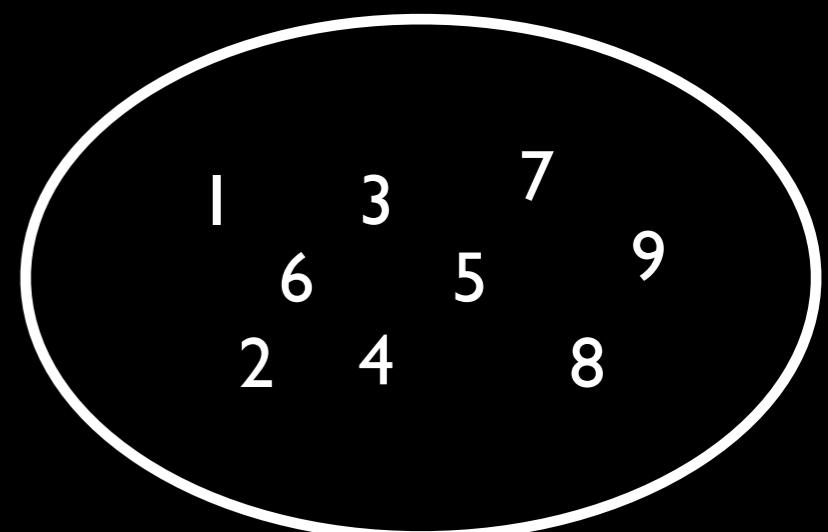
bootstrapping with combinatorial design

N=9, m=3, b=6



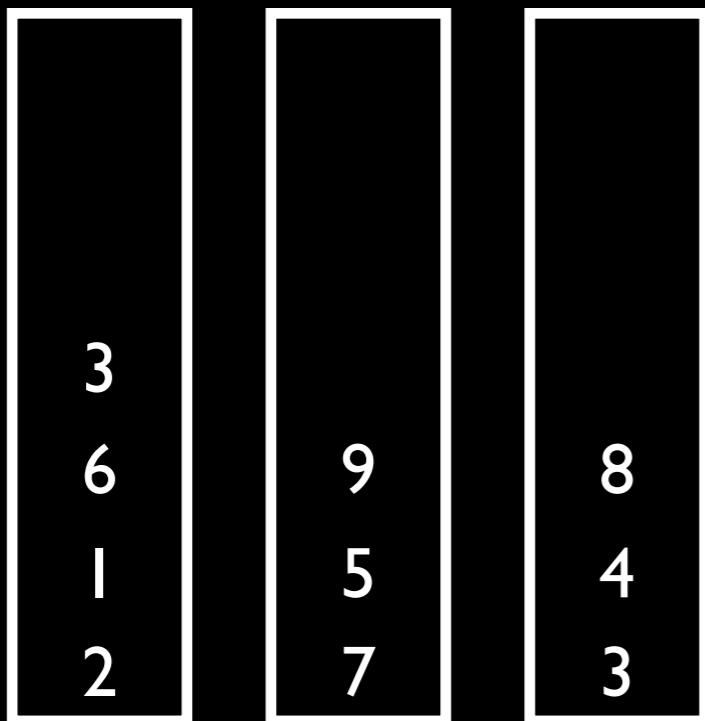
bootstrap: 1 2 3

universe



bootstrapping with combinatorial design

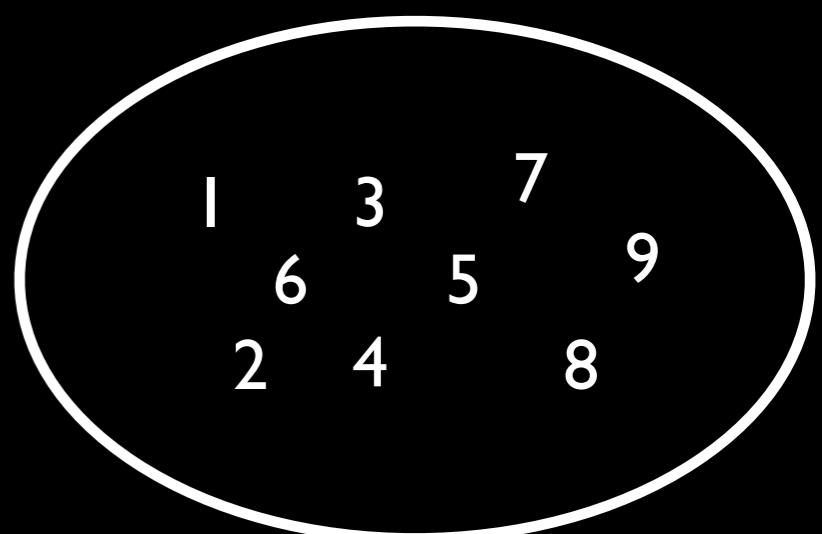
N=9, m=3, b=6



bootstrap: 1 2 3

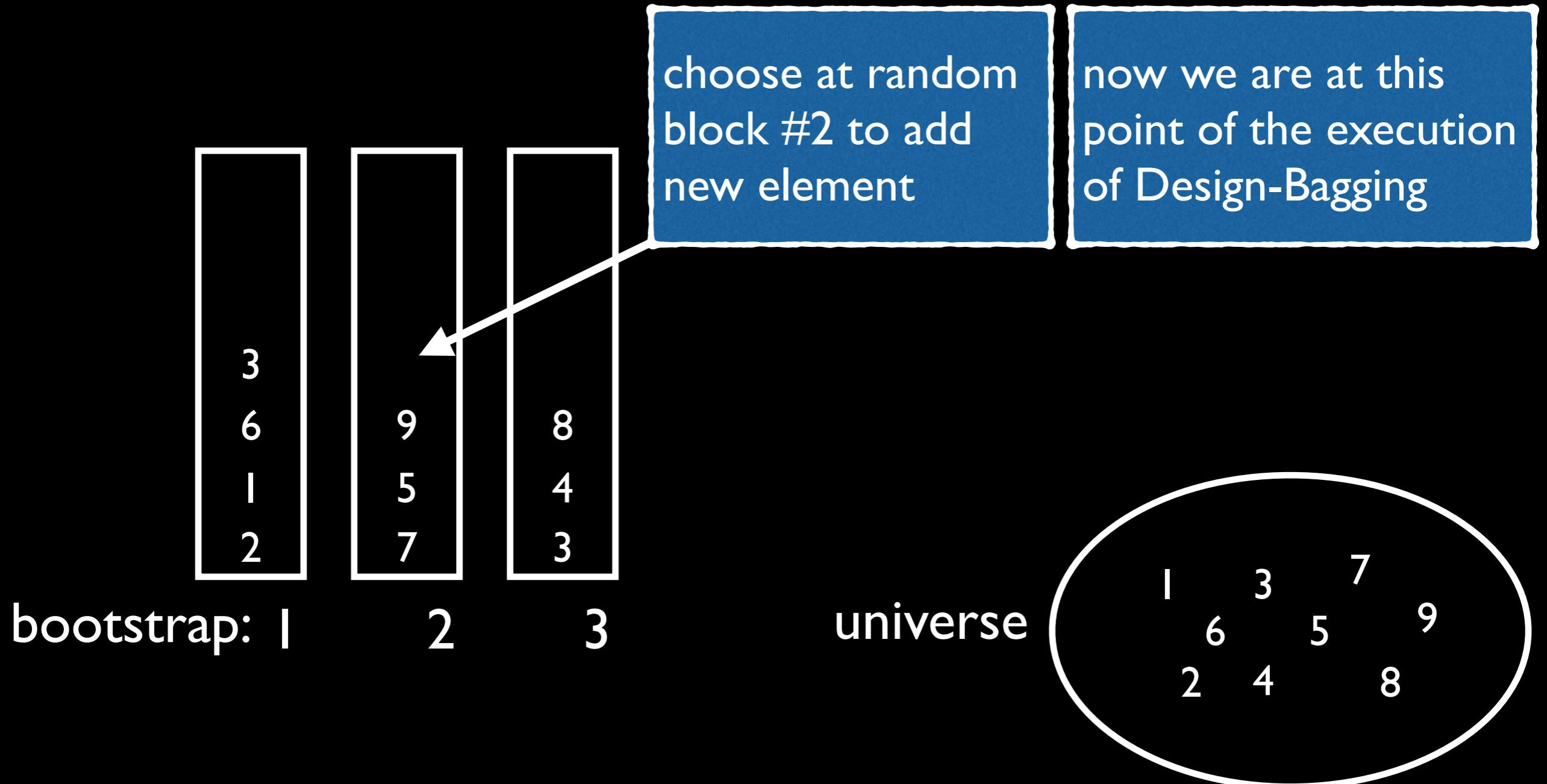
universe

now we are at this
point of the execution
of Design-Bagging



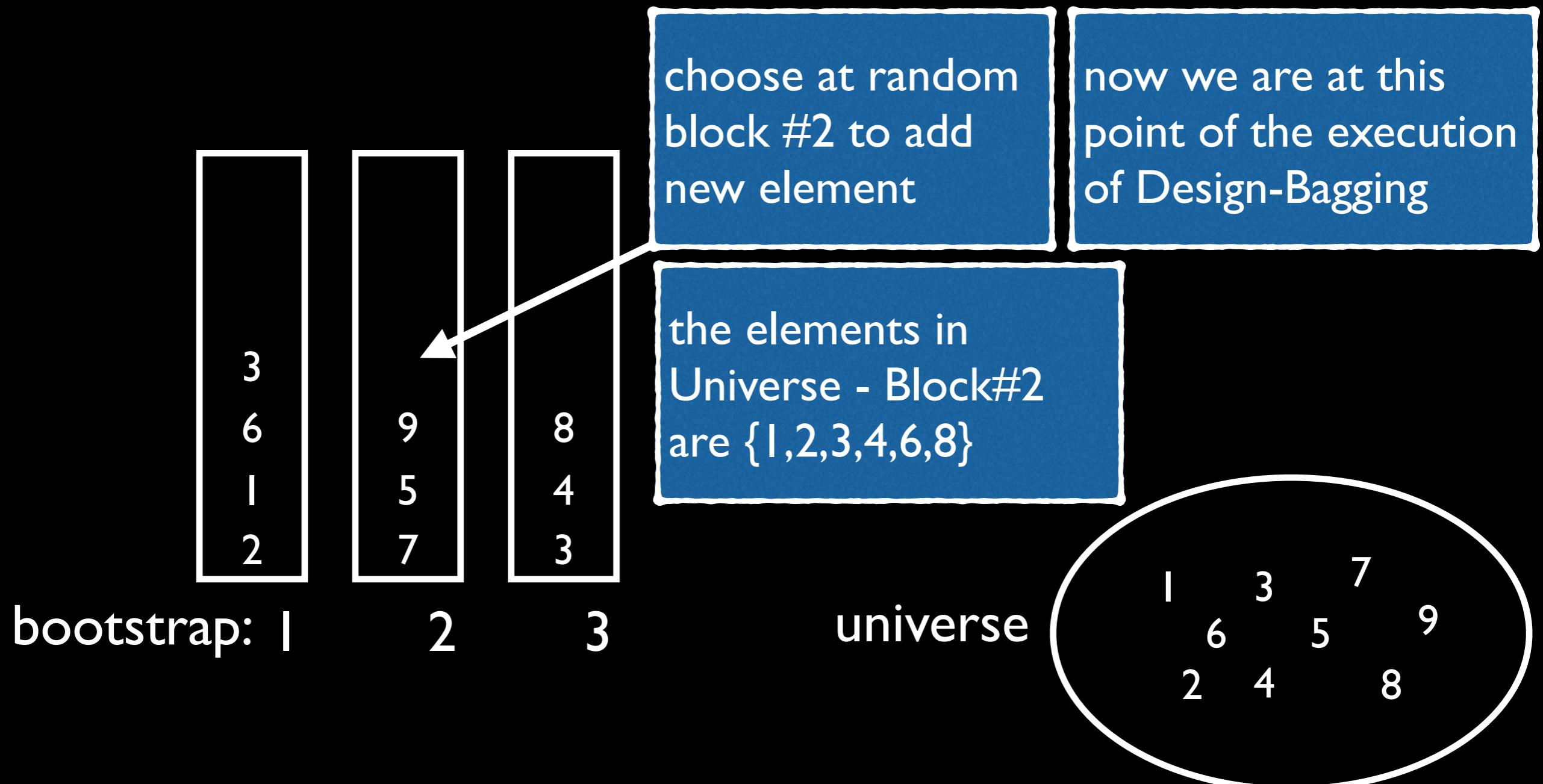
bootstrapping with combinatorial design

N=9, m=3, b=6



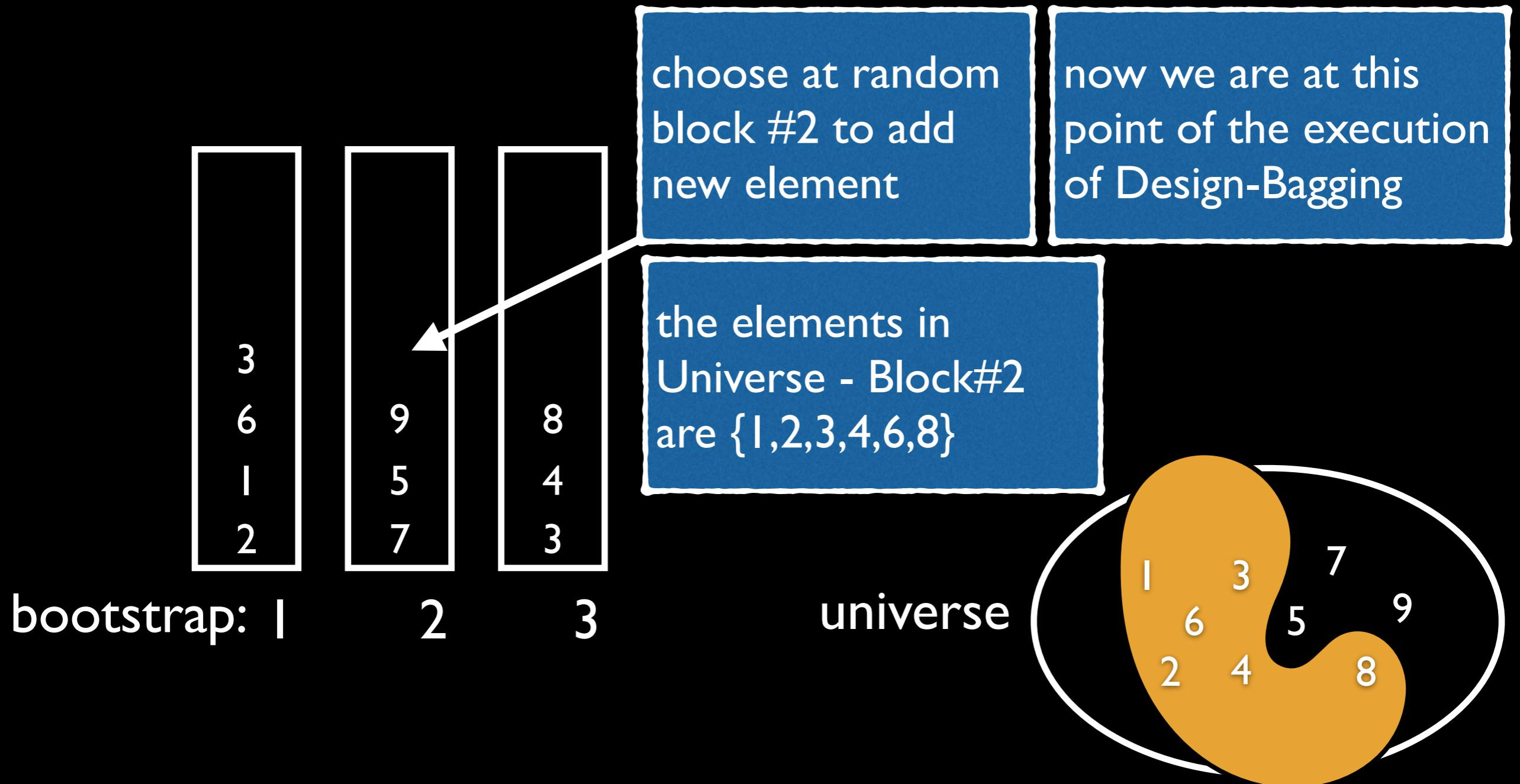
bootstrapping with combinatorial design

N=9, m=3, b=6



bootstrapping with combinatorial design

N=9, m=3, b=6



bootstrapping with combinatorial design

N=9, m=3, b=6

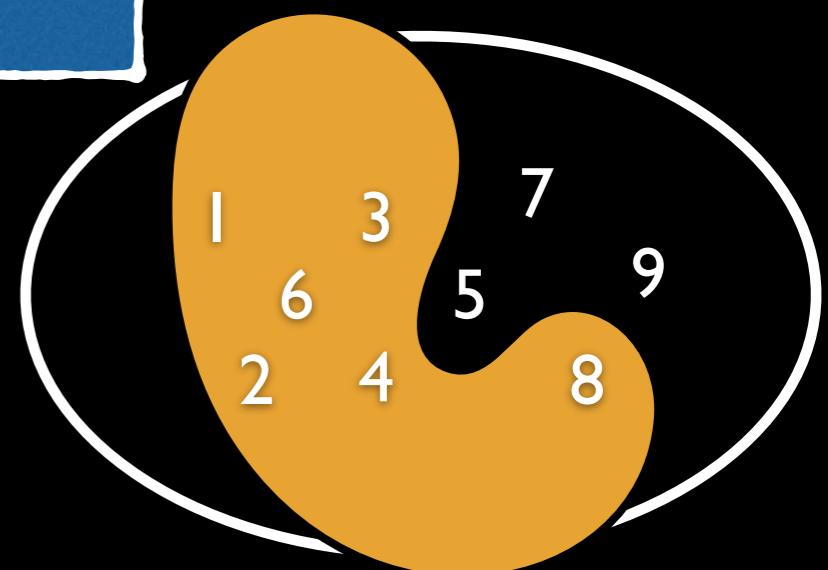
3	9	8
6	5	4
1	7	3
2		

bootstrap: 1 2 3

among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

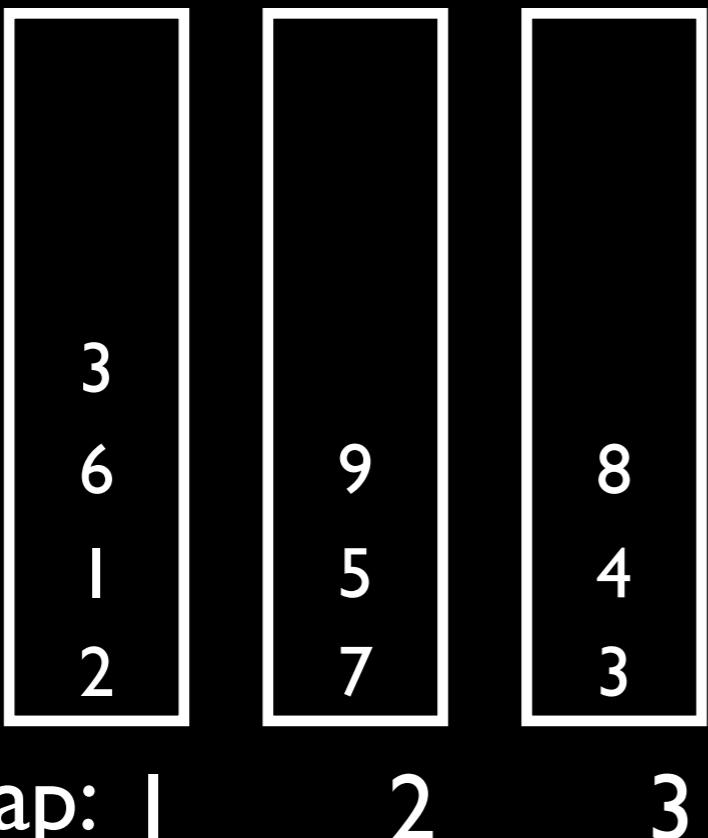
the elements in
Universe - Block#2
are {1,2,3,4,6,8}

universe



bootstrapping with combinatorial design

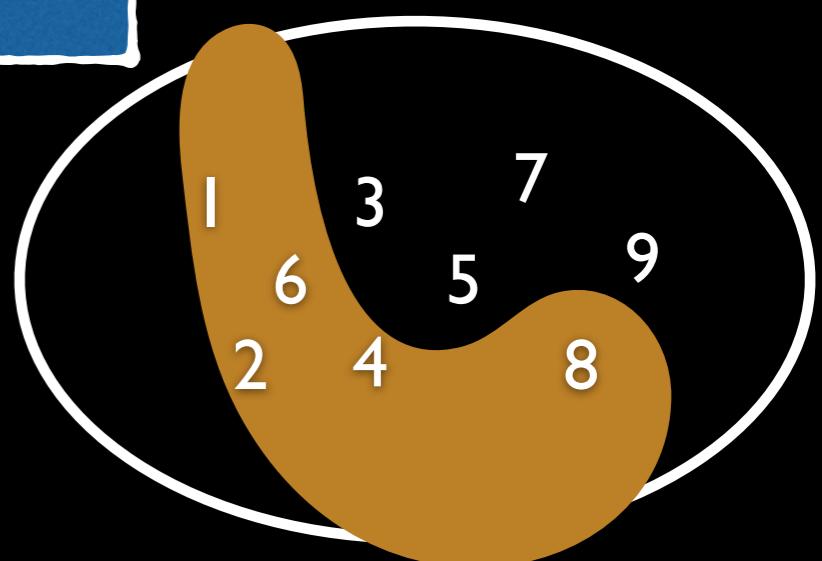
N=9, m=3, b=6



among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

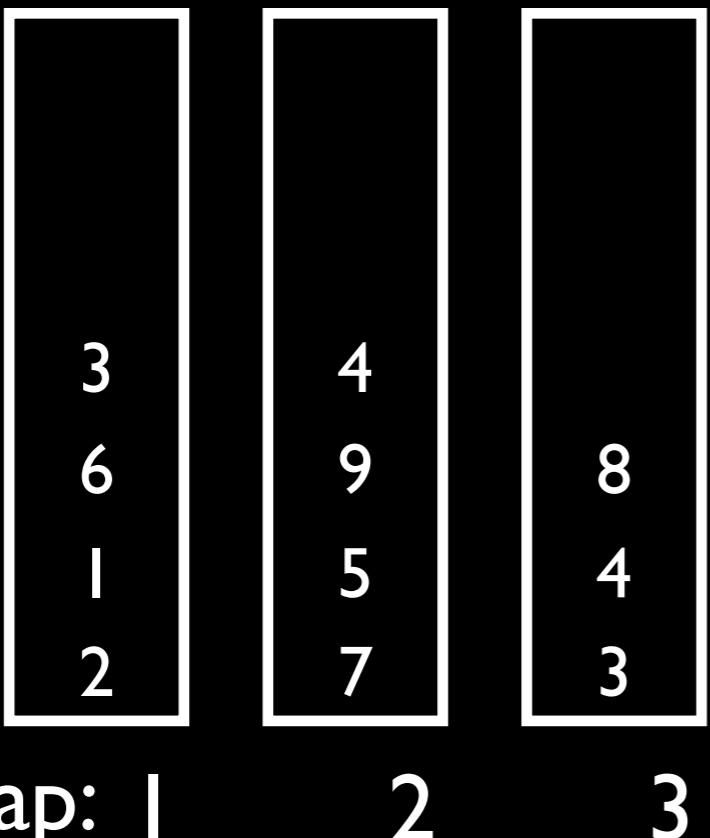
the elements in
Universe - Block#2
are {1,2,3,4,6,8}

universe



bootstrapping with combinatorial design

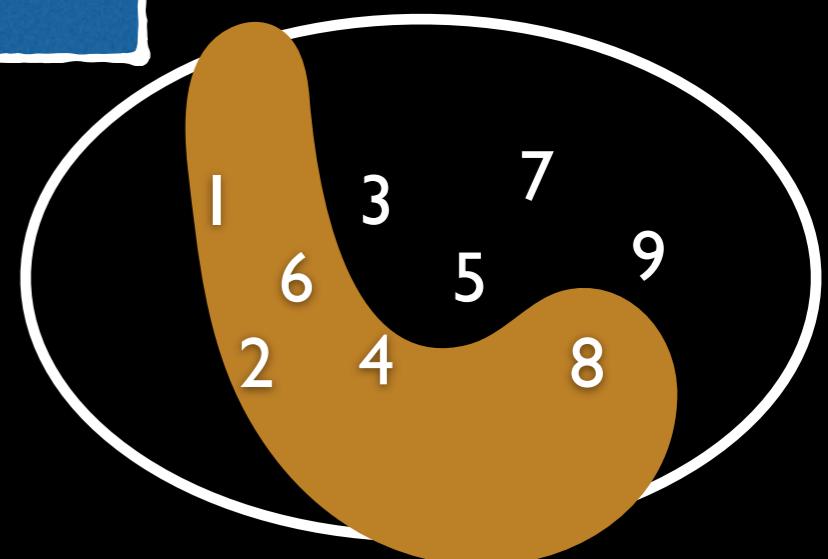
N=9, m=3, b=6



among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

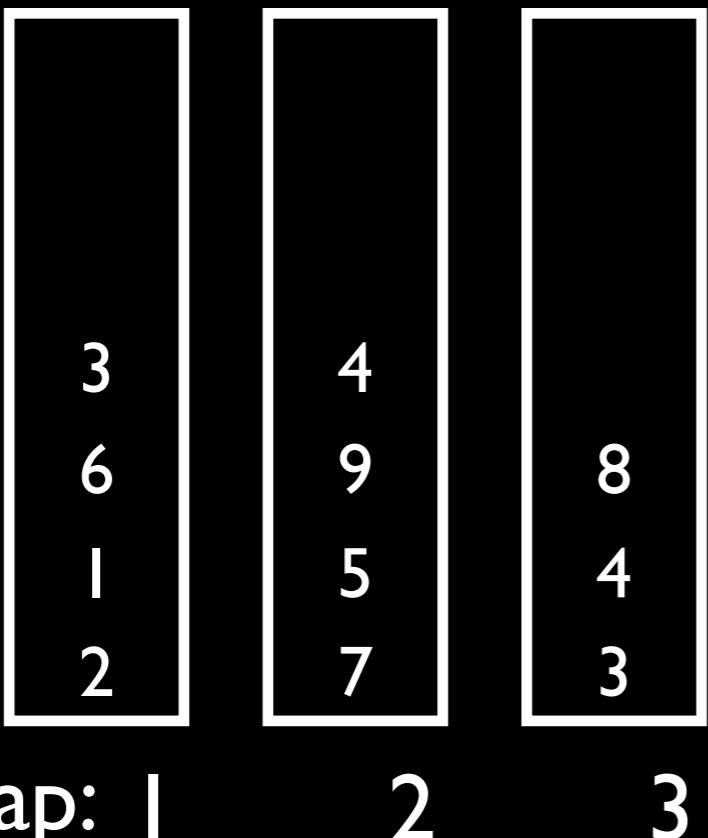
sampled: 4
— chosen uniformly at
random from {1,2,4,6,8} —

universe



bootstrapping with combinatorial design

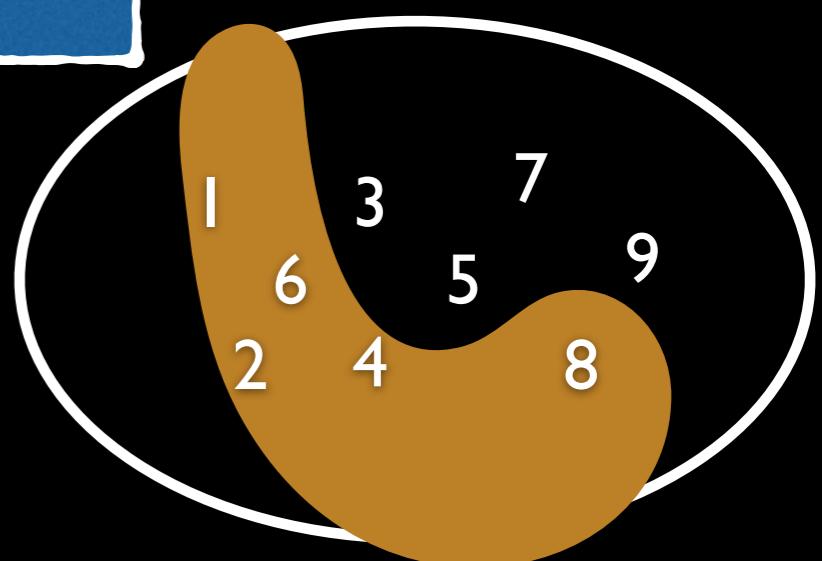
N=9, m=3, b=6



among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

sampled: 4
— chosen uniformly at
random from {1,2,4,6,8} —

universe



this is a biased sampling process

bootstrapping with combinatorial design

N=9, m=3, b=6

Question #24: enhance bootstrapping

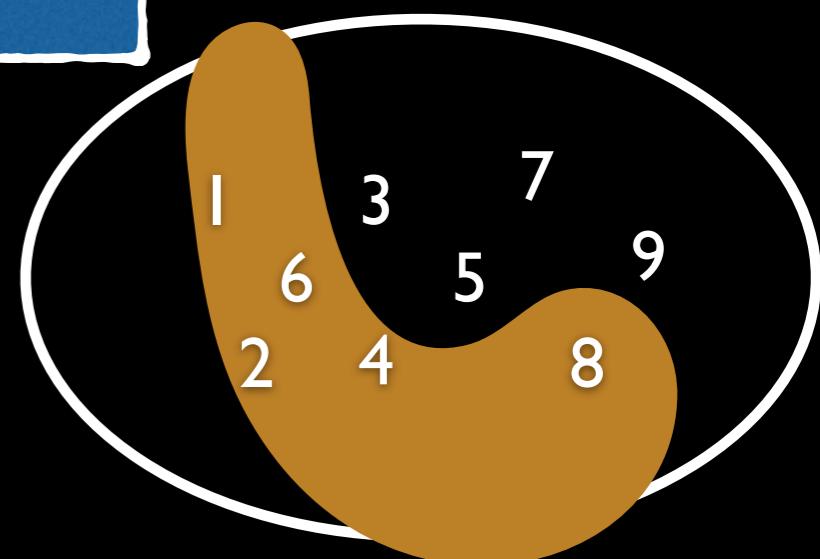
3	4	8
6	9	4
1	5	4
2	7	3

bootstrap: 1 2 3

among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

sampled: 4
— chosen uniformly at
random from {1,2,4,6,8} —

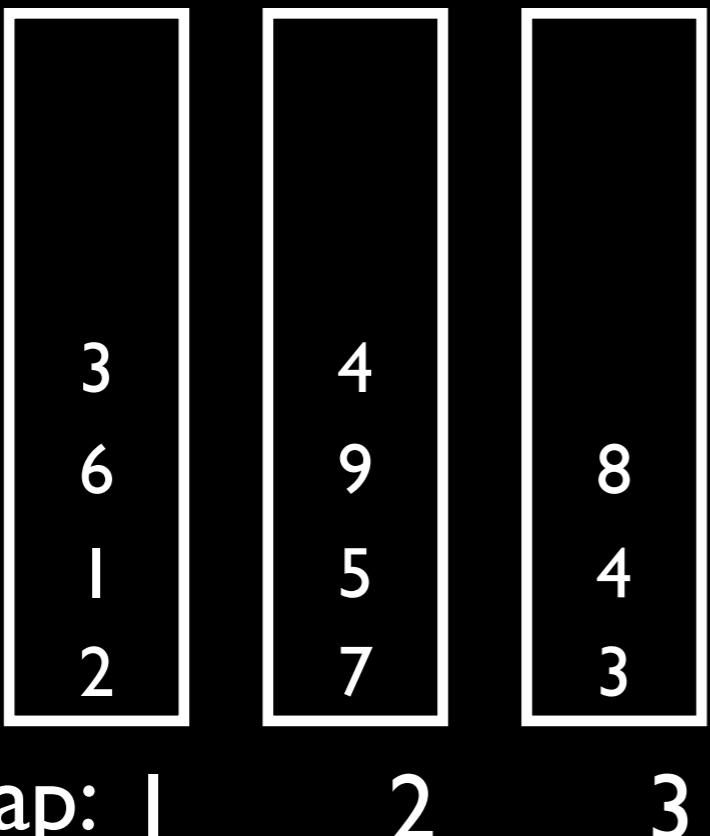
universe



this is a biased sampling process

bootstrapping with combinatorial design

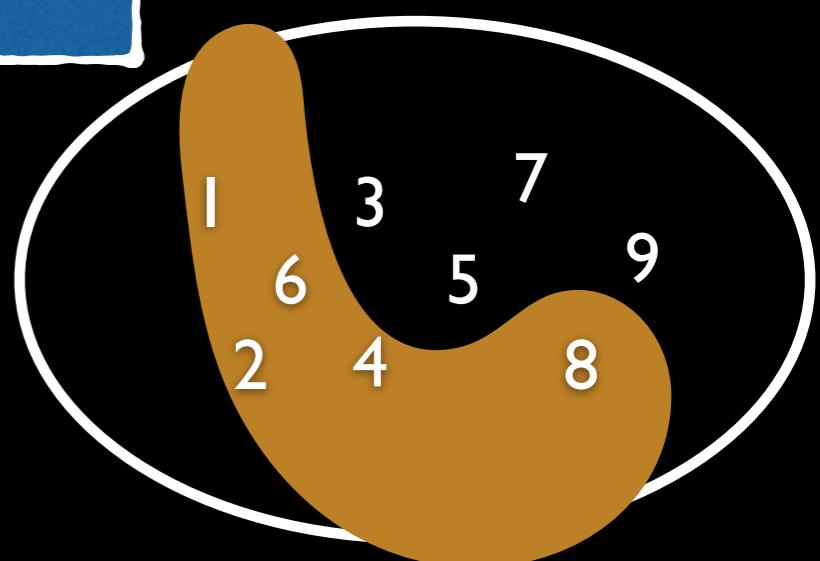
N=9, m=3, b=6



among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

sampled: 4
— chosen uniformly at
random from {1,2,4,6,8} —

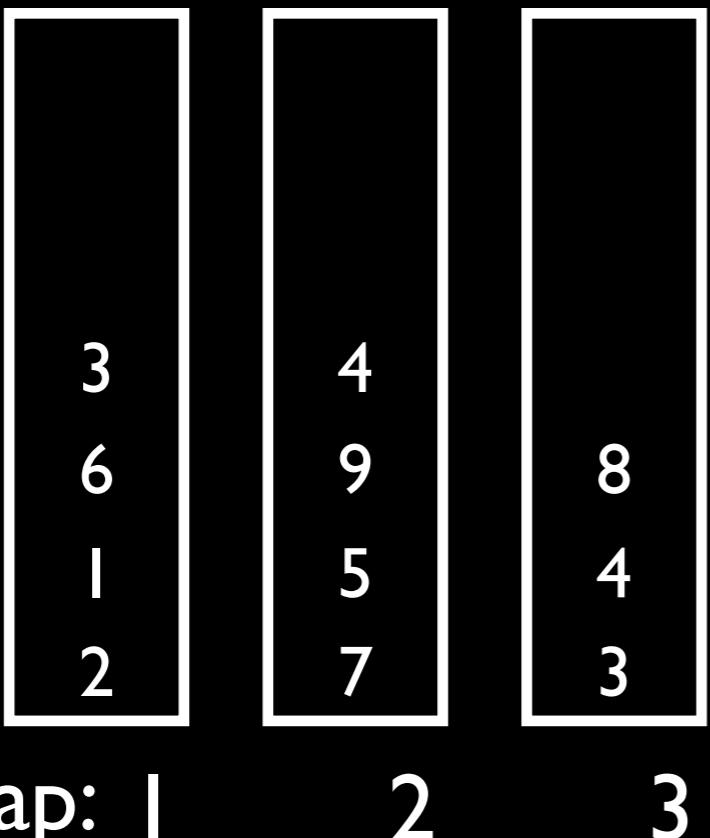
universe



this is a biased sampling process

bootstrapping with combinatorial design

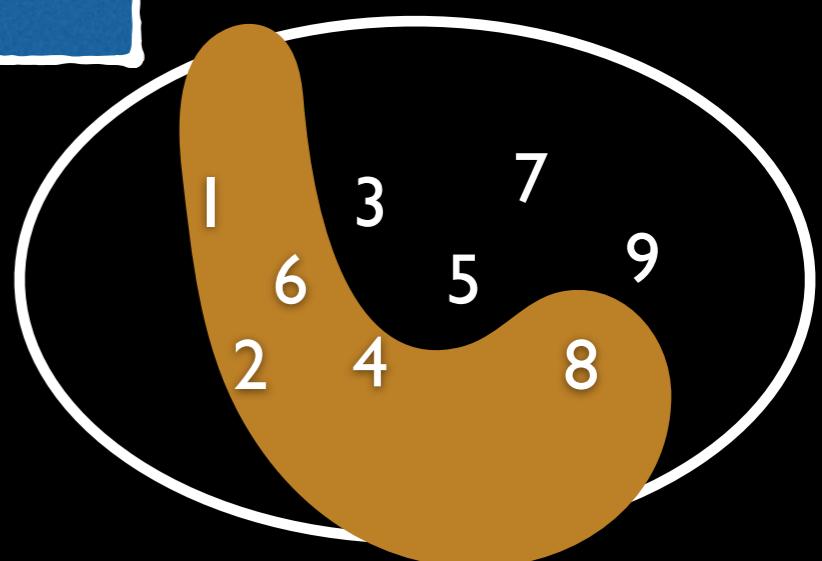
N=9, m=3, b=6



among {1,2,3,4,6,8} those
that appear the least
are {1,2,4,6,8}

sampled: 4
— chosen uniformly at
random from {1,2,4,6,8} —

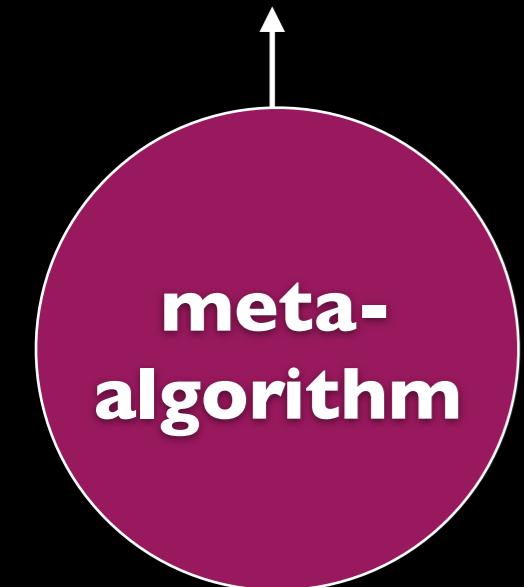
universe



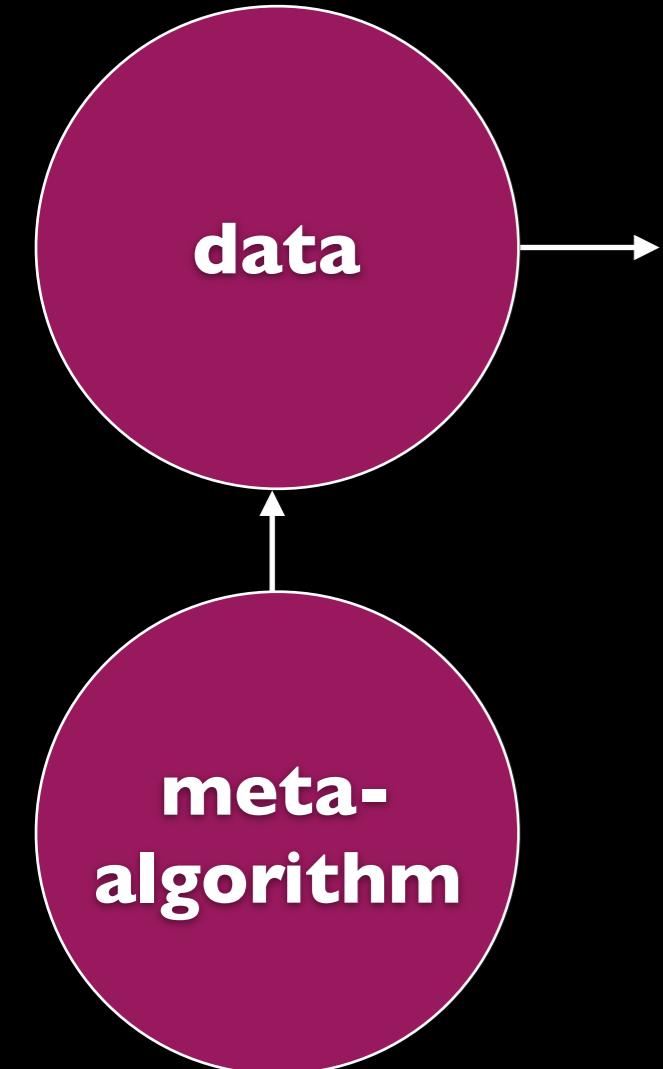
this is a biased sampling process

machine translation components

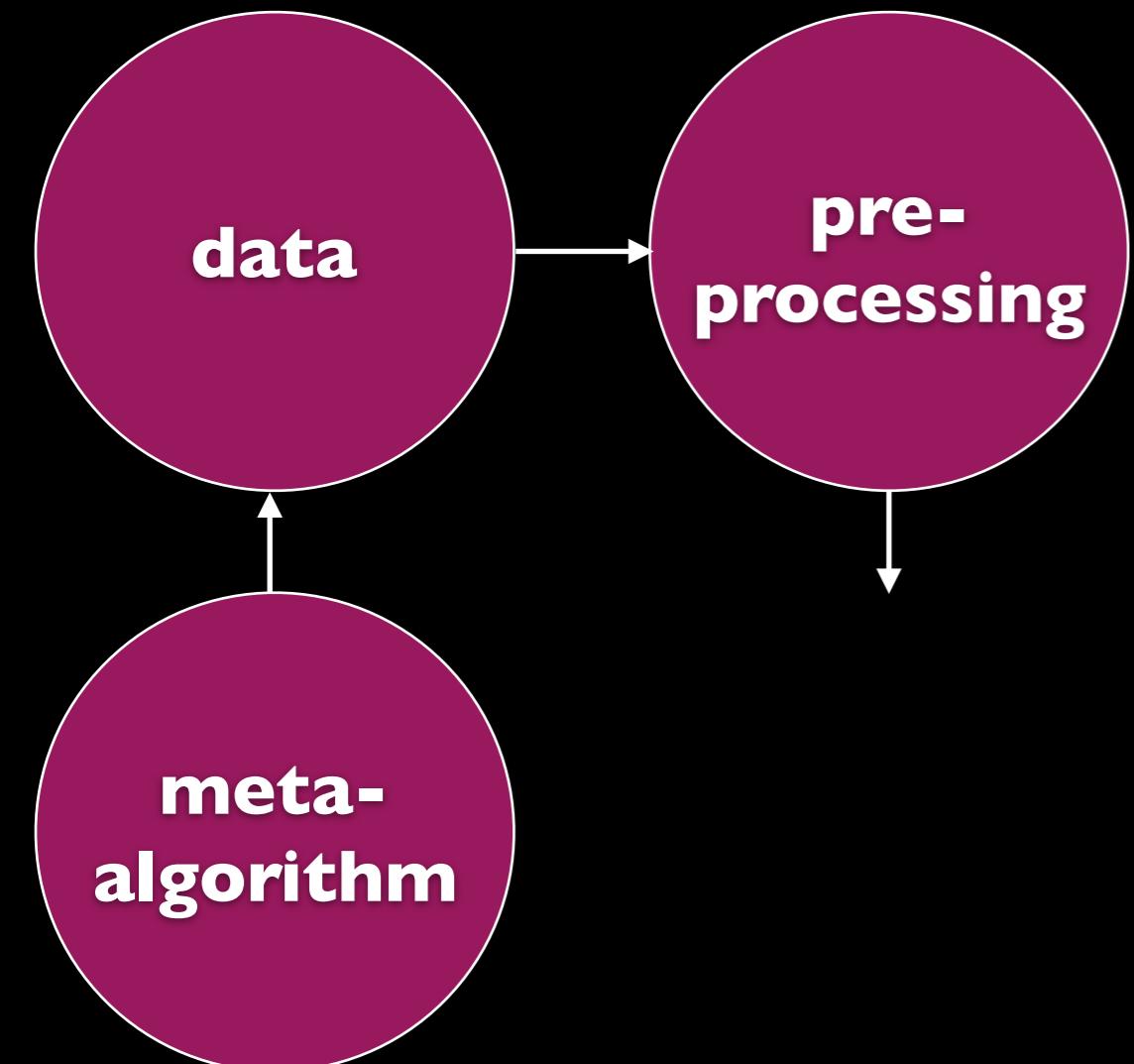
machine translation components



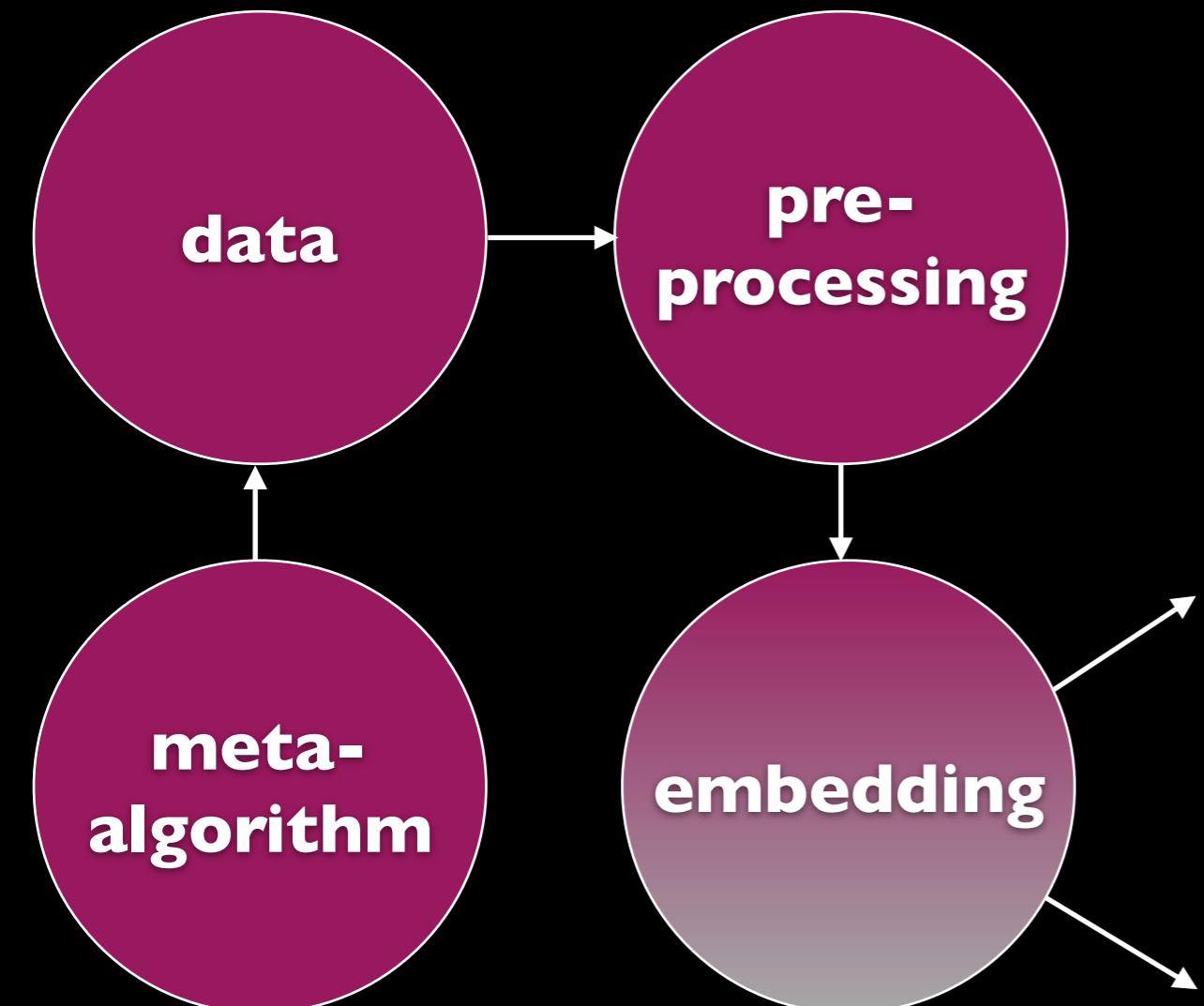
machine translation components



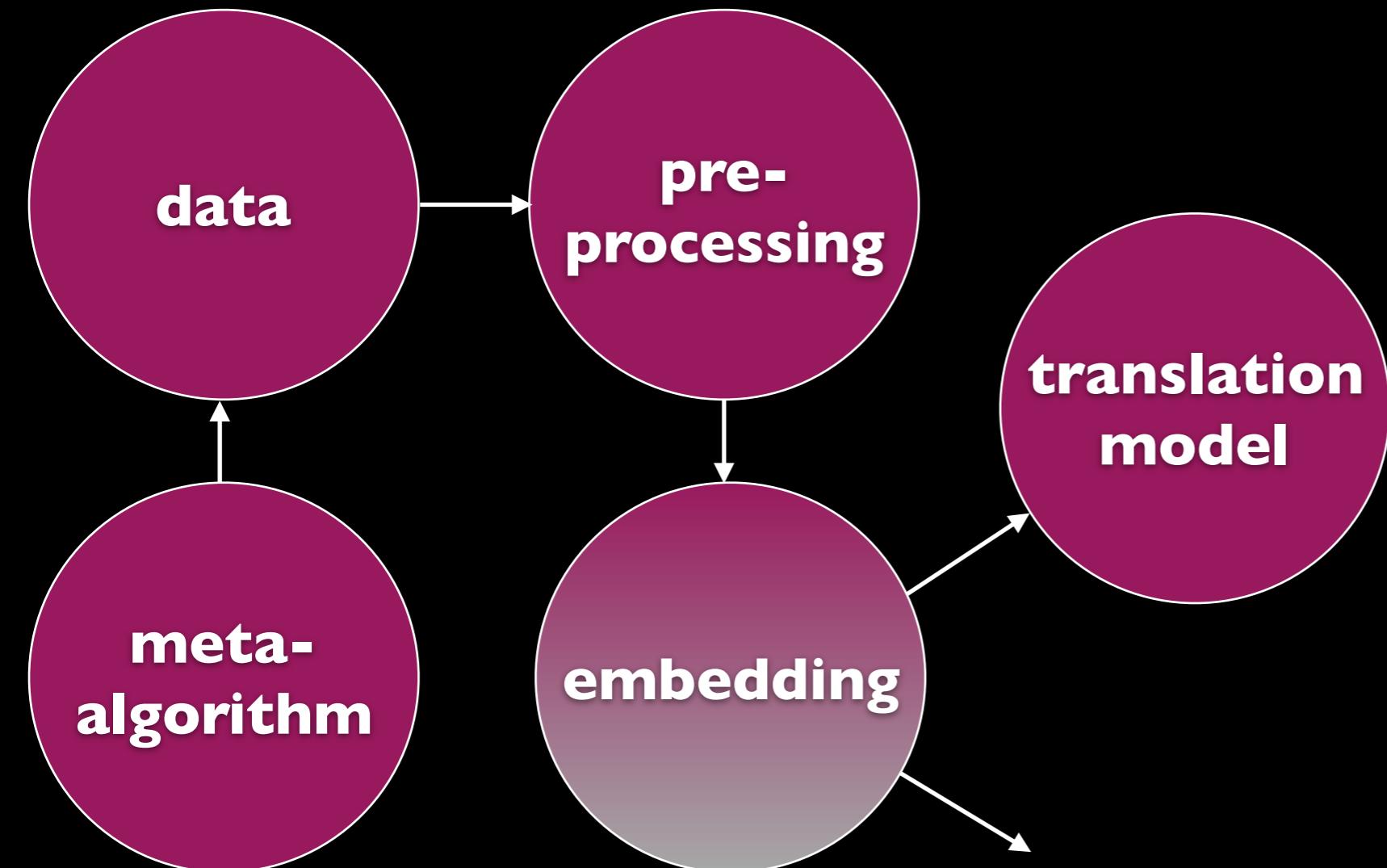
machine translation components



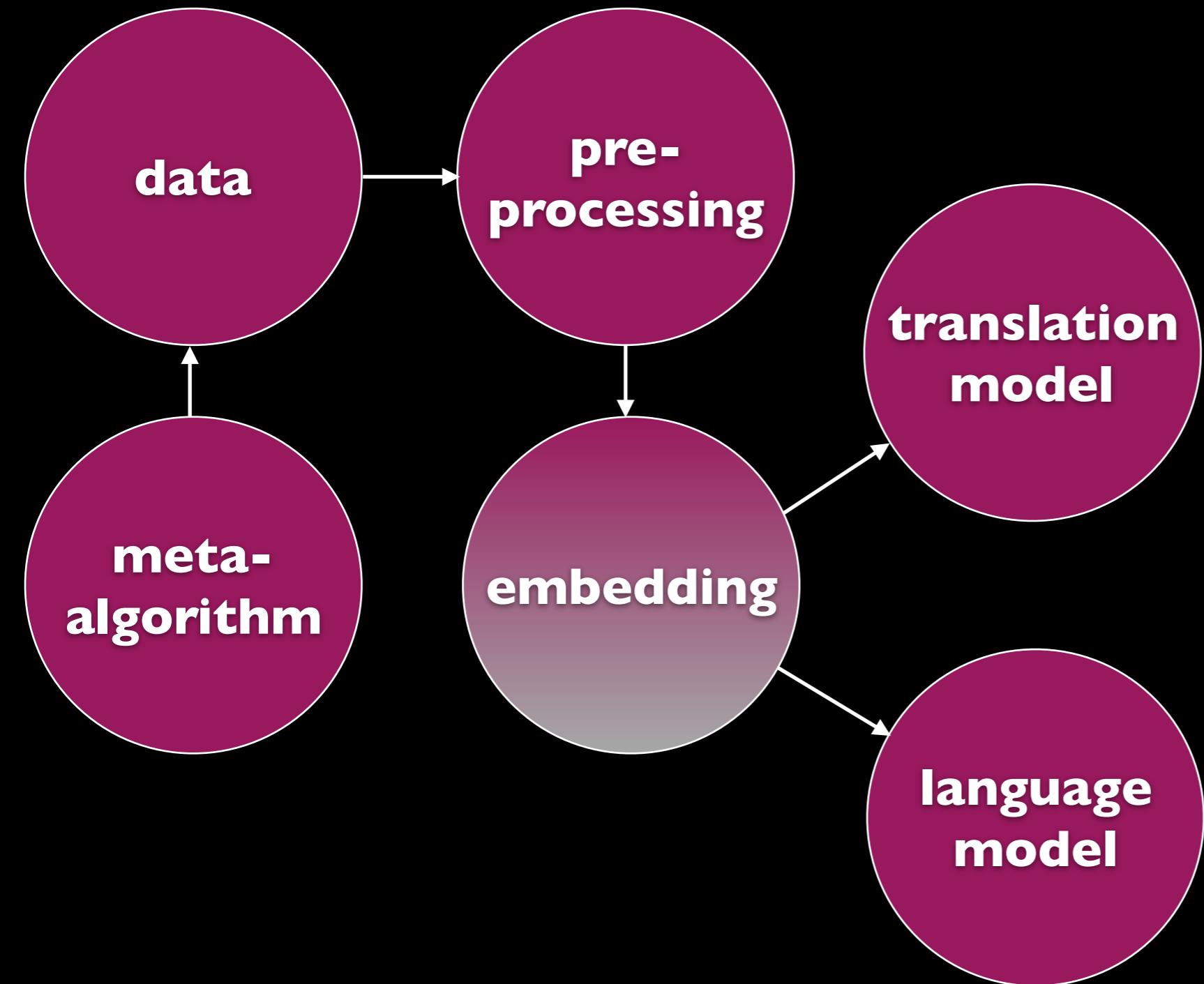
machine translation components



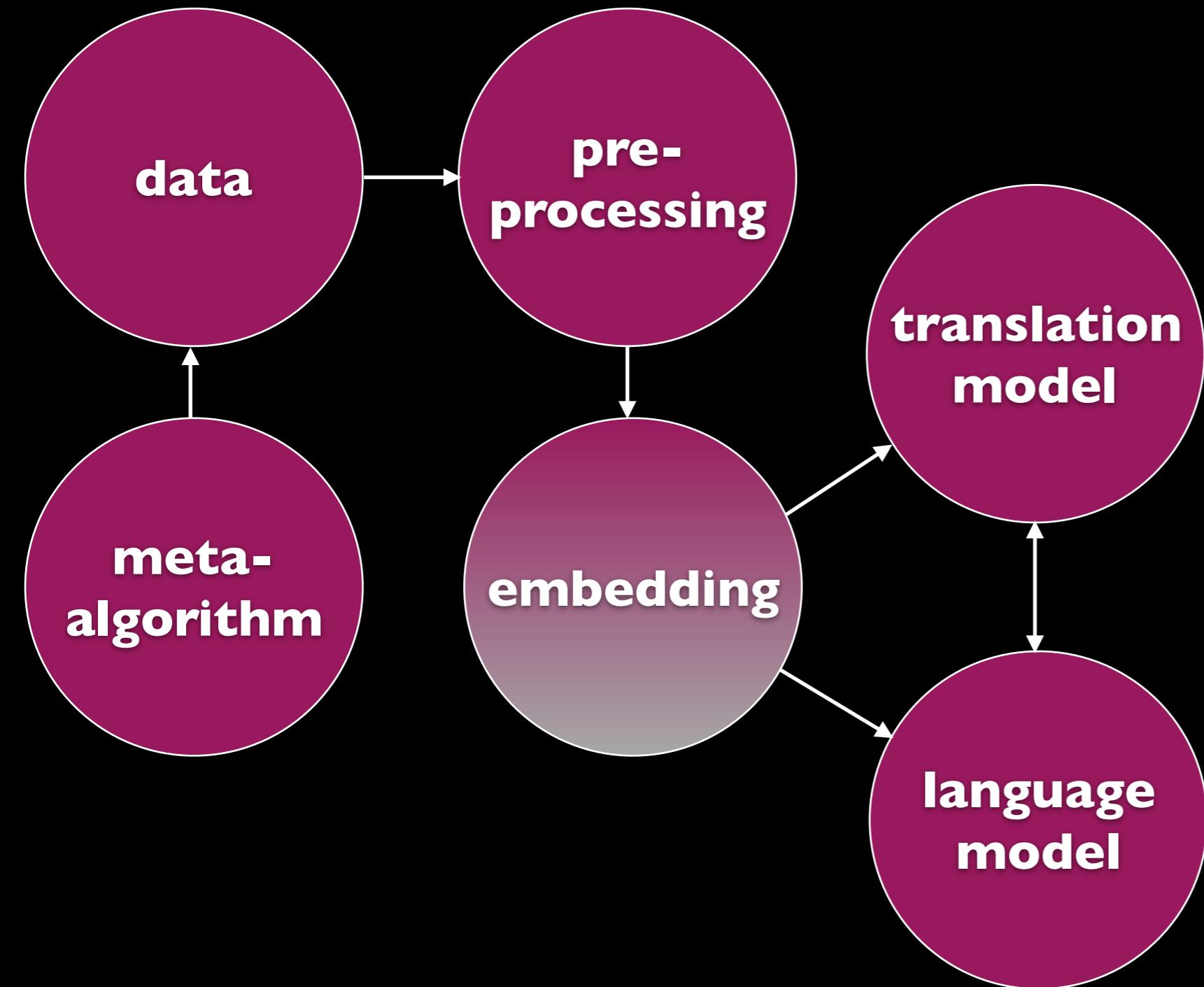
machine translation components



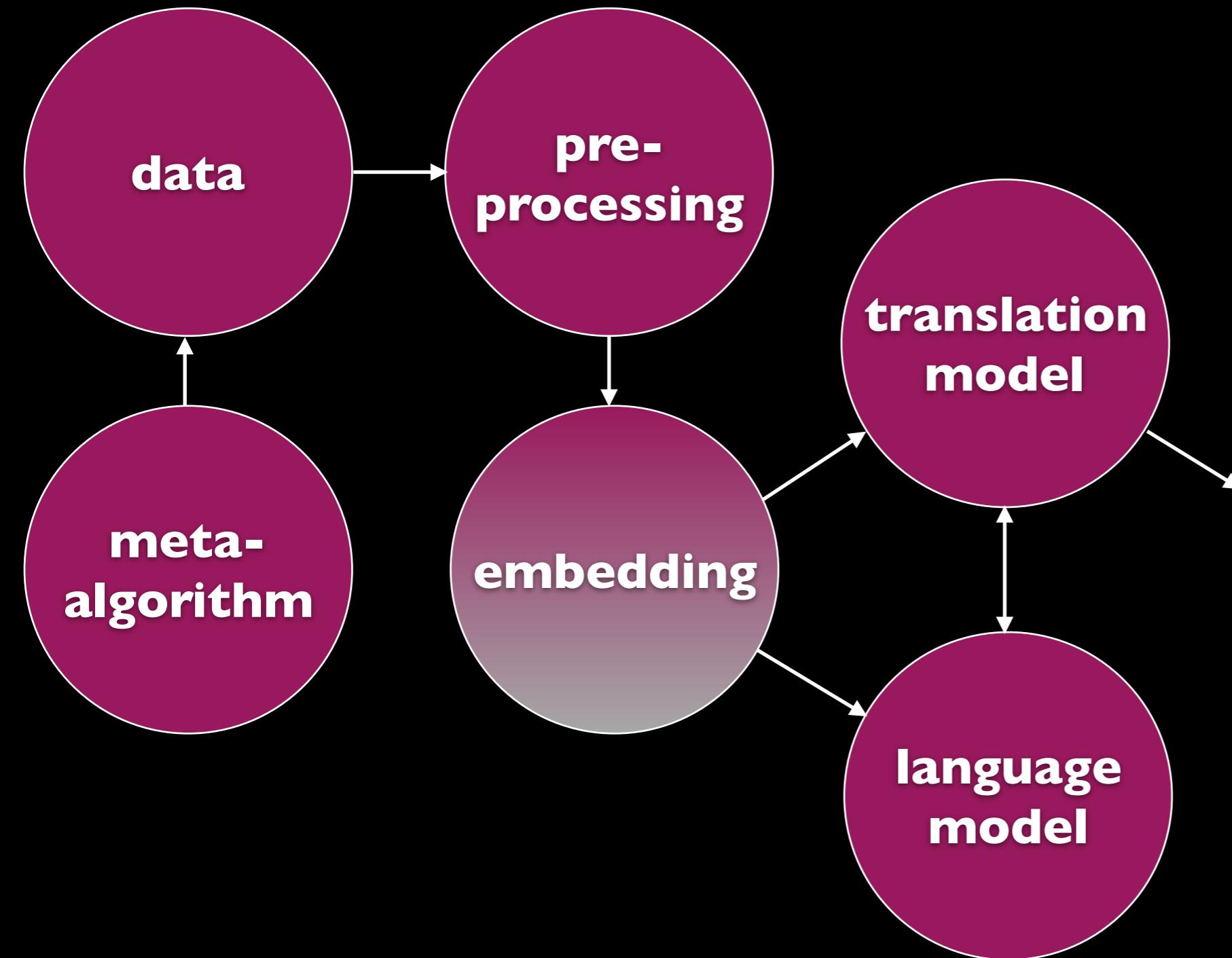
machine translation components



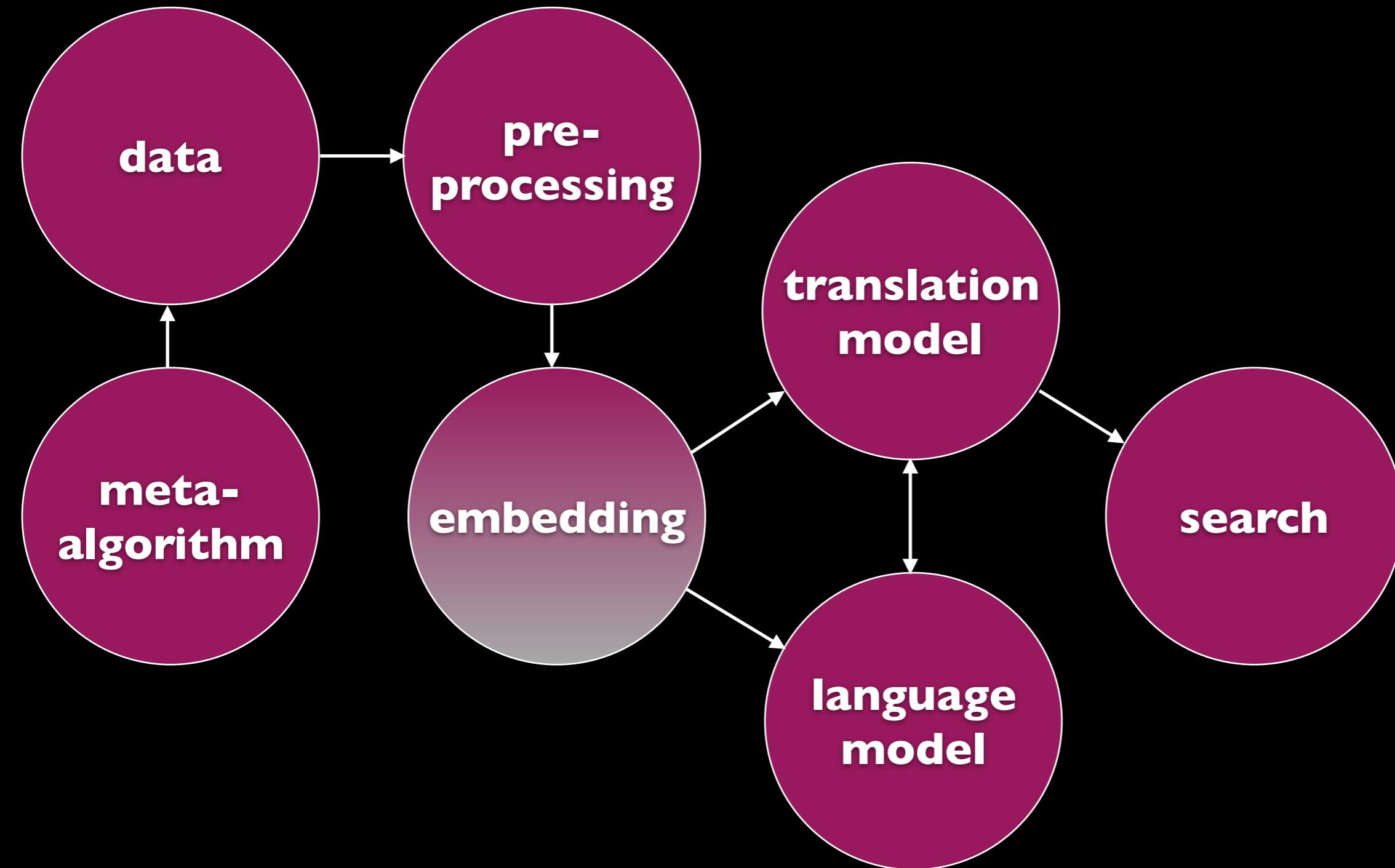
machine translation components



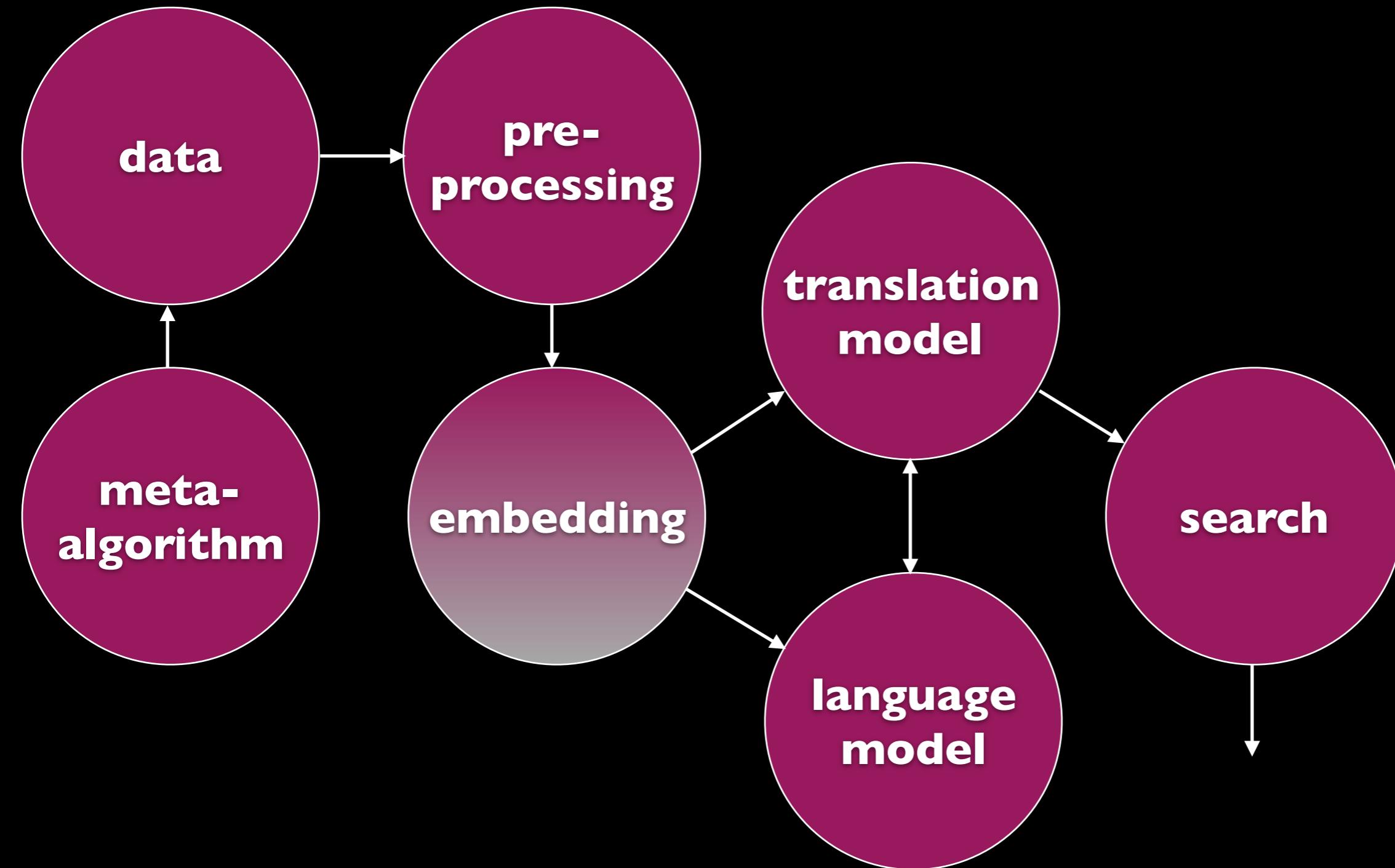
machine translation components



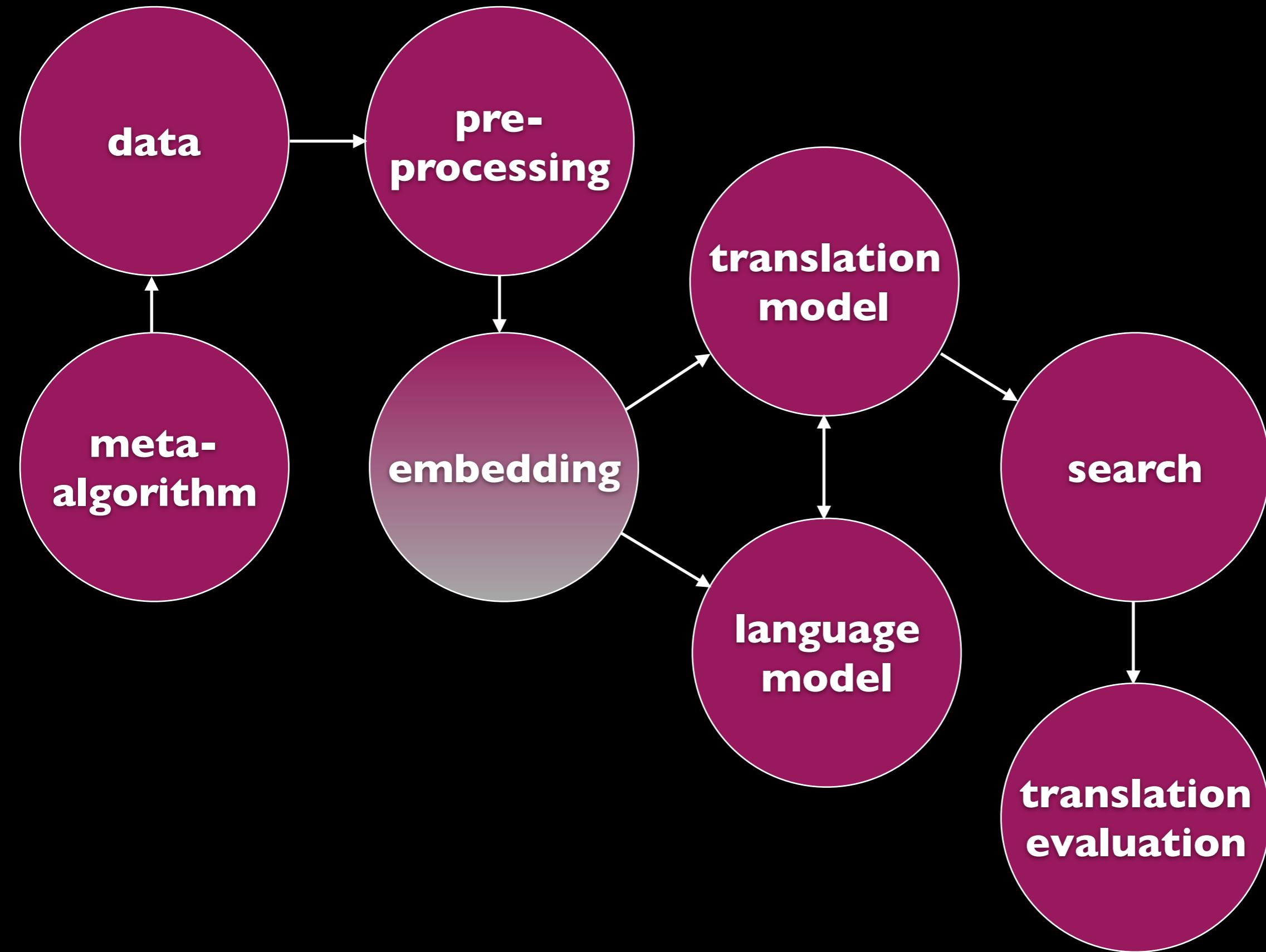
machine translation components



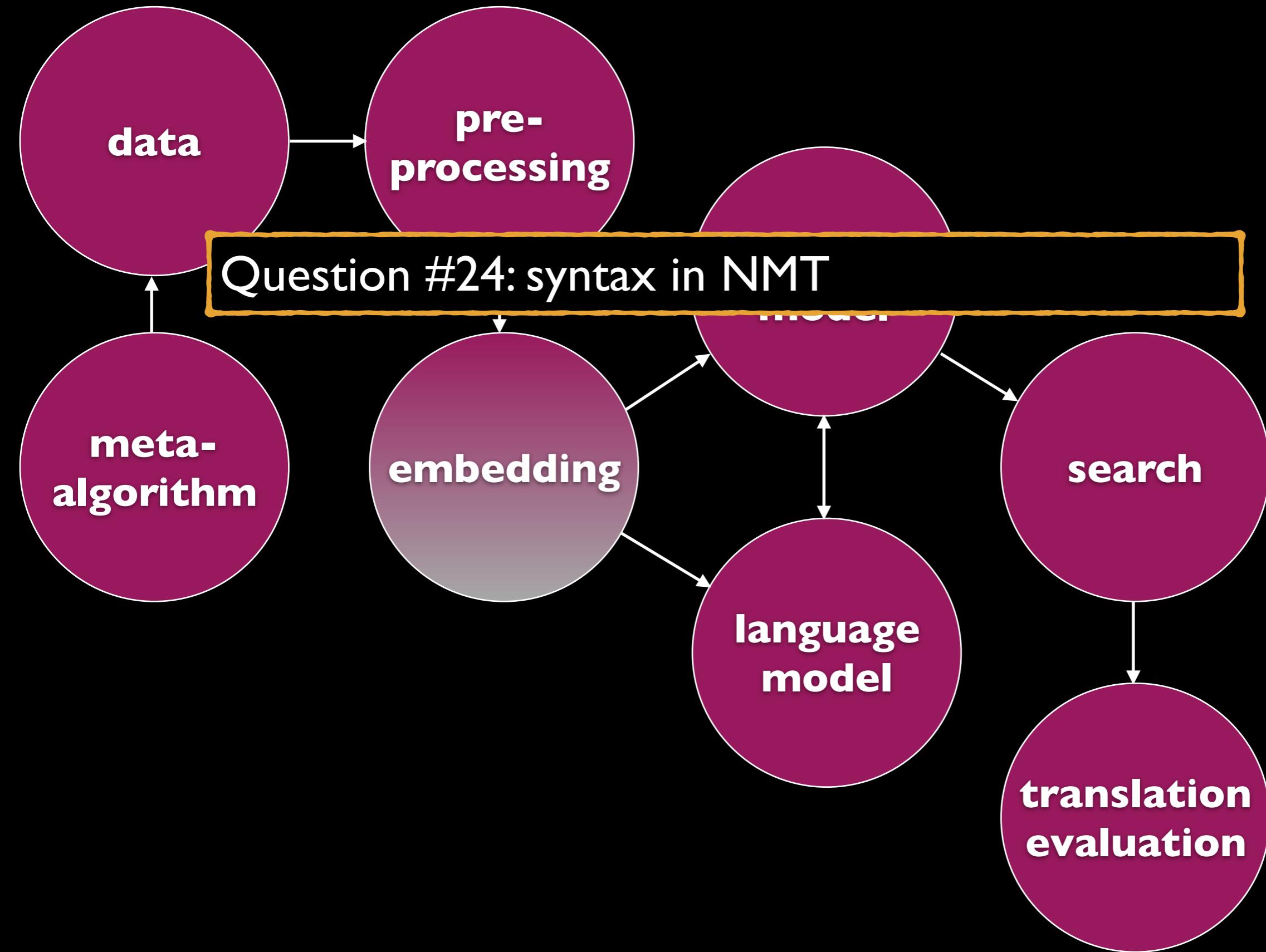
machine translation components



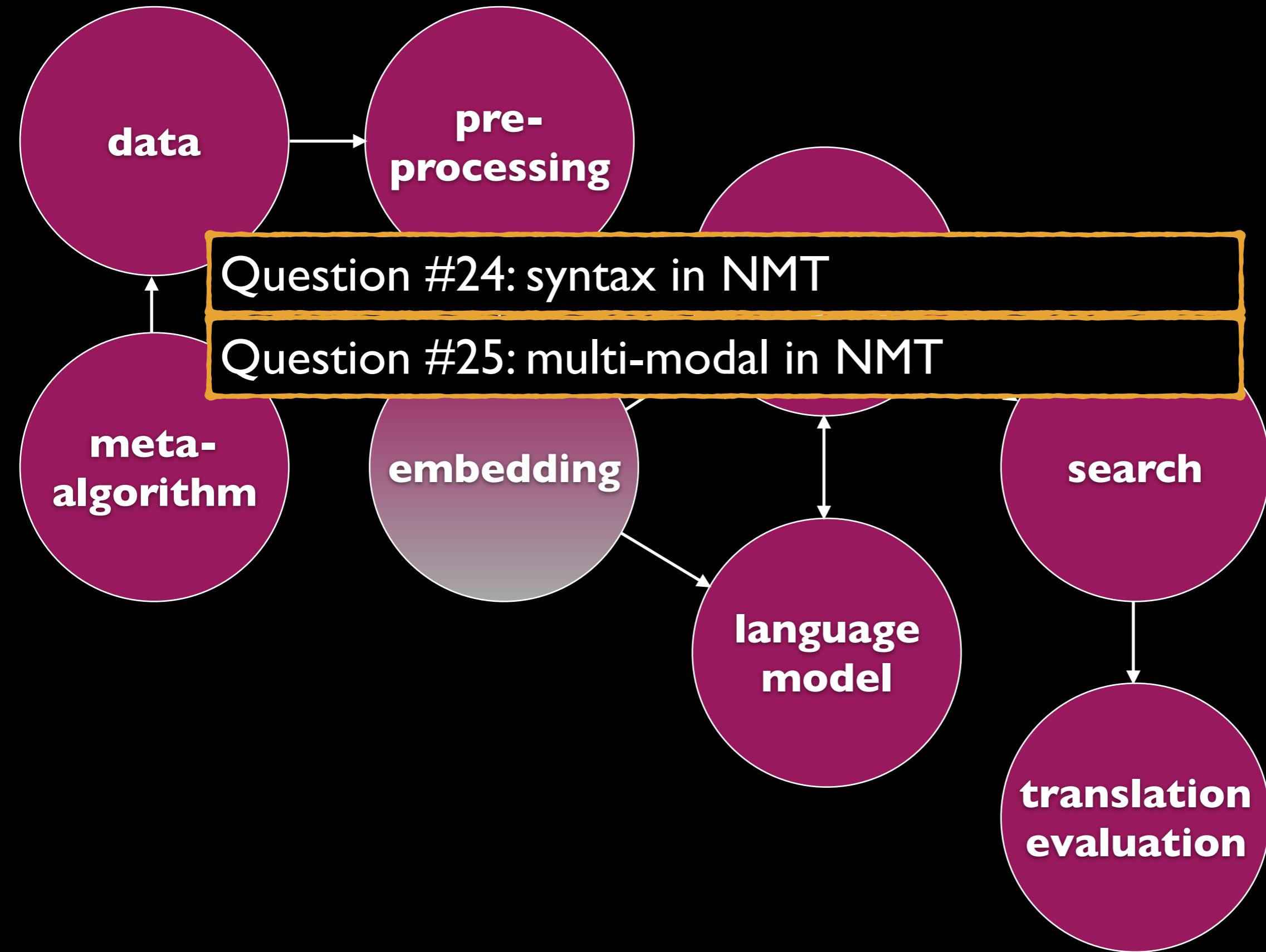
machine translation components



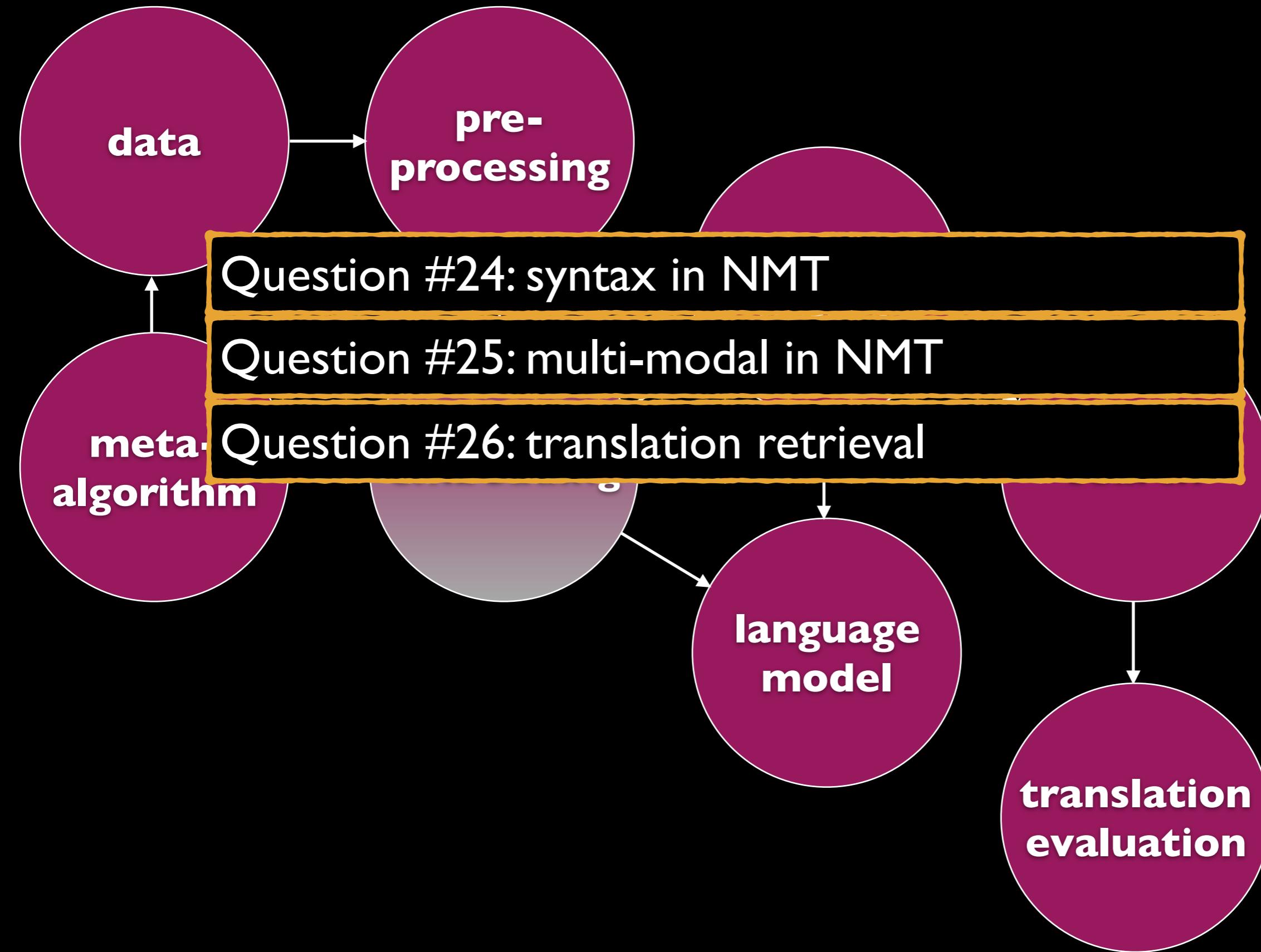
machine translation components



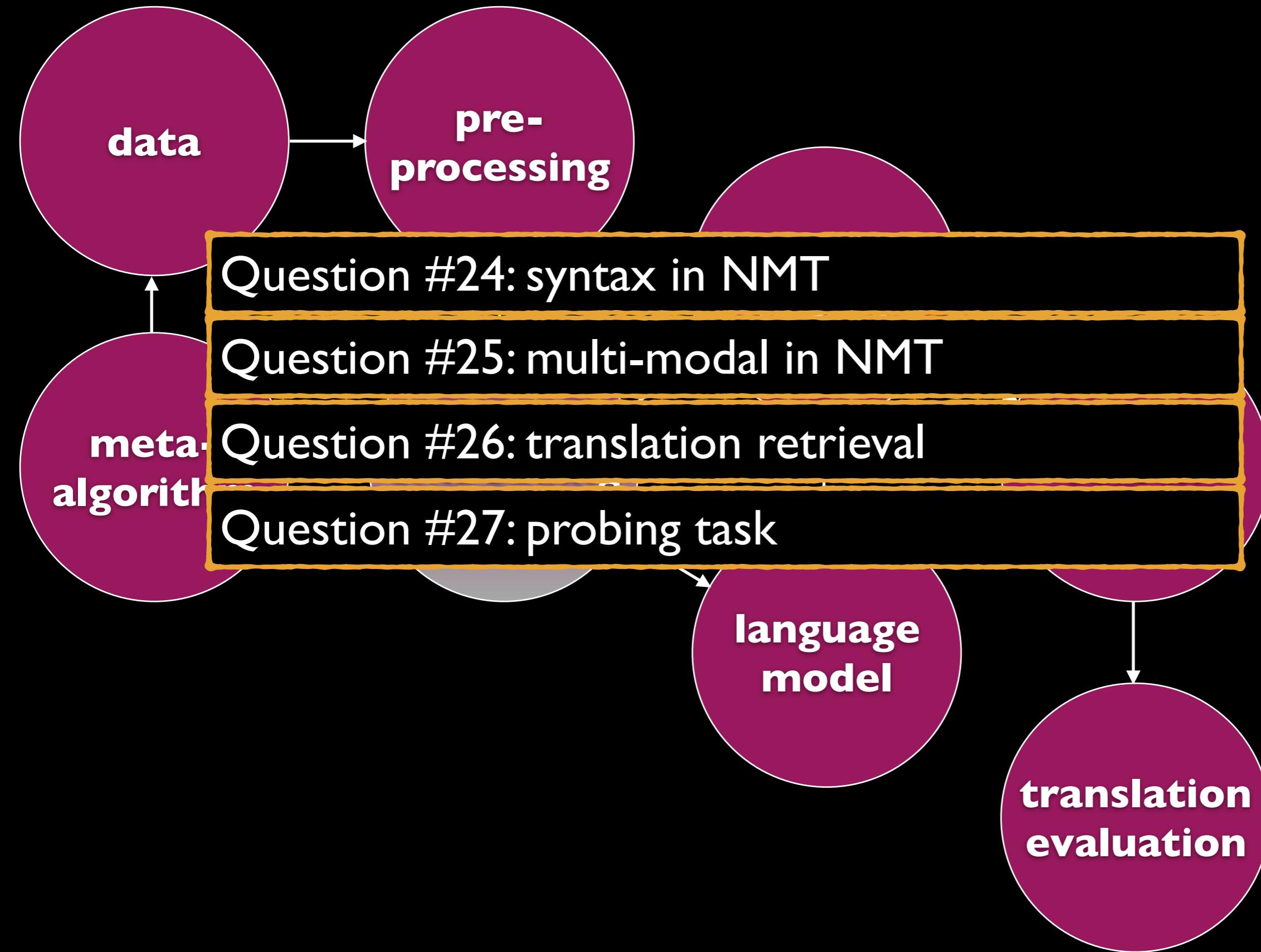
machine translation components



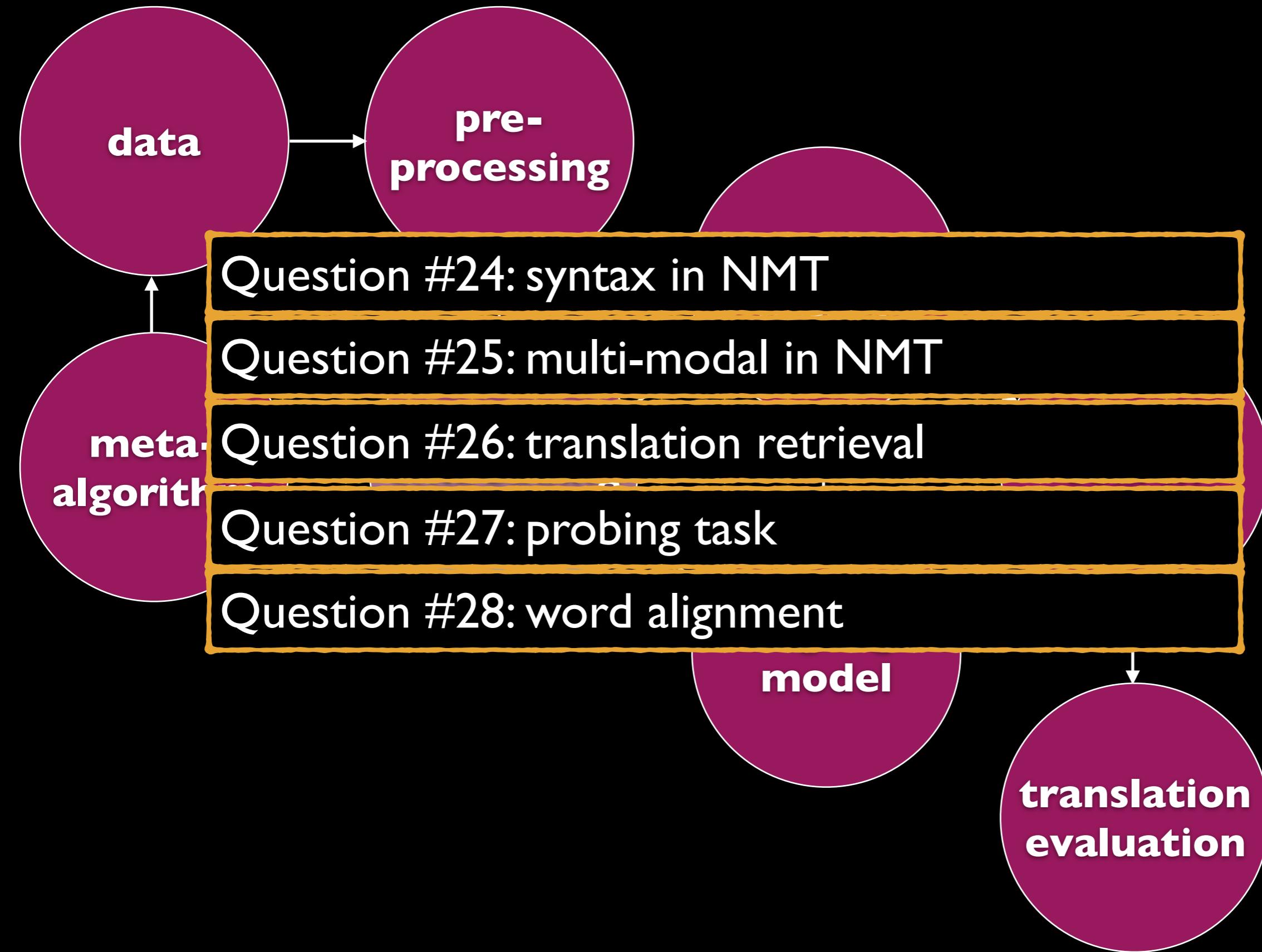
machine translation components



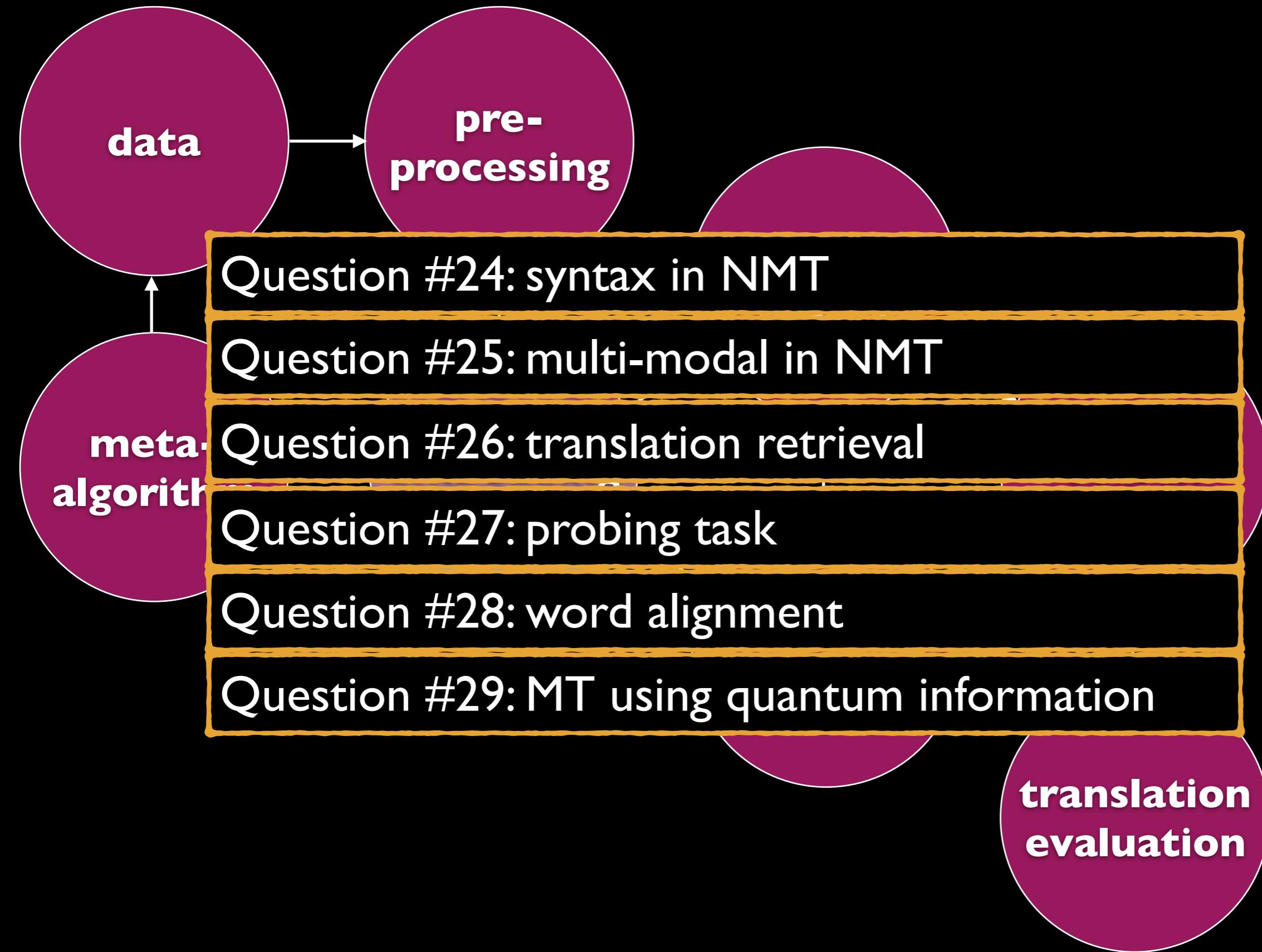
machine translation components



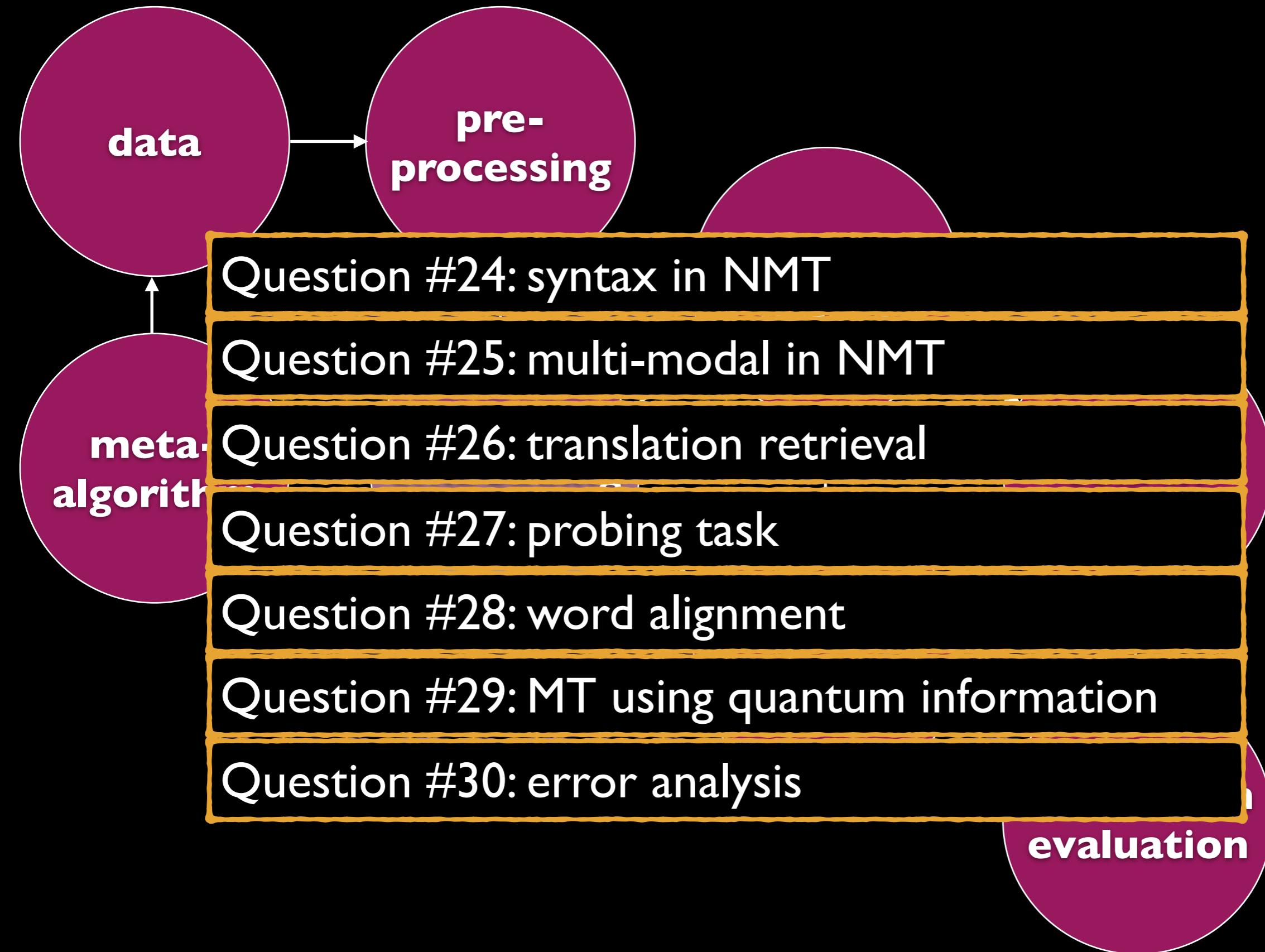
machine translation components



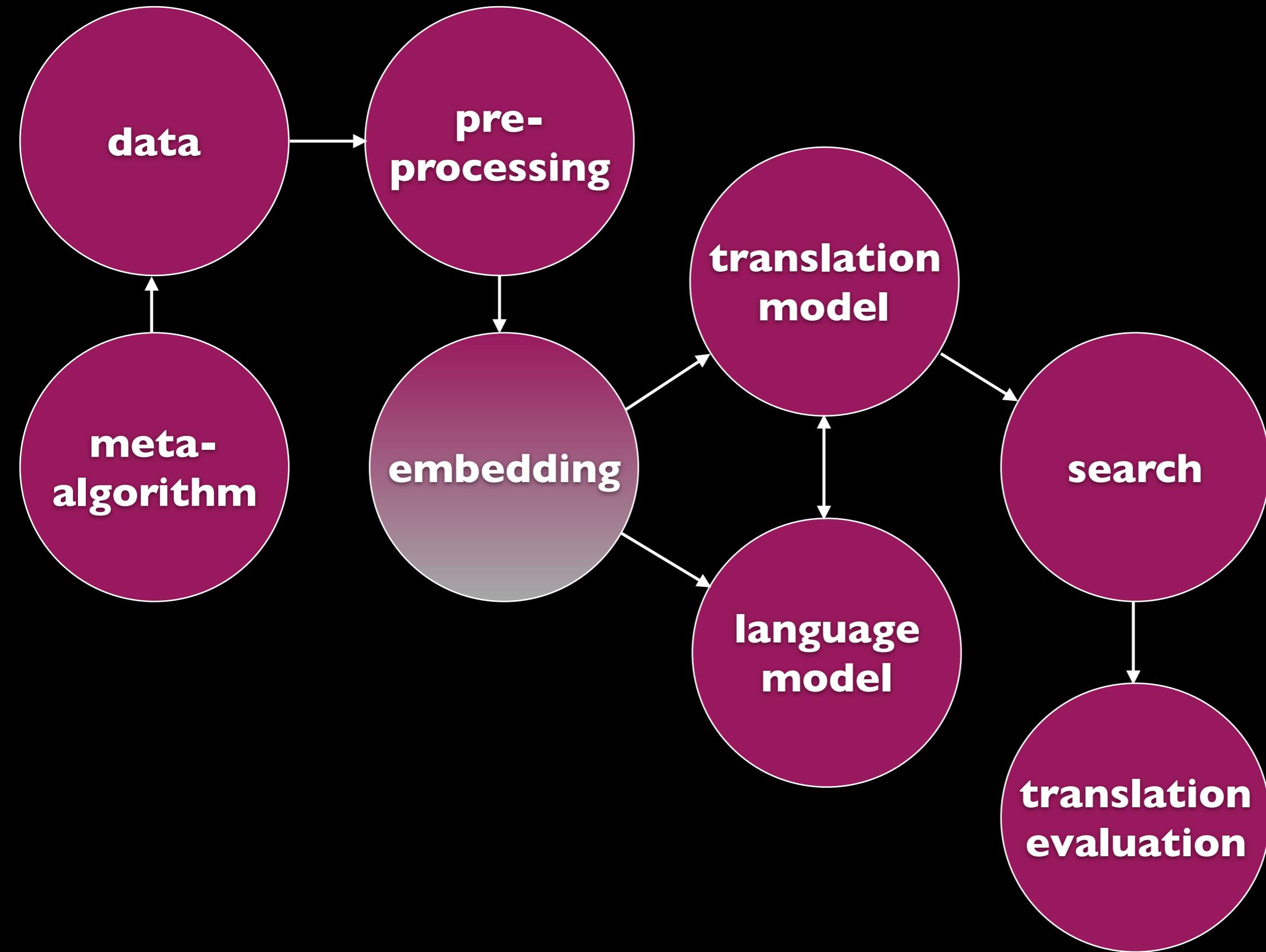
machine translation components



machine translation components



machine translation components



30 Questions

Question #1: how to enhance NMT robustness?

Question #2: how to increase interpretability?

Question #3: better text embedding/representation?

BERT, ELMO, GloVec, FastText, ...

Question #4: contextual memory in language model

Question #5: affective neuron activation function

Question #6: better training criterion?

Maximum Likelihood, squared error, MAP, cross-entropy, minimum risk, ..

Question #7: better training algorithm?

error back propagation, contrastive estimation, ...

Question #8: more efficient or controlled search? binary NMT, constraint

Question #9: higher correlation with human judgement? rich literature

Question #10: better quality estimation?

Question #11: text normalization

Question #12: better subword?

Question #13: monolingual and bilingual sentence segmentation

Question #14: domain adaptation

Question #15: what can we borrow from statistical MT?

Question #16: better subword e.g. with morphology?

Question #17: unseen words?

Question #18: named entities?

Question #19: higher quality in unsupervised MT?

Question #20: back translation

Question #21: pivot translation

Question #22: multi-lingual and zero resource

Question #23: interlingua exists? [Lu, et.al., 18]

Question #24: syntax in NMT

Question #25: multi-modal in NMT

Question #26: translation retrieval

Question #27: probing task

Question #28: word alignment

Question #29: MT using quantum information

Question #30: error analysis

questions?