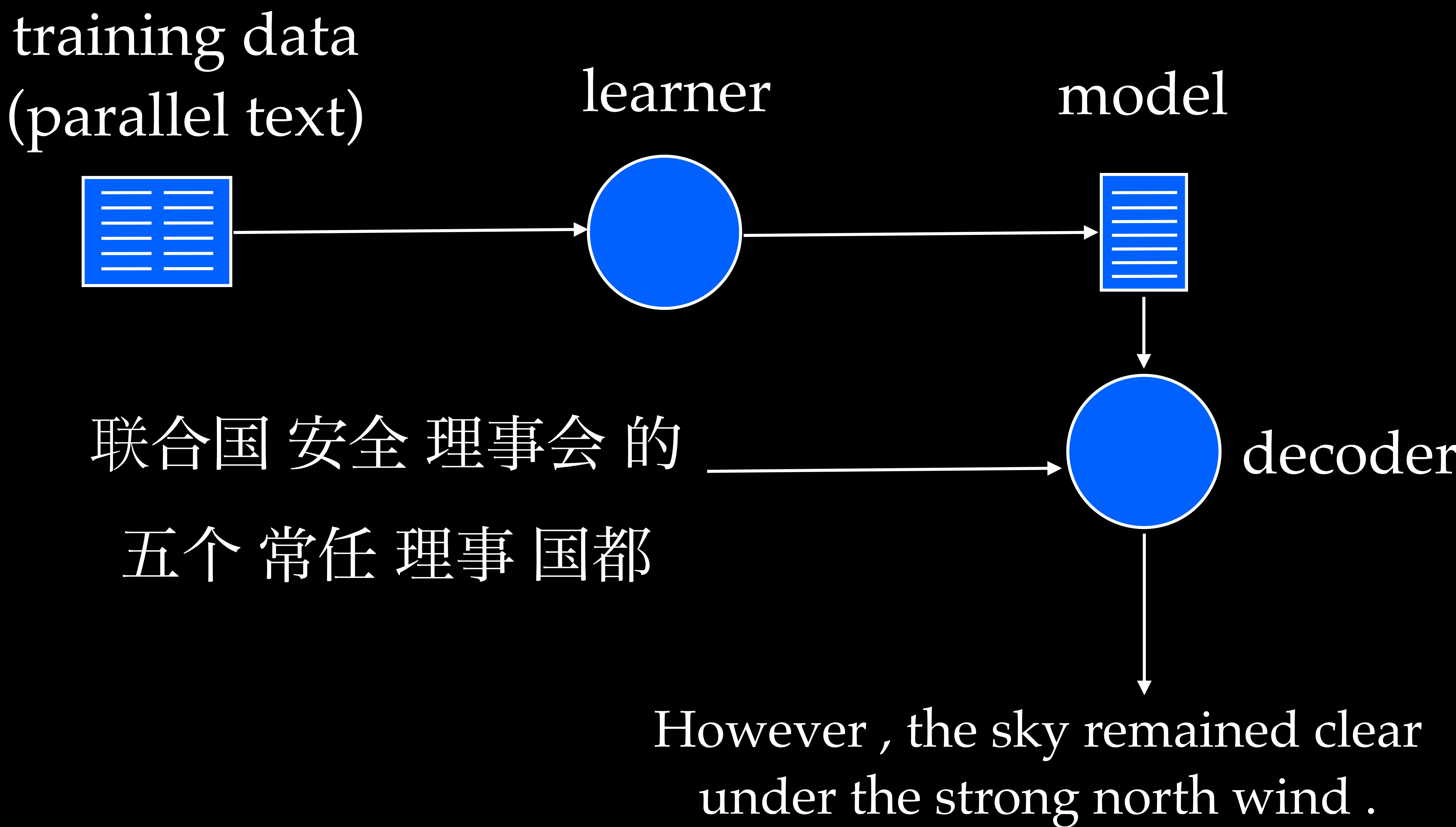
Learning (continued), Prediction, and Phrase Modeling

Adam Lopez
Johns Hopkins

Quick Recap


training data
(parallel text)
learner
model
decoder
联合国 安全 理事会 的
五个 常任 理事 国都
However , the sky remained clear
under the strong north wind .

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

$$p(English\ length|Chinese\ length)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

$$p(English\ length|Chinese\ length)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

$$p(Chinese\ word\ position)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

However

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However

$$p(English\ word|Chinese\ word)$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

However

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

However

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However ,

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

However ,

IBM Model 1

Although north wind howls , but sky still very clear .

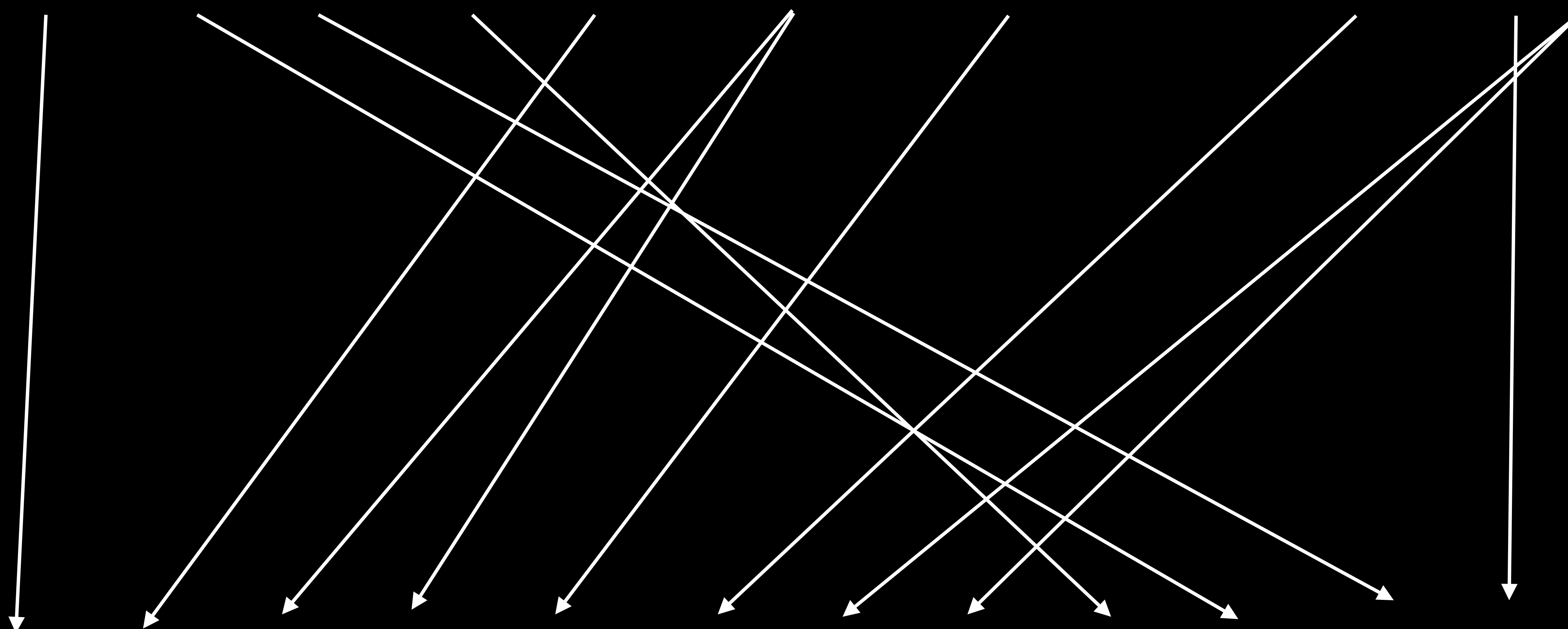
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However , the

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε



However , the sky remained clear under the strong north wind .

IBM Model 1

$$p(despite|\text{虽然})$$

$$p(however|\text{虽然})$$

$$p(although|\text{虽然})$$

$$p(northern|\text{北})$$

$$p(north|\text{北})$$

IBM Model 1

$p(despite|\text{虽然})$???

$p(however|\text{虽然})$???

$p(although|\text{虽然})$???

$p(northern|\text{北})$???

$p(north|\text{北})$???

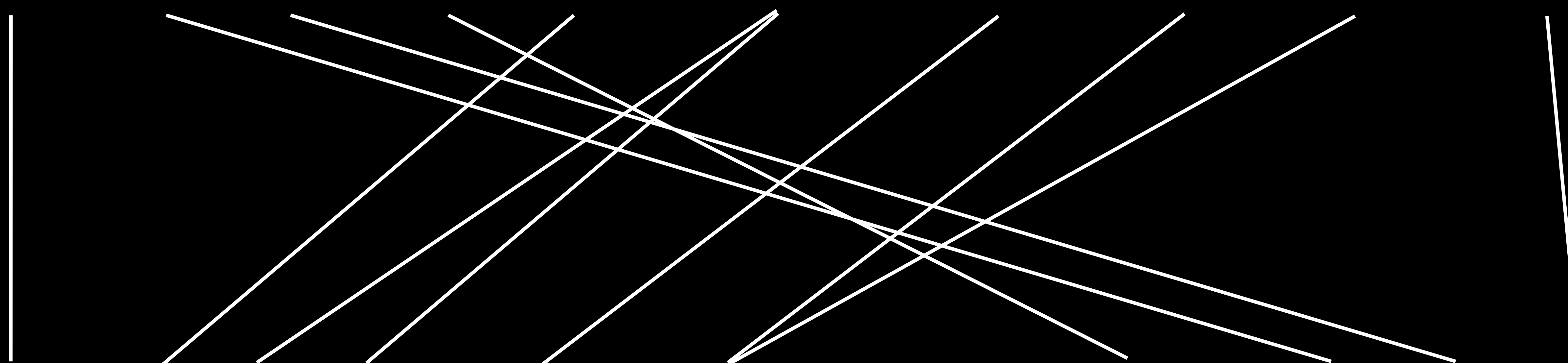
IBM Model 1

$$\theta \left\{ \begin{array}{ll} p(despite|\text{虽然}) & ??? \\ p(however|\text{虽然}) & ??? \\ p(although|\text{虽然}) & ??? \\ & \\ p(northern|\text{北}) & ??? \\ p(north|\text{北}) & ??? \end{array} \right.$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

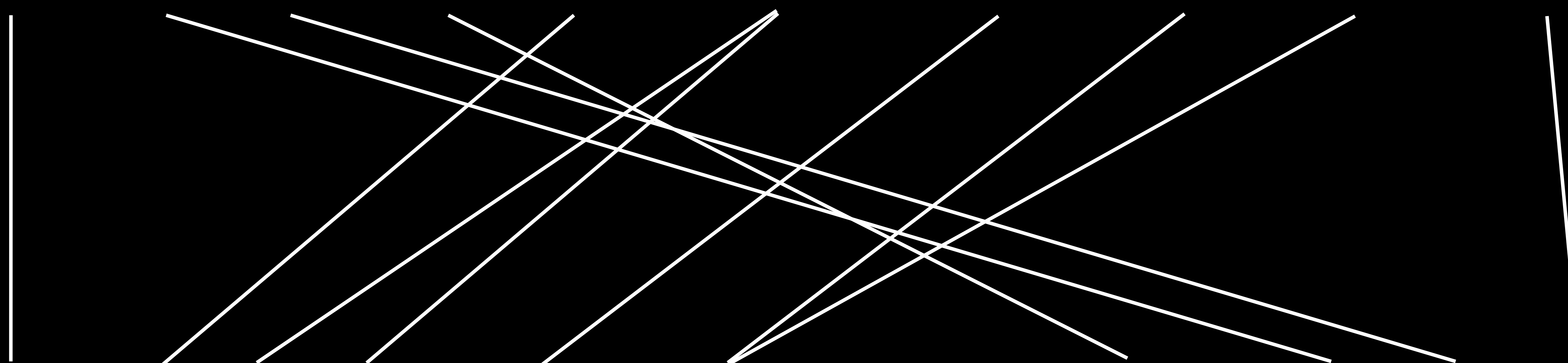
However , the sky remained clear under the strong north wind .

$$p(however|\text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

IBM Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

$$p(however|\text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

- Optimization

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)}) \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

number of sentences

alignment of French word at position *i*

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(p(I^{(n)} | J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)} | J^{(n)}) \cdot p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

French, English sentence lengths

French, English word pair

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \prod_{n=1}^{N} \left(\underbrace{p(I^{(n)}|J^{(n)}) \prod_{i=1}^{I^{(n)}} p(a_i^{(n)}|J^{(n)})}_{\text{constant!}} \cdot p(f_i^{(n)}|e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)})$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

$$\log(a) < \log(b) \iff a < b$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{count(\langle f,e \rangle)} \right)$$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$

log of product = sum of logs

MLE for IBM Model 1 (observed)

$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e)$$

$$\underbrace{- \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)}_{\text{Lagrange multiplier expresses normalization constraint}}$$

MLE for IBM Model 1 (observed)

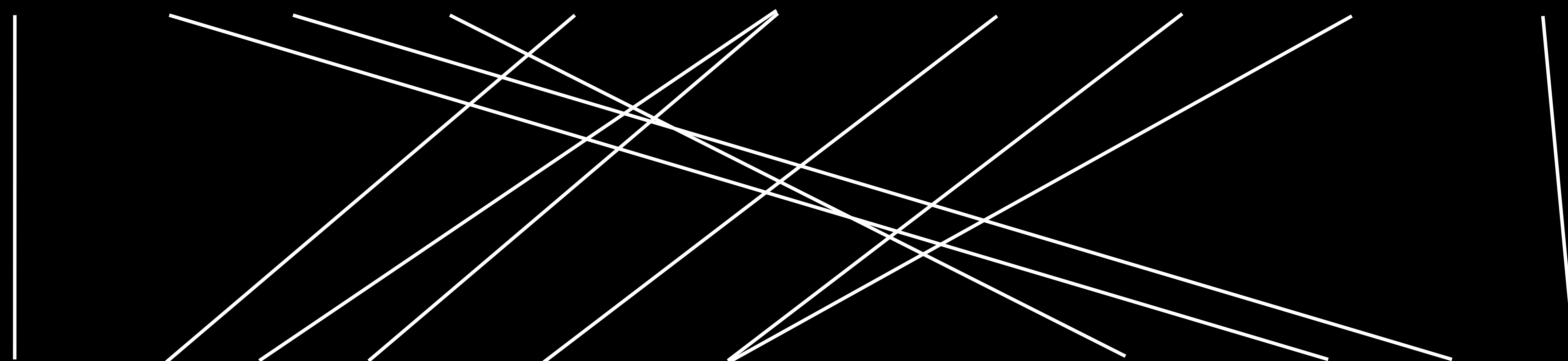
$$\Lambda(\theta, \lambda) = \log C + \sum_{f,e} count(\langle f, e \rangle) \log p(f|e) - \sum_{e} \lambda_e \left(\sum_{f} p(f|e) - 1 \right)$$

derivative $$\frac{\partial \Lambda(\theta, \lambda)}{\partial p(f|e)} = \frac{count(\langle f, e \rangle)}{p(f|e)} - \lambda_e$$

MLE for IBM Model 1 (observed)

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

MLE for IBM Model 1 (unobserved)

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$p(however|\text{虽然}) = ???$

MLE for IBM Model 1 (observed)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \prod_{n=1}^{N} \sum_{a} \prod_{i=1}^{I^{(n)}} p(f_i^{(n)} | e_{a_i}^{(n)}) \right)$$

marginalize over alignments:

$$p(f|e) = \sum_{a} p(f, a|e)$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

Not constant! Depends on parameters,
no analytic solution.

MLE for IBM Model 1 (unobserved)

$$\hat{\theta} = \arg\max_{\theta} \log \left(C \cdot \prod_{f,e} p(f|e)^{\mathbb{E}[count(\langle f,e \rangle)]} \right)$$

Not constant! Depends on parameters, no analytic solution.

But it does strongly imply an iterative solution.

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

Parameters and alignments are both unknown.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。 ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。 ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

Likelihood Estimation for Model 1

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。 ε

Parameters and alignments are both unknown.

If we knew the alignments, we could calculate the values of the parameters.

If we knew the parameters, we could calculate the likelihood of the data.

However , the sky remained clear under the strong north wind .

$p(English\ word|Chinese\ word)$ unobserved!

The Plan: Bootstrapping

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.
- Guarantee: likelihood will be monotonically nondecreasing.

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

However , the sky remained clear under the strong north wind .

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

if we had observed the alignment, this line would either be here (count 1) or it wouldn't (count 0).

However , the sky remained clear under the strong north wind .

The Plan: Bootstrapping

Although north wind howls , but sky still very clear .

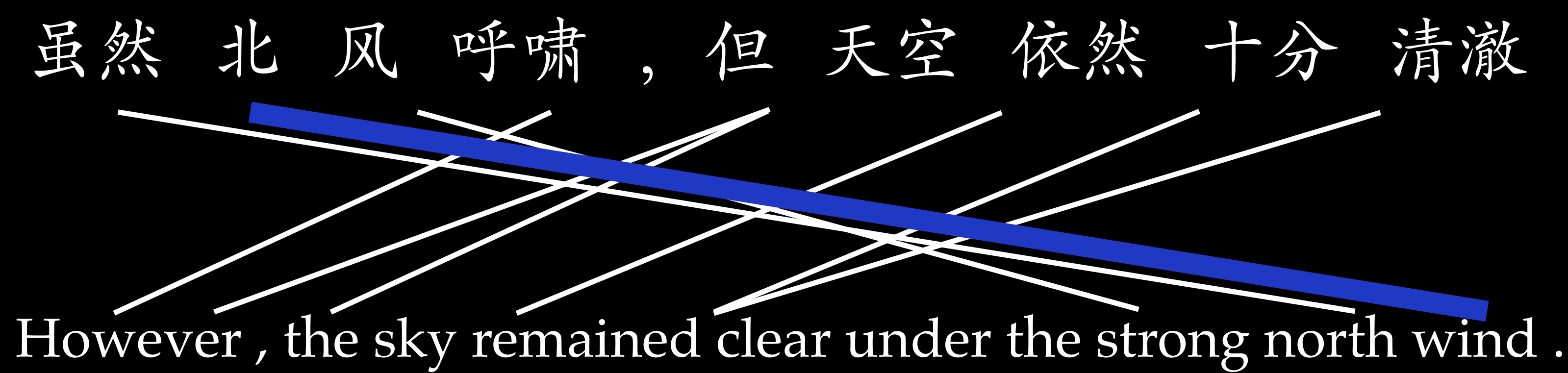
虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。ε

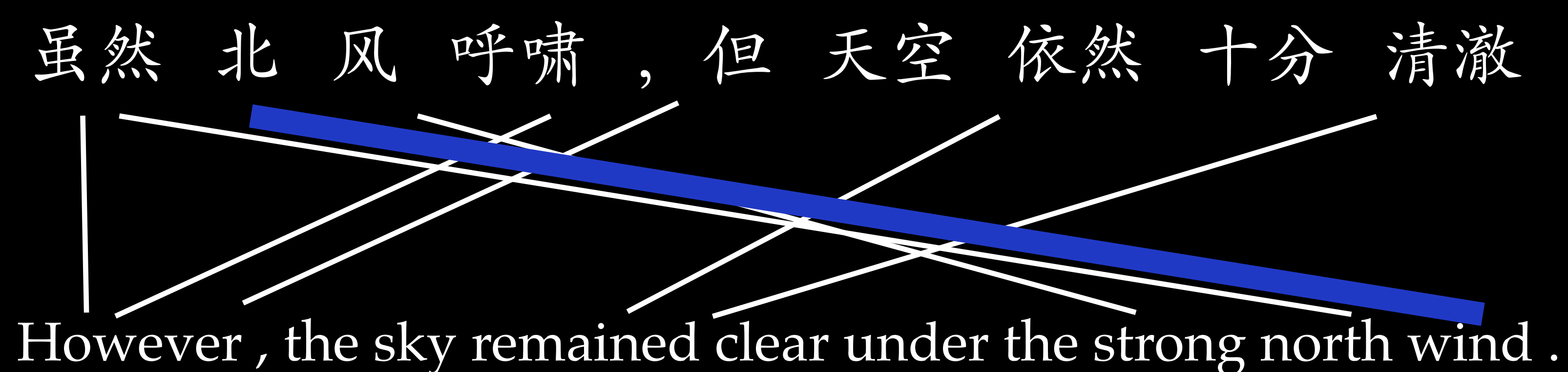
if we had observed the alignment, this line would either be here (count 1) or it wouldn't (count 0).

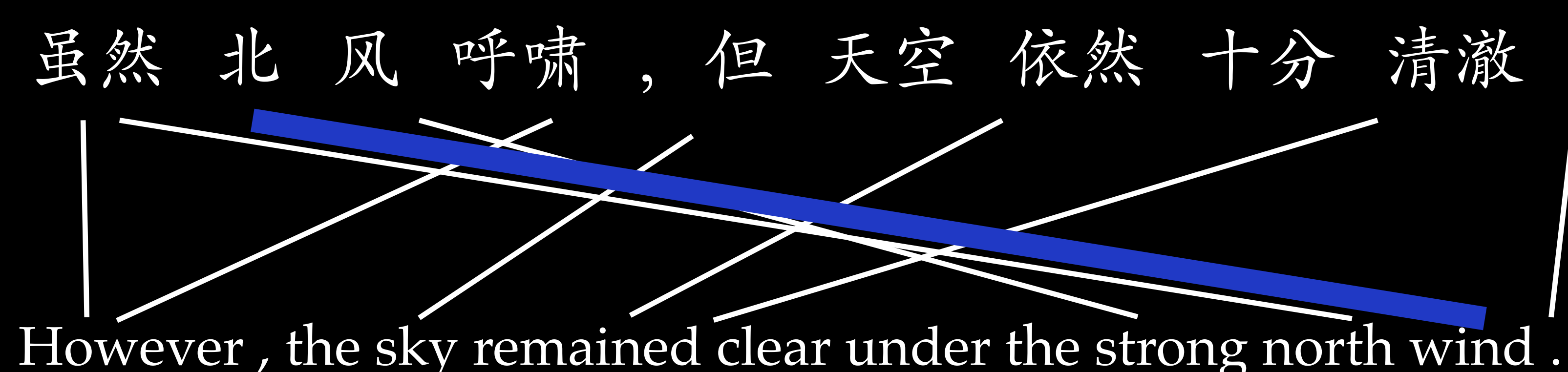
since we didn't observe the alignment, we calculate the probability that it's there.

However , the sky remained clear under the strong north wind .

Marginalize: sum all alignments containing the link

p() +

p() +

p()

Divide by sum of all *possible* alignments

p(虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .) +

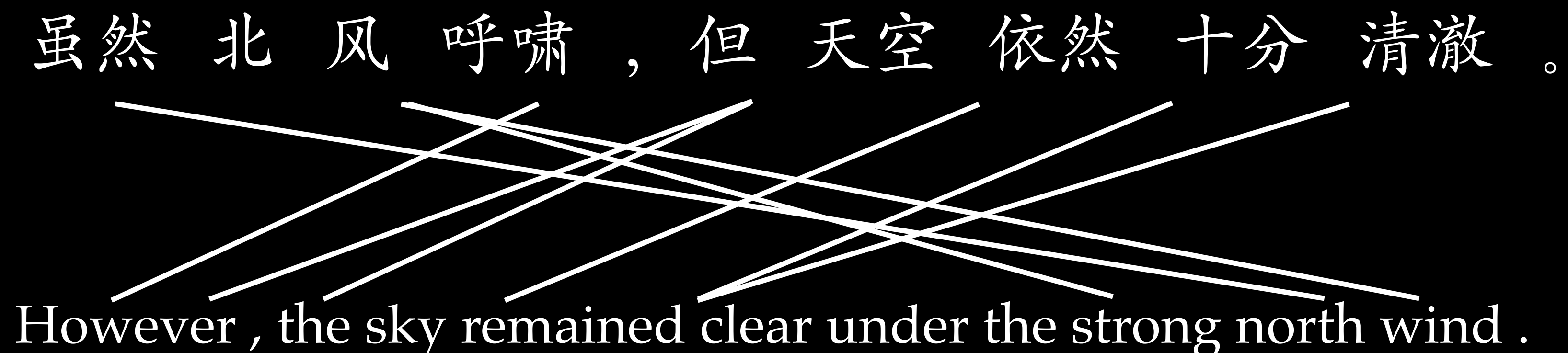
p(虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .) +

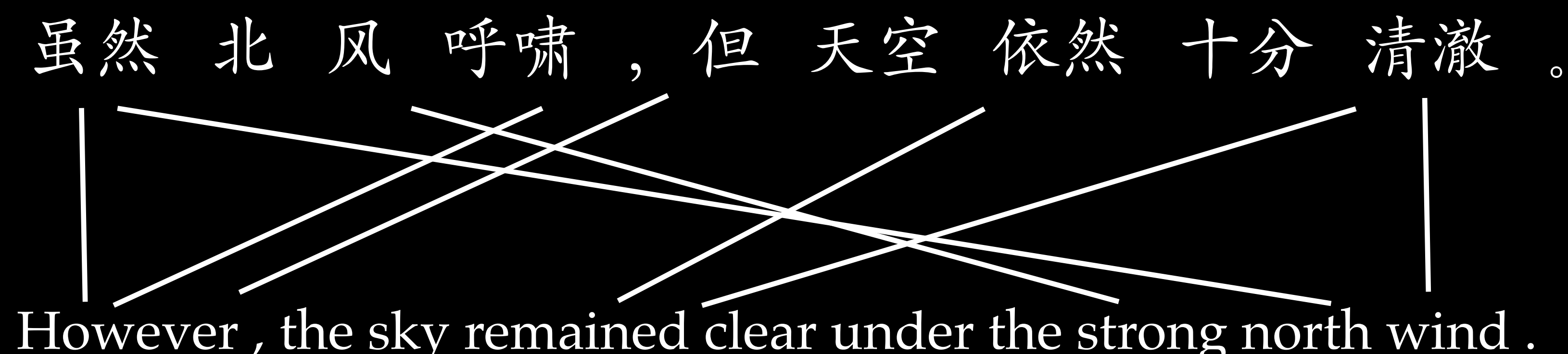
p(虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .)

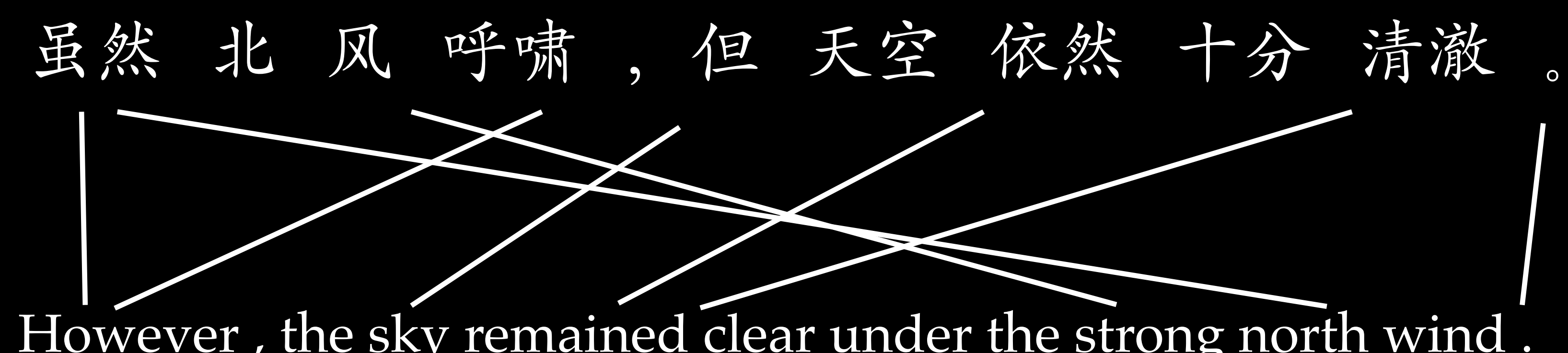
Divide by sum of all *possible* alignments

p(

) +

p(

) +

p(

)

Is this hard? How many alignments are there?

Expectation Maximization

probability of an alignment.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) p(f_i|e_j)$$

Expectation Maximization

probability of an alignment.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) p(f_i|e_j)$$

observed

uniform

Expectation Maximization

probability of an alignment.

factors across words.

$$p(F, A|E) = p(I|J) \prod_{a_i} p(a_i = j) \boxed{p(f_i|e_j)}$$

observed

uniform

Expectation Maximization

marginal probability of
alignments containing link

$$\sum_{a \in A : 北 \leftrightarrow north} p(north | 北) \cdot p(rest\ of\ a)$$

Expectation Maximization

marginal probability of
alignments containing link

$$p(north|北) \sum_{a \in A: 北 \leftrightarrow north} p(rest\ of\ a)$$

Expectation Maximization

marginal probability of
alignments containing link

$$\frac{p(north|北) \displaystyle\sum_{a \in A: 北 \leftrightarrow north} p(rest\ of\ a)}{\displaystyle\sum_{c \in Chinese\ words} p(north|c) \sum_{a \in A: \mathrm{c} \leftrightarrow north} p(rest\ of\ a)}$$

marginal probability of all
alignments

Expectation Maximization

marginal probability of
alignments containing link

$$\frac{p(north|\text{北}) \displaystyle\sum_{a \in A:\text{北} \leftrightarrow north} p(rest\ of\ a)}{\displaystyle\sum_{c \in Chinese\ words} p(north|c) \sum_{a \in A:\, c \leftrightarrow north} p(rest\ of\ a)}$$

marginal probability of all
alignments

Expectation Maximization

marginal probability of
alignments containing link

$$\frac{p(north|北)\sum_{a\in A:北\leftrightarrow north} p(rest\ of\ a)}{\sum_{c\in Chinese\ words} p(north|c)\sum_{a\in A:c\leftrightarrow north} p(rest\ of\ a)}$$

identical!

marginal probability of all
alignments

Expectation Maximization

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Expectation Maximization

marginal probability (expected count) of an alignment containing the link

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Expectation Maximization

marginal probability (expected count) of an alignment containing the link

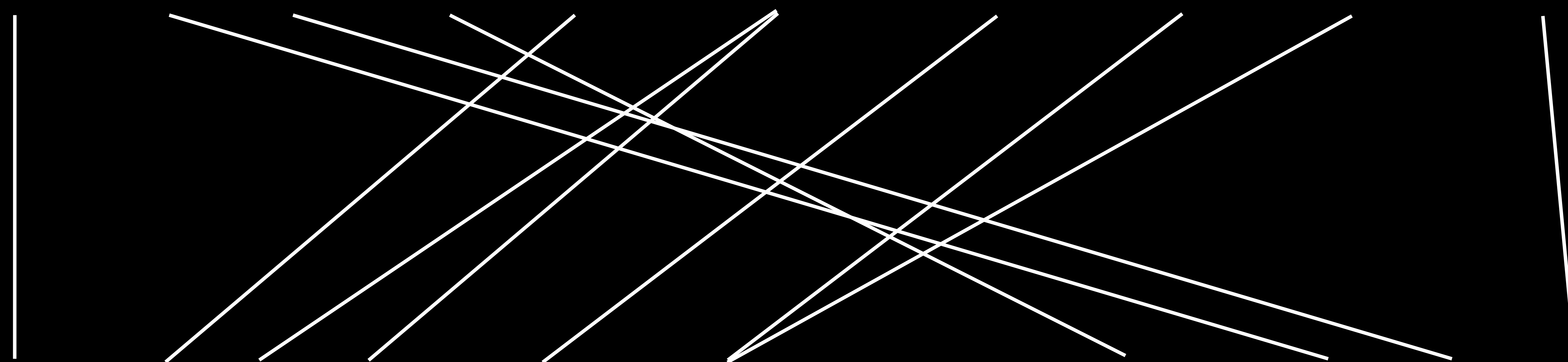
$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

For each sentence, use this quantity instead of 0 or 1

Translation Models

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。



However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

Translation Models

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ， 但 天空 依然 十分 清澈 。

However , the sky remained clear under the strong north wind .

$$p(however \mid \text{虽然}) = \frac{\textit{Expected}\ \text{\# of times 虽然 aligns to However}}{\text{\# of times 虽然 occurs}}$$

Expectation Maximization

Why does this even work?

$$\frac{p(north|\text{北})}{\sum_{c \in Chinese\ words} p(north|c)}$$

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

MLE: choose parameters that maximize this expression.

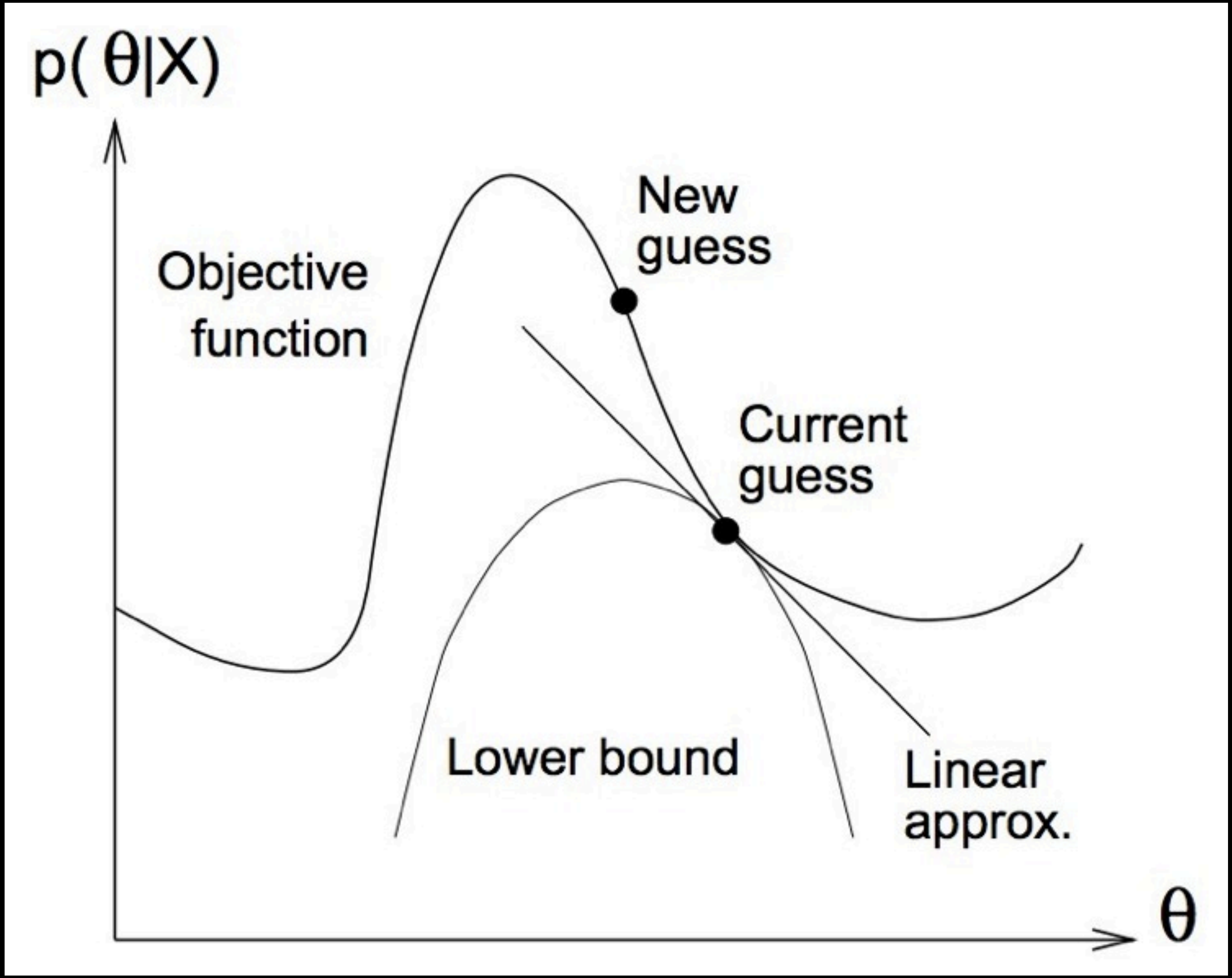
Expectation Maximization

Observation 1: We are still solving a maximum likelihood estimation problem.

$$p(Chinese|English) = \sum_{alignments} p(Chinese, alignment|English)$$

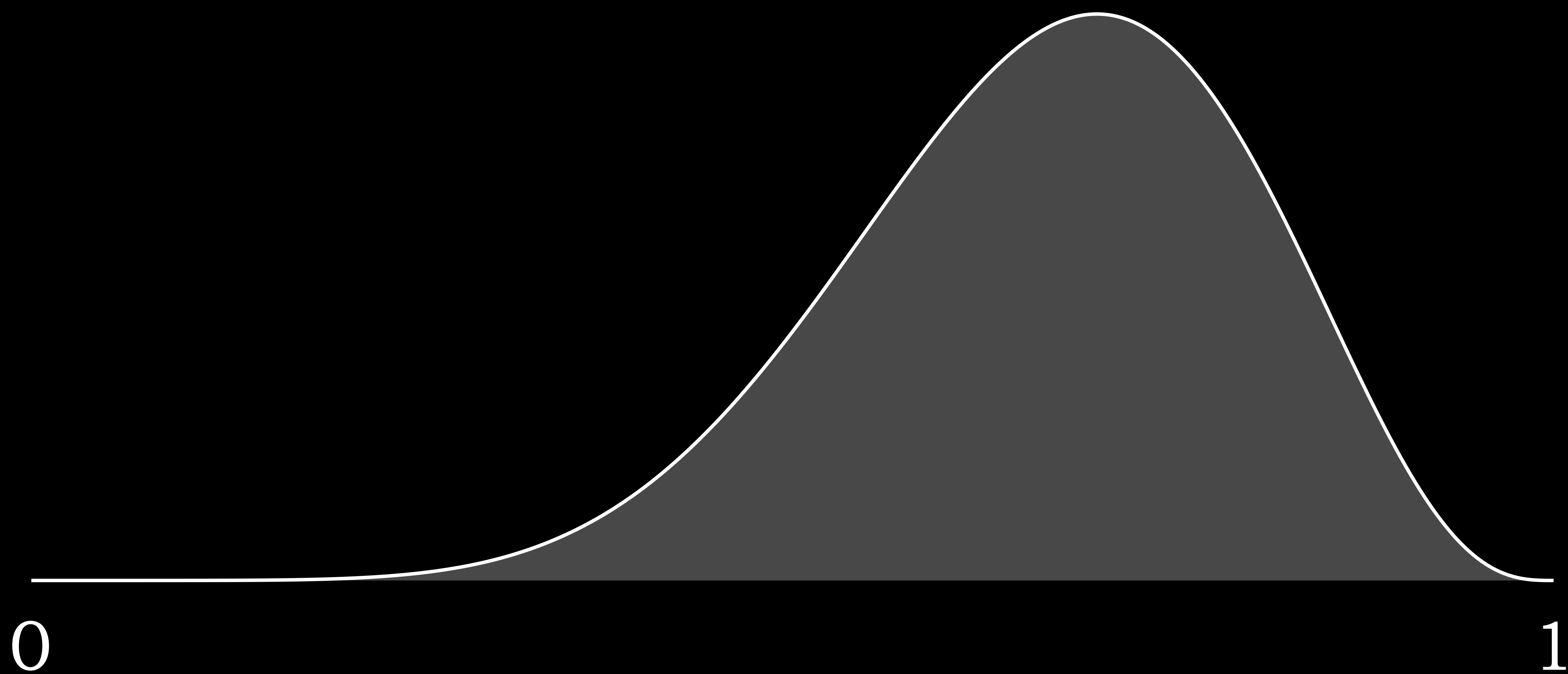
MLE: choose parameters that maximize this expression.

Minor problem: there is no analytic solution.



(from Minka ’98)

... and, likelihood is ***convex*** for this model:



Exercises!

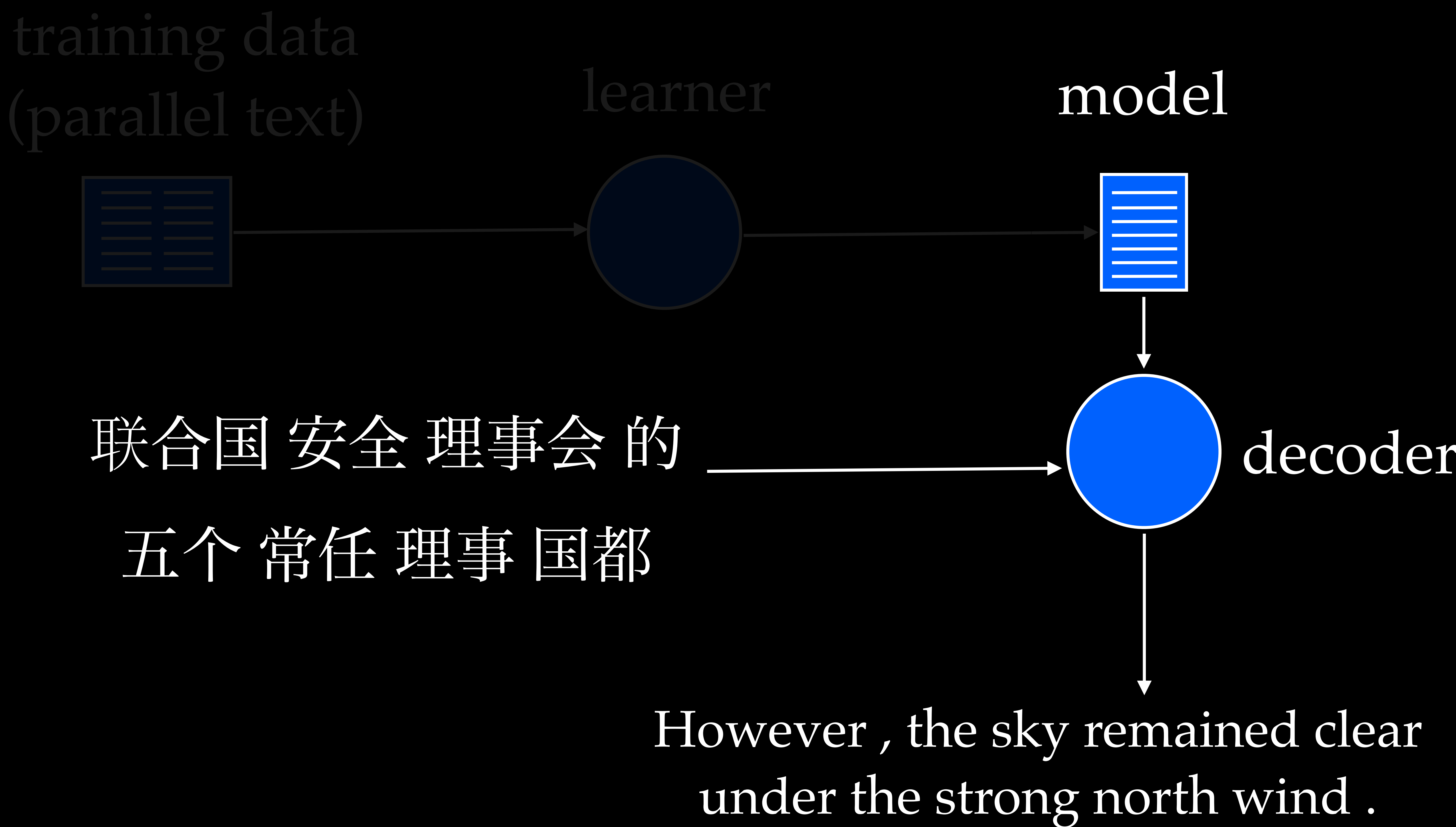
- Totally optional.
- Most effective way to understand concepts is to apply them!
- I'm happy to answer questions.

http://www.cs.jhu.edu/~alopez/nasslli2012/exercise1.html

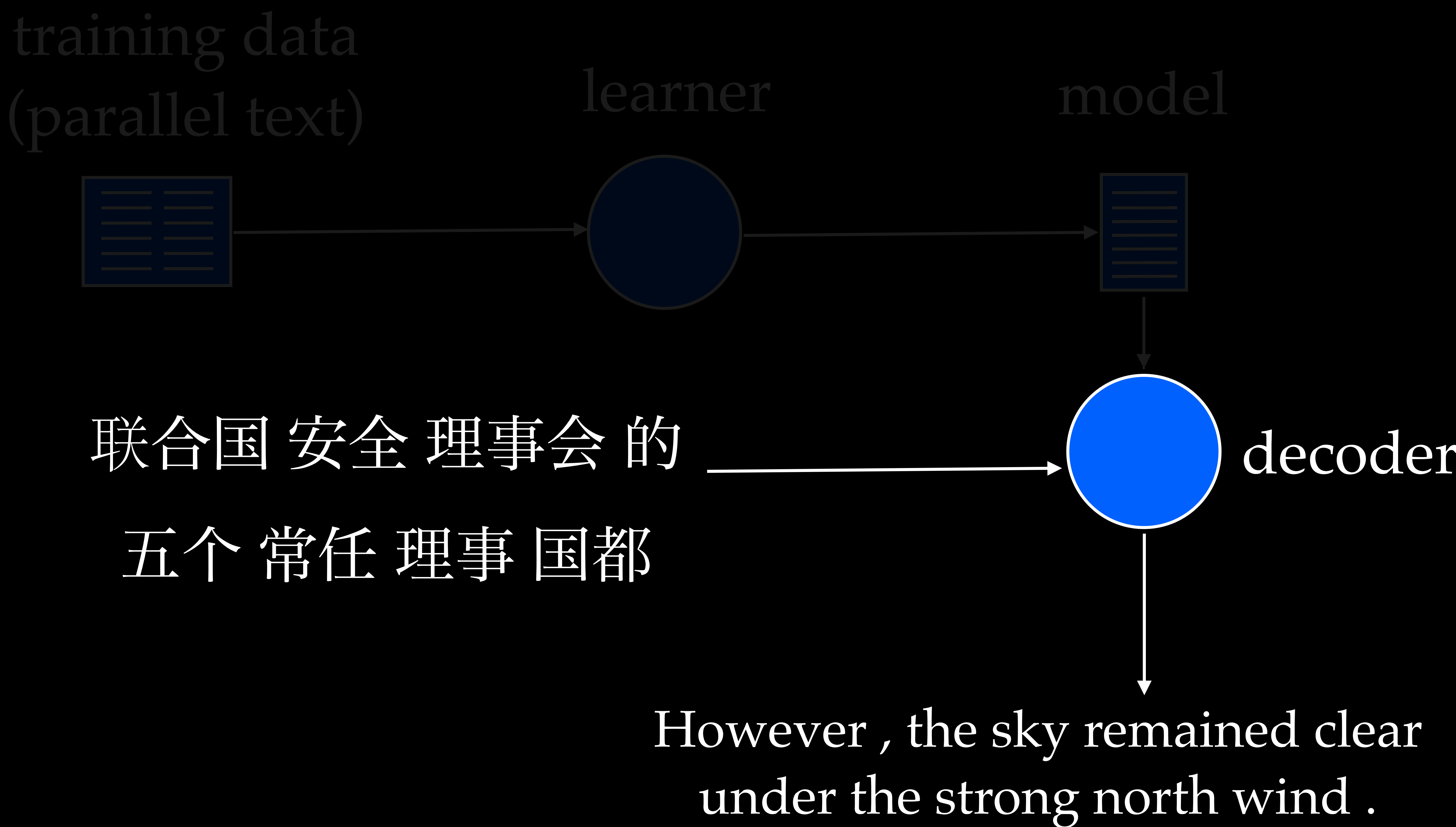
Summary

- ***Learning is optimization***: choose parameters that optimize some function, such as likelihood.
- Supervised: maximum likelihood.
  - Beware of overfitting.
- Unsupervised: expectation maximization.
- Many, many, many other algorithms.
- Next up: prediction, better models.

Overview


training data
(parallel text)
learner
model
decoder
联合国 安全 理事会 的
五个 常任 理事 国都
However , the sky remained clear
under the strong north wind .

Overview


training data
(parallel text)
learner
model
联合国 安全 理事会 的
五个 常任 理事 国都
decoder
However , the sky remained clear
under the strong north wind .

Quick Recap

$$p(English|Chinese) =$$

$$\frac{p(English) \times p(Chinese|English)}{p(Chinese)}$$

language model

translation model

normalization term (ensures we're working with valid probabilities).

Quick Recap

$$p(English|Chinese) \sim$$

$$p(English) \times p(Chinese|English)$$

language model

translation model

Decoding

Probability models enable us to ***make predictions***: Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?

Decoding

Probability models enable us to ***make predictions***: Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?

In math, we want to solve:

$$\text{argmax}_{English}\, p(English|Chinese)$$

Decoding

Probability models enable us to ***make predictions***:
Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?

In math, we want to solve:

$$\mathrm{argmax}_{English}\, p(English|Chinese)$$

problem: there are a lot of English sentences to choose from!


THE WØRD

- Optimization

北 风 呼啸 。

北 风 呼啸 。

substitutions

permutations

北 风 呼啸 。

substitutions $O(5^n)$

permutations

北 风 呼啸 。

substitutions	$O(5^n)$
permutations	$O(n!)$

北 风 呼啸 。

substitutions $O(5^n)$

permutations $O(n!)$

15,000 possibilities!

北 风 呼啸 。

Can we do this without enumerating $O(5^n n!)$ pairs?

北 风 呼啸 。

the strong north wind .

Can we do this without enumerating $O(5^n n!)$ pairs?

北 风 呼啸 。

the strong north wind .

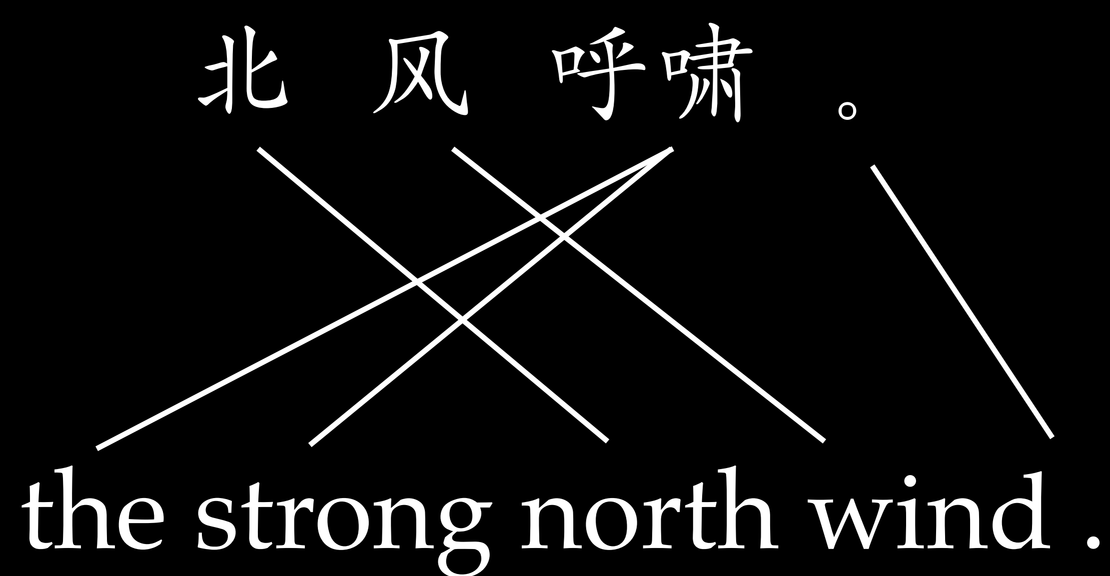
Can we do this without enumerating $O(5^n n!)$ pairs?

Given a sentence pair and an alignment, we can easily calculate

$$p(English, alignment|Chinese)$$

Can we do this without enumerating $O(5^n n!)$ pairs?

Key Idea



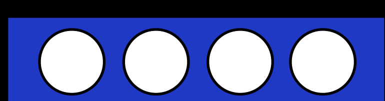
There are $O(5^n n!)$ target sentences.

But there are only $O(5^n)$ ways to start them.

Key Idea

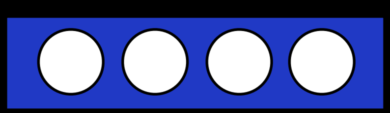
北 风 呼啸 。

Key Idea



北 风 呼啸 。

Key Idea



coverage vector

北 风 呼啸 。

Key Idea

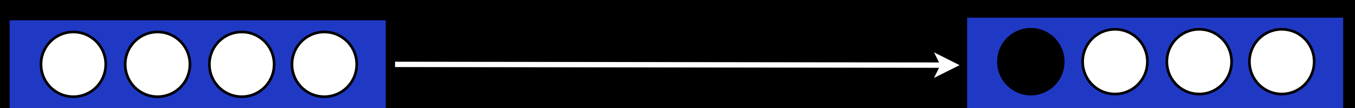
coverage vector

北 风 呼啸 。

Key Idea

$$p(north|START) \cdot p(北\,|north)$$

north

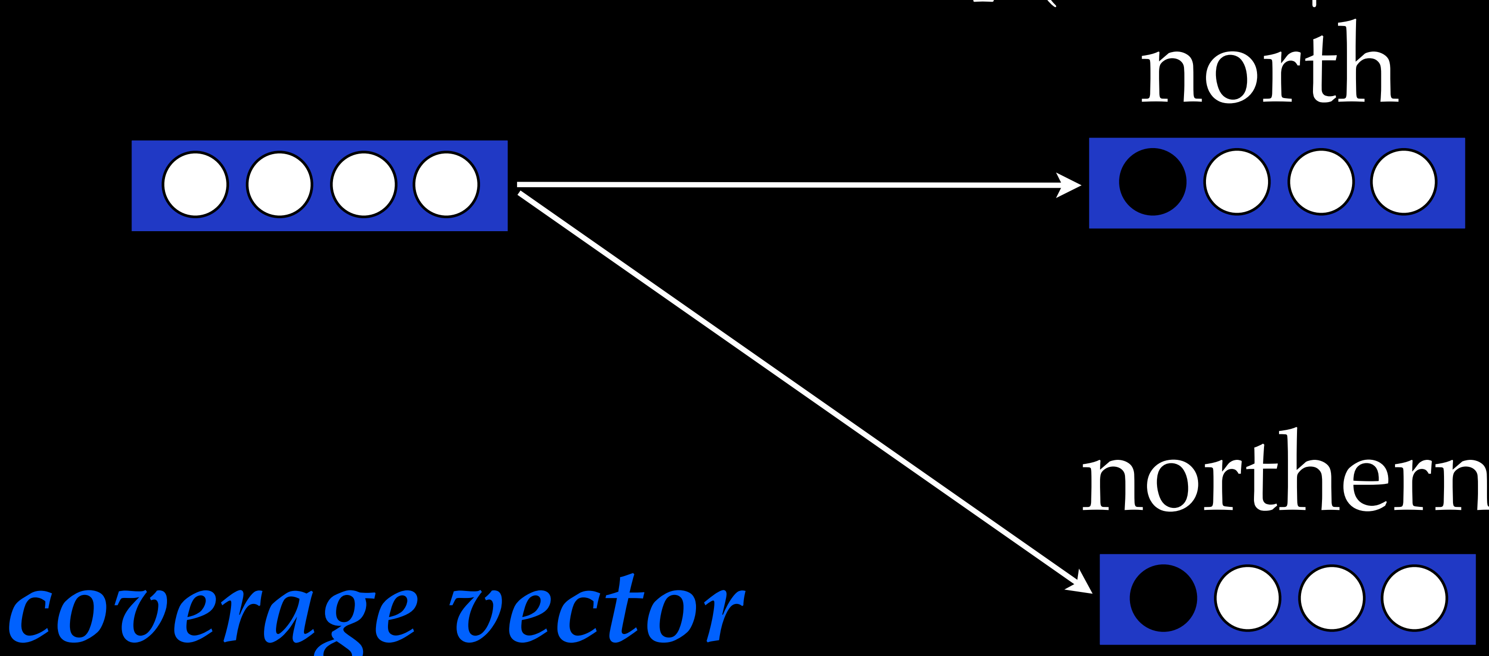


coverage vector

北 风 呼啸 。

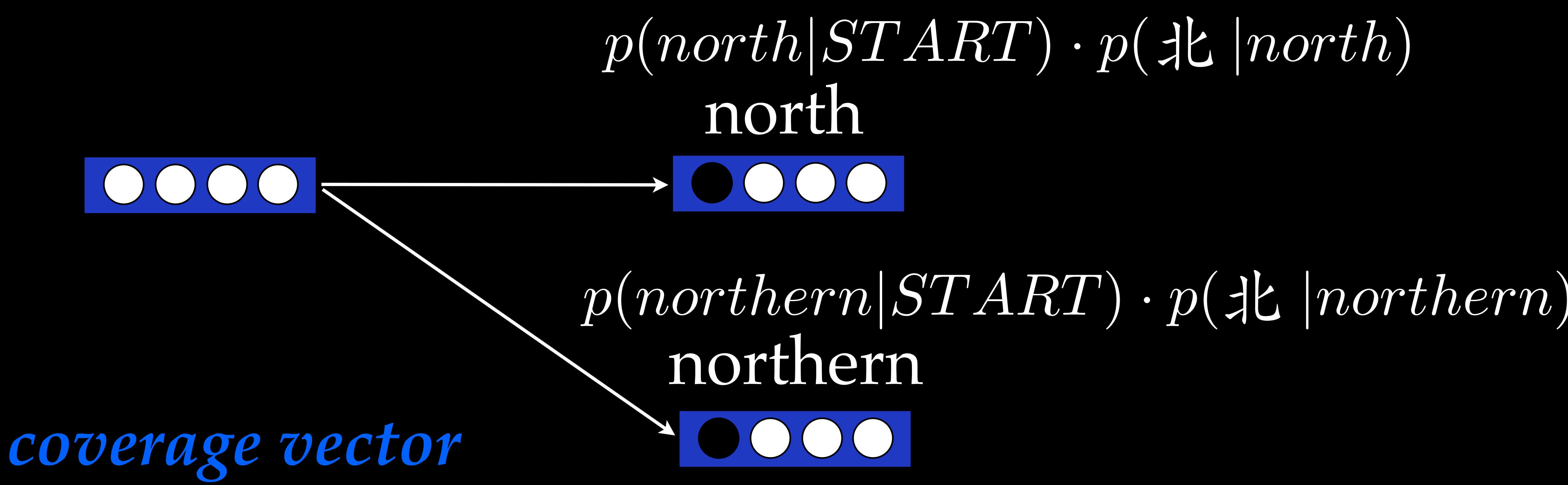
Key Idea

$$p(north|START) \cdot p(\text{北}\,|north)$$



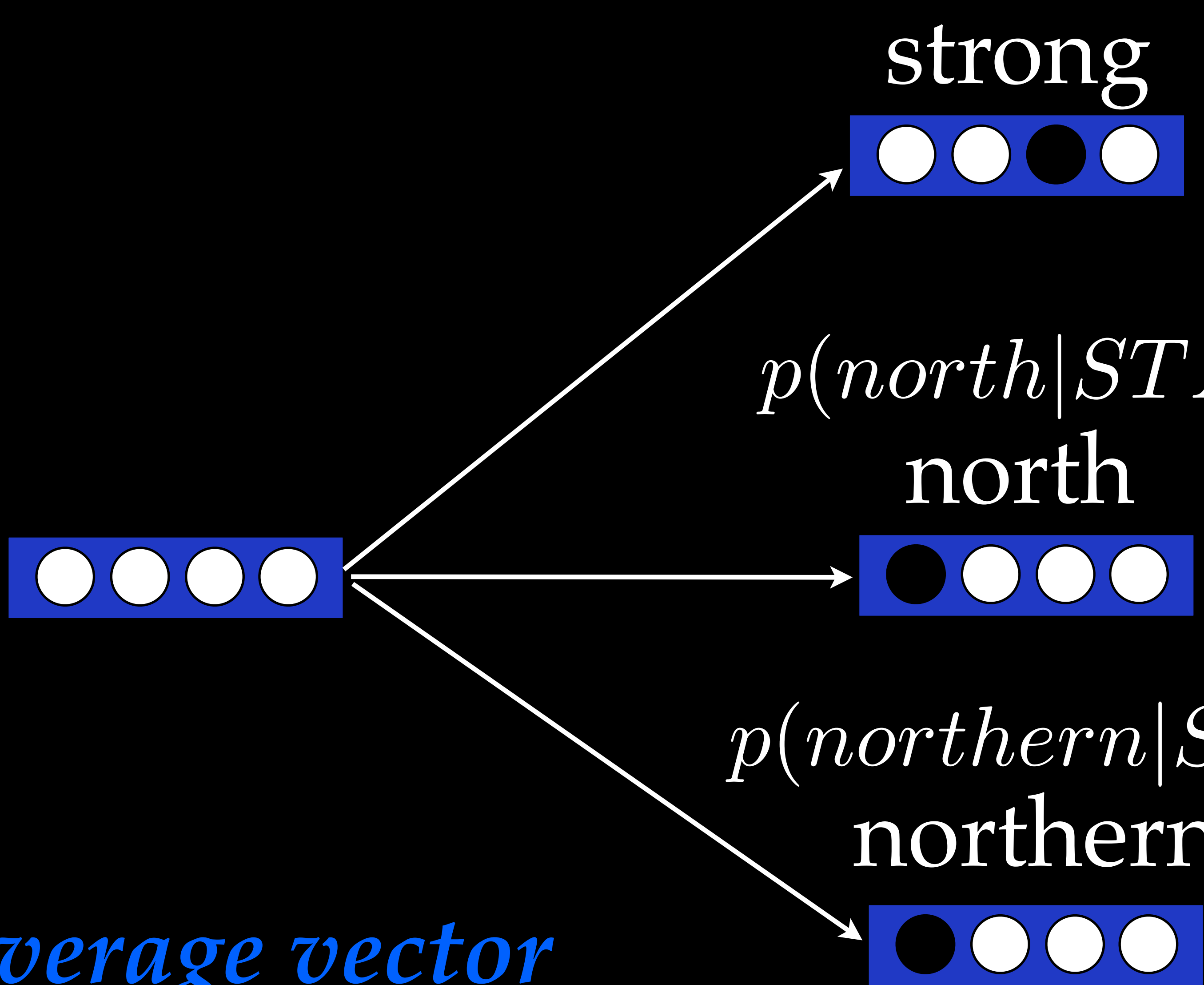
北 风 呼啸 。

Key Idea



北 风 呼啸 。

Key Idea



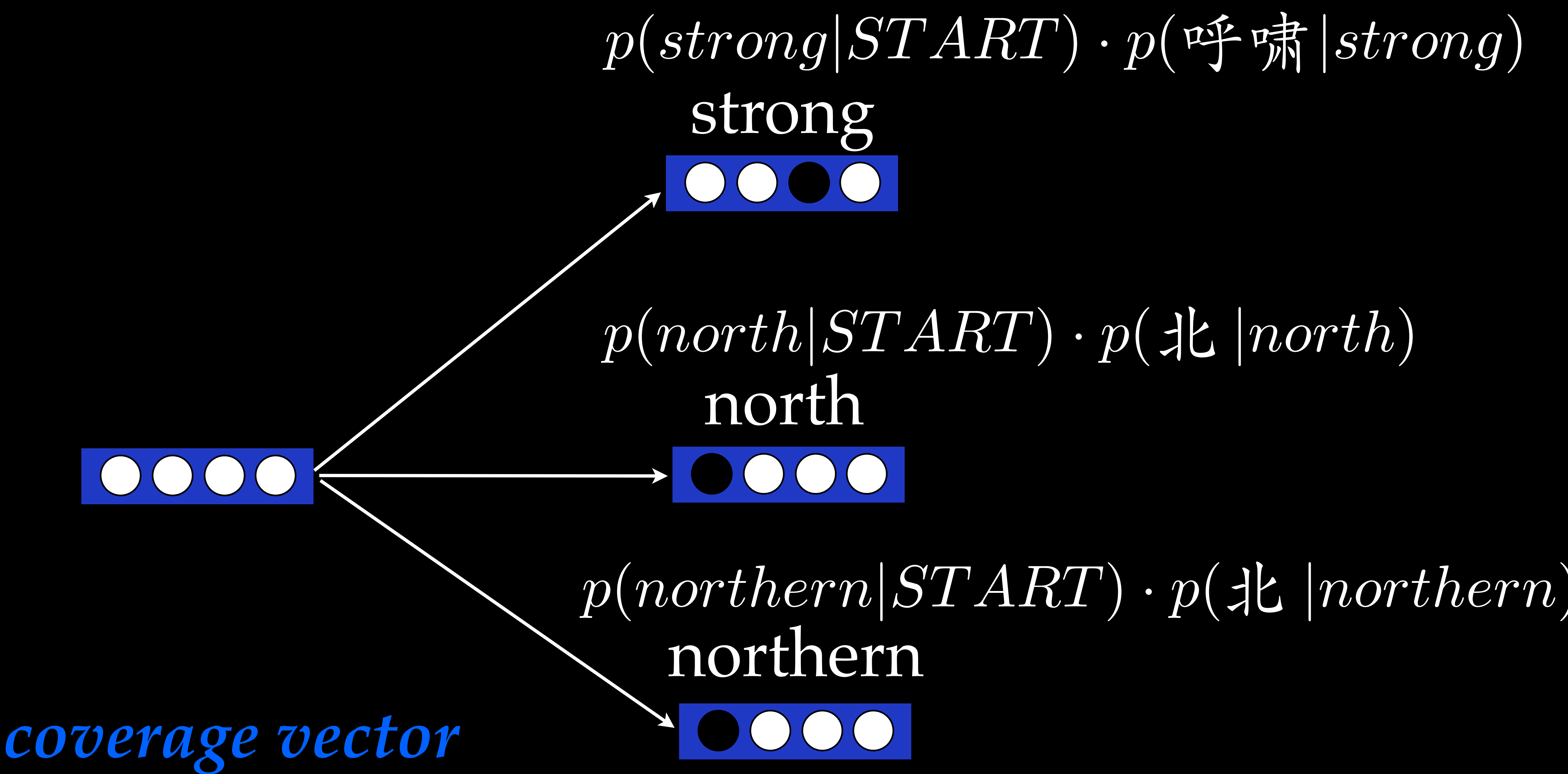
$$p(north|START) \cdot p(北|north)$$

$$p(northern|START) \cdot p(北|northern)$$

coverage vector

北 风 呼啸 。

Key Idea

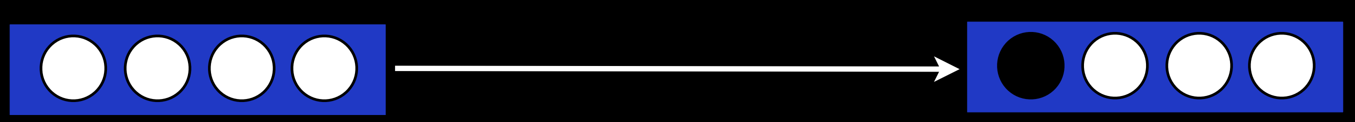


北 风 呼啸 。

Key Idea

$$p(north|START) \cdot p(\text{北}|north)$$

north

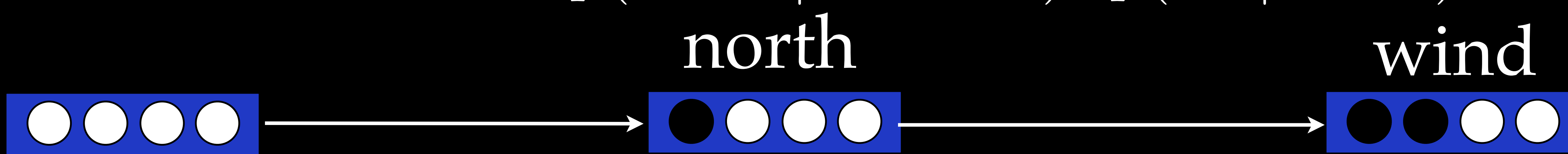


coverage vector

北 风 呼啸 。

Key Idea

$$p(north|START) \cdot p(\text{北}\,|north)$$



coverage vector

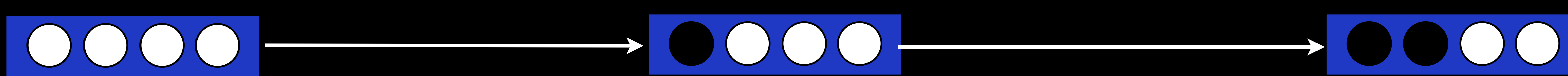
北 风 呼啸 。

Key Idea

$p(north|START) \cdot p(北|north)$

north

wind

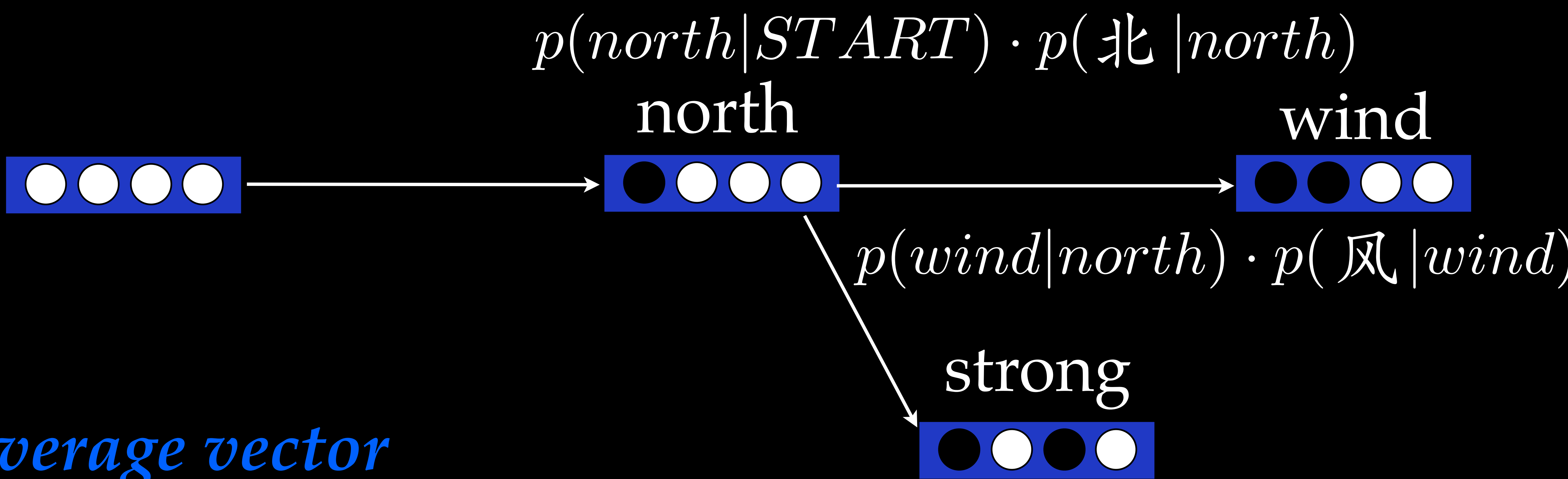


$p(wind|north) \cdot p(风|wind)$

coverage vector

北 风 呼啸 。

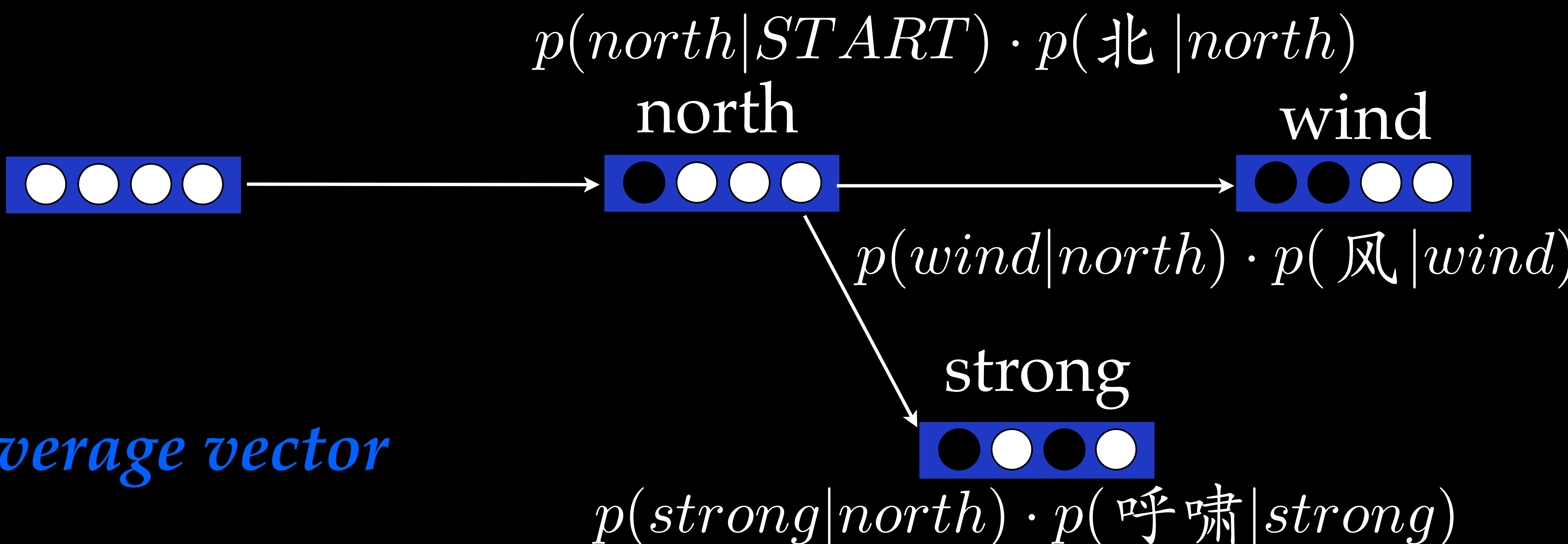
Key Idea



coverage vector

北 风 呼啸 。

Key Idea

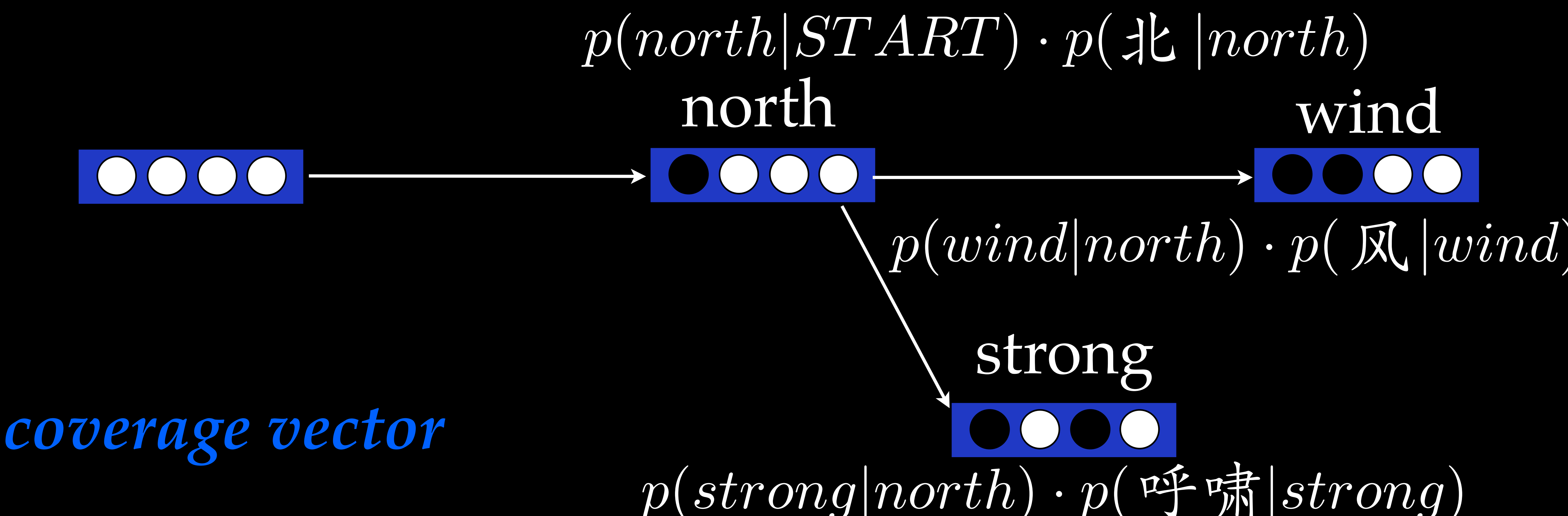


coverage vector

北 风 呼啸 。

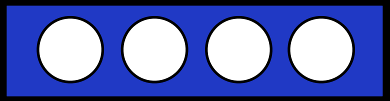
Key Idea

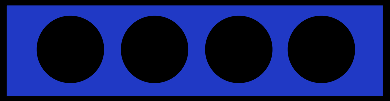
Work done at sentence beginnings is shared across many possible output sentences!



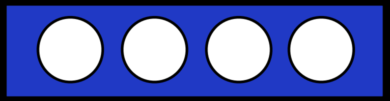
北　风　呼啸　。

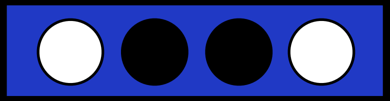
Key Idea

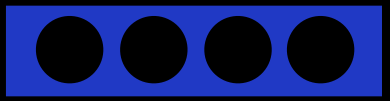




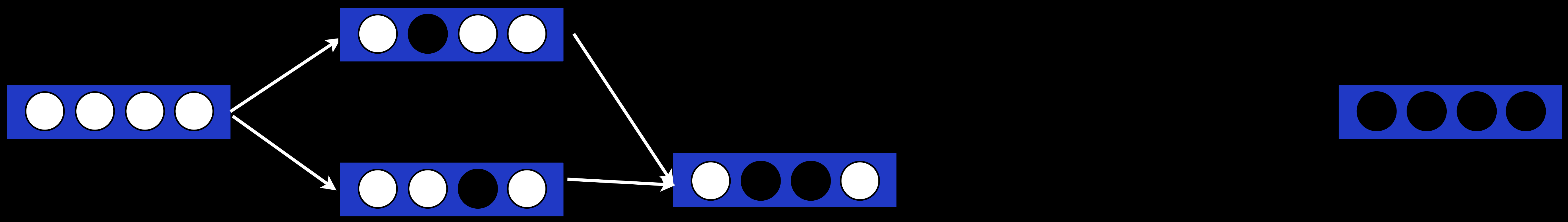
Key Idea



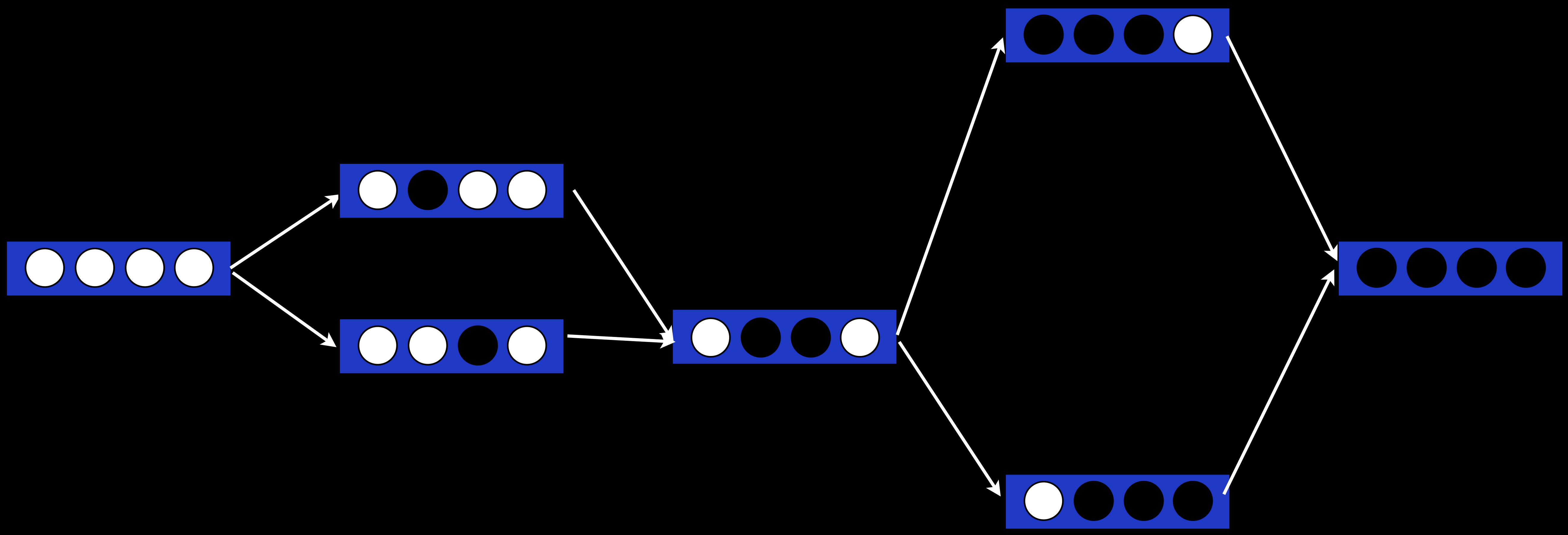




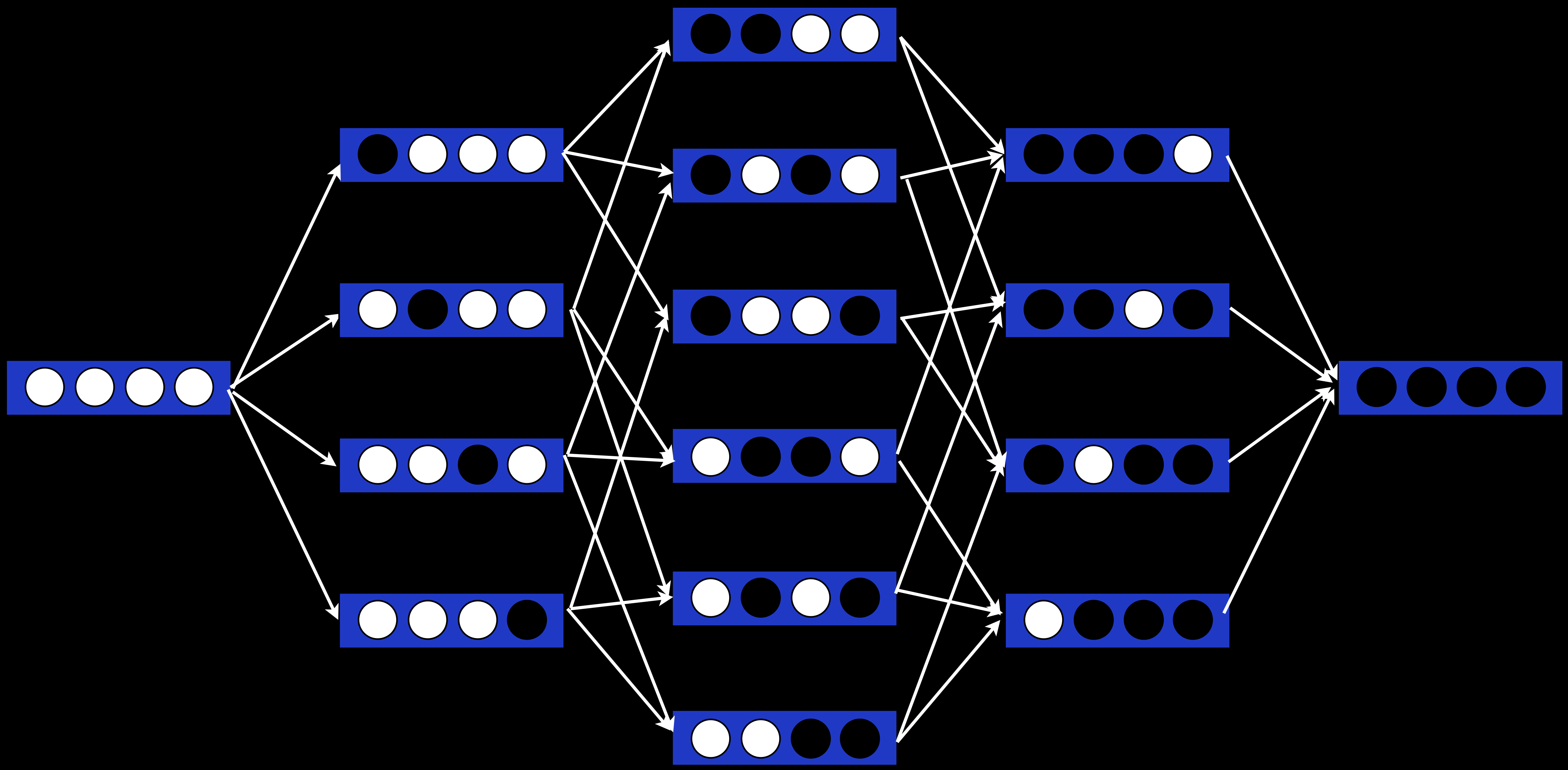
Key Idea



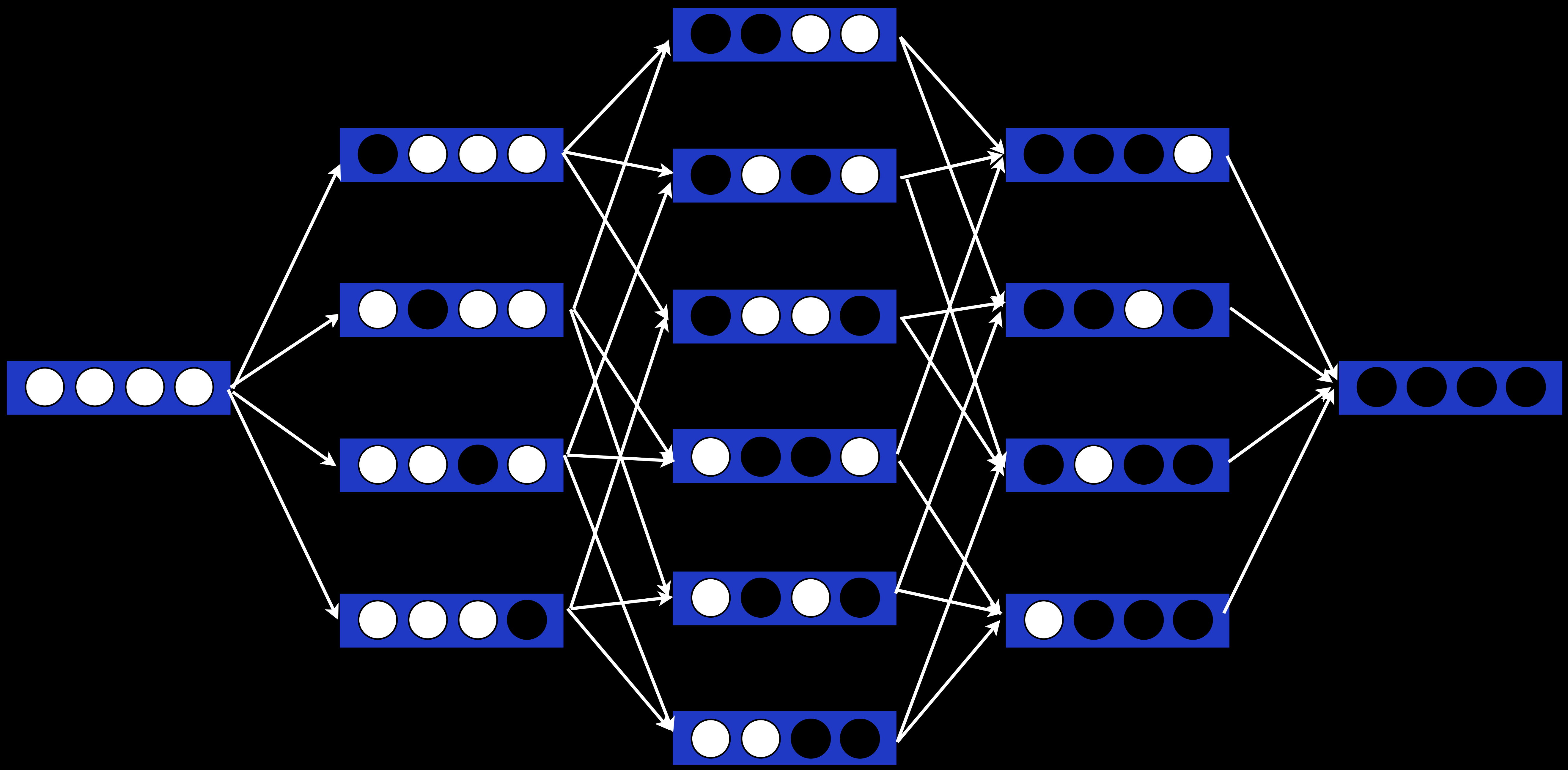
Key Idea



Key Idea



Key Idea

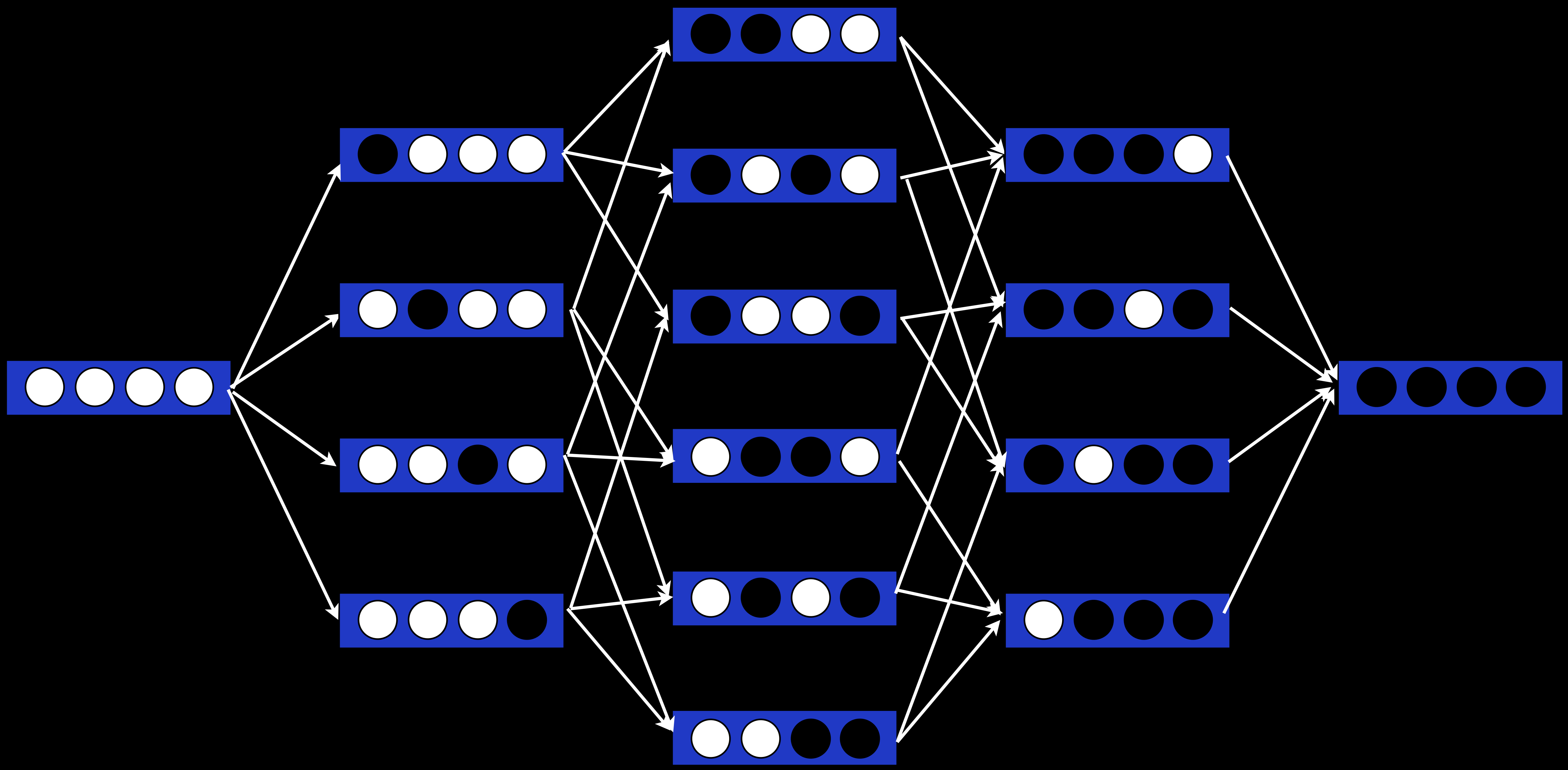


Dynamic Programming

Key Idea

amount of work:

$$O(5^n 2^n)$$



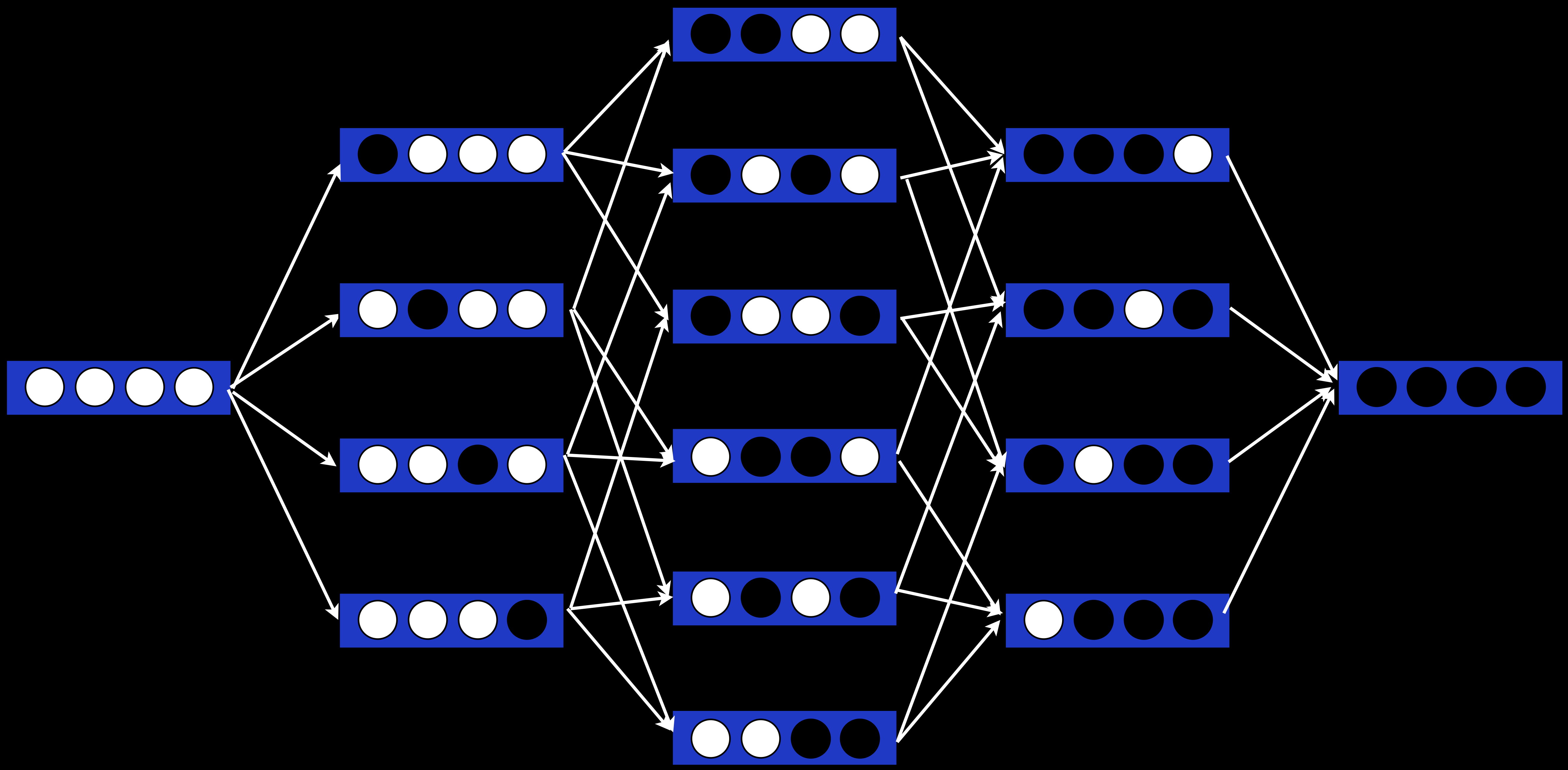
Dynamic Programming

Key Idea

amount of work:

$O(5^n 2^n)$

bad, but much better than $O(5^n n!)$

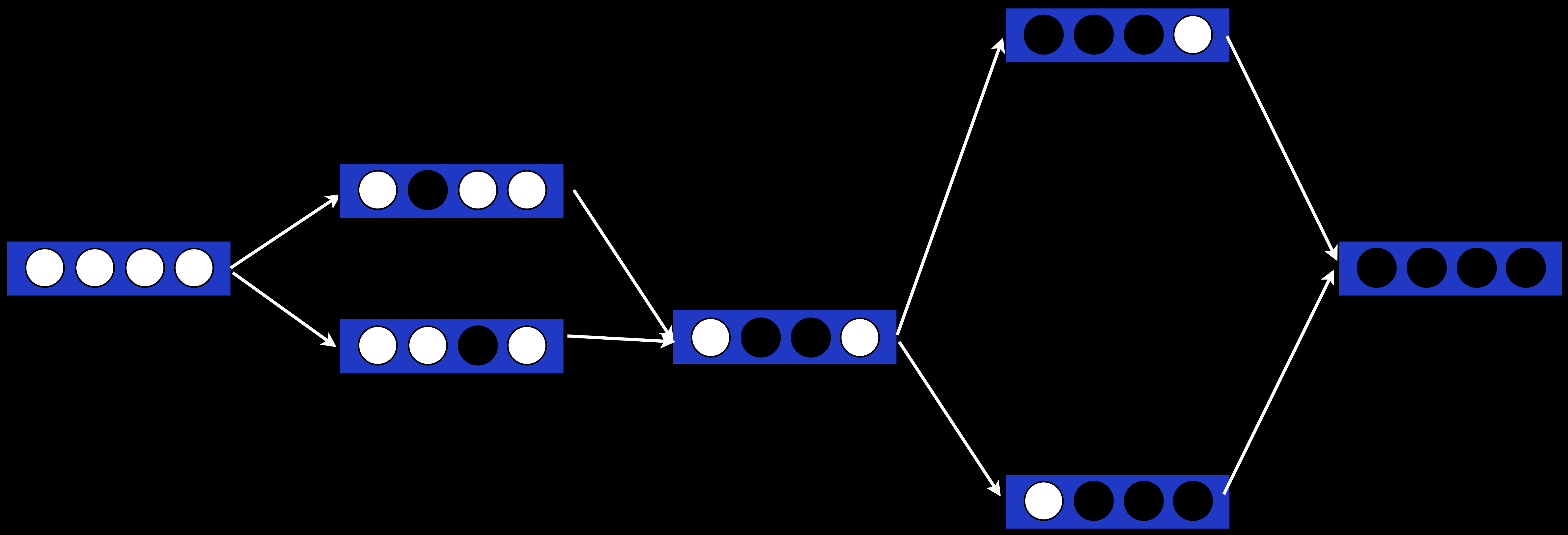


Dynamic Programming

Key Idea

amount of work: $O(5^n 2^n)$

bad, but much better than $O(5^n n!)$

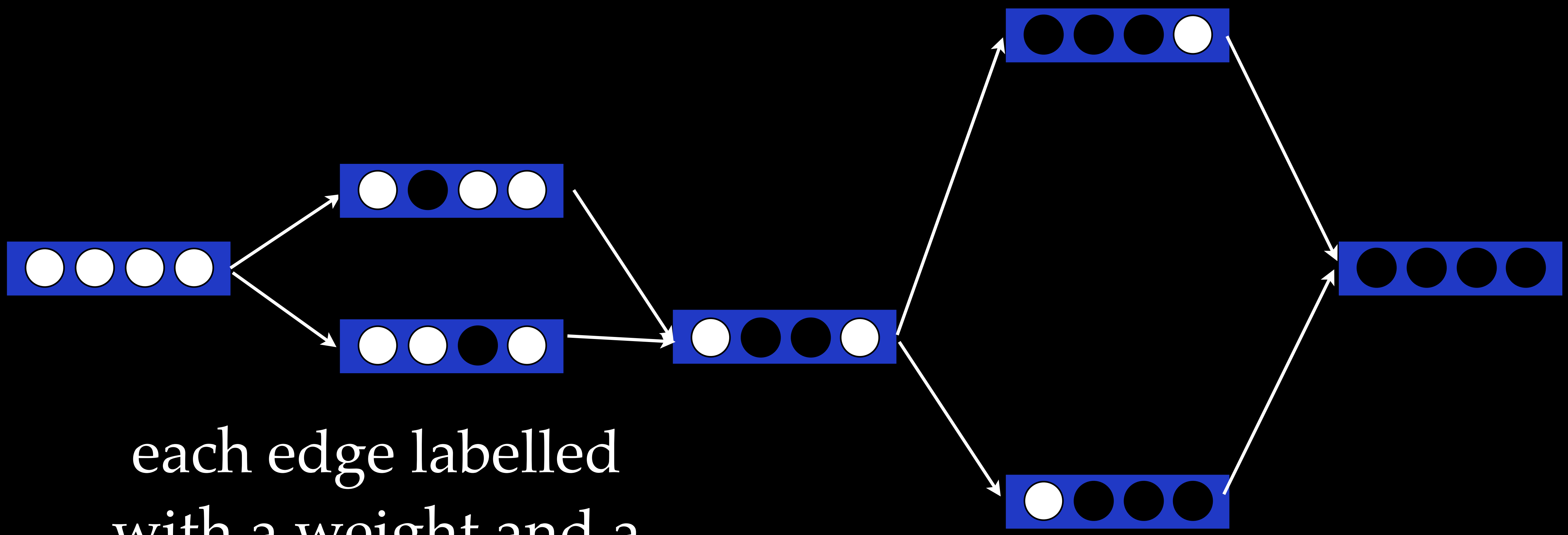


Dynamic Programming

Key Idea

amount of work: $O(5^n 2^n)$

bad, but much better than $O(5^n n!)$



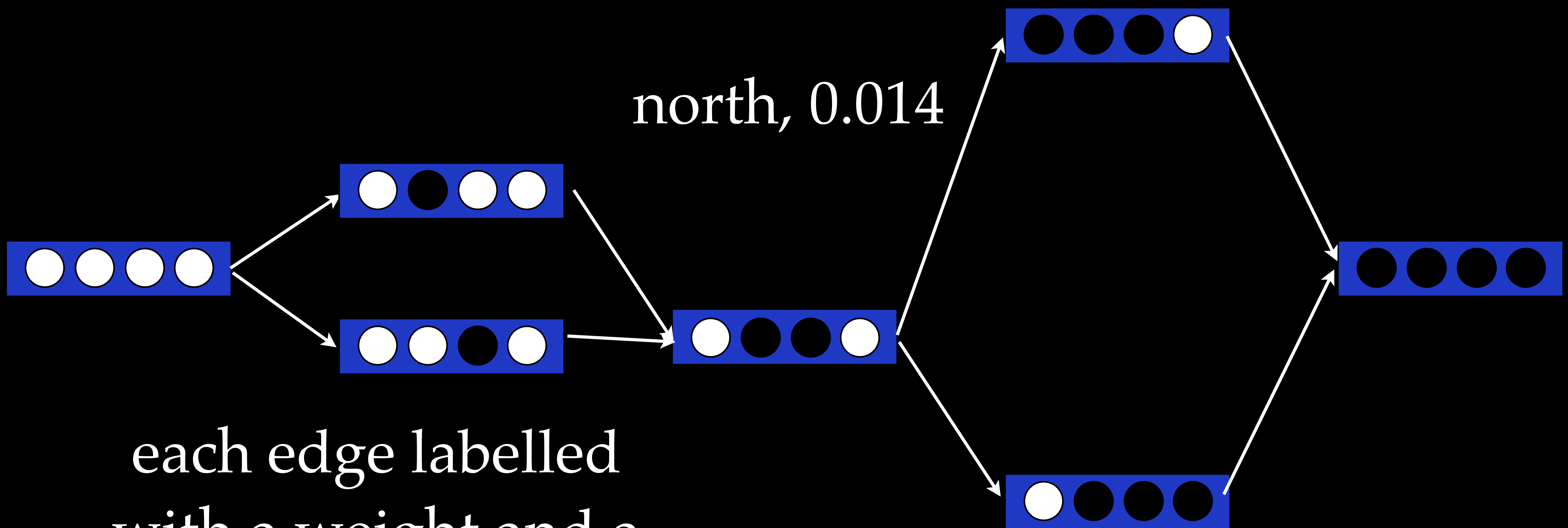
each edge labelled with a weight and a word (or words)

Dynamic Programming

Key Idea

amount of work: $O(5^n 2^n)$

bad, but much better than $O(5^n n!)$



each edge labelled with a weight and a word (or words)

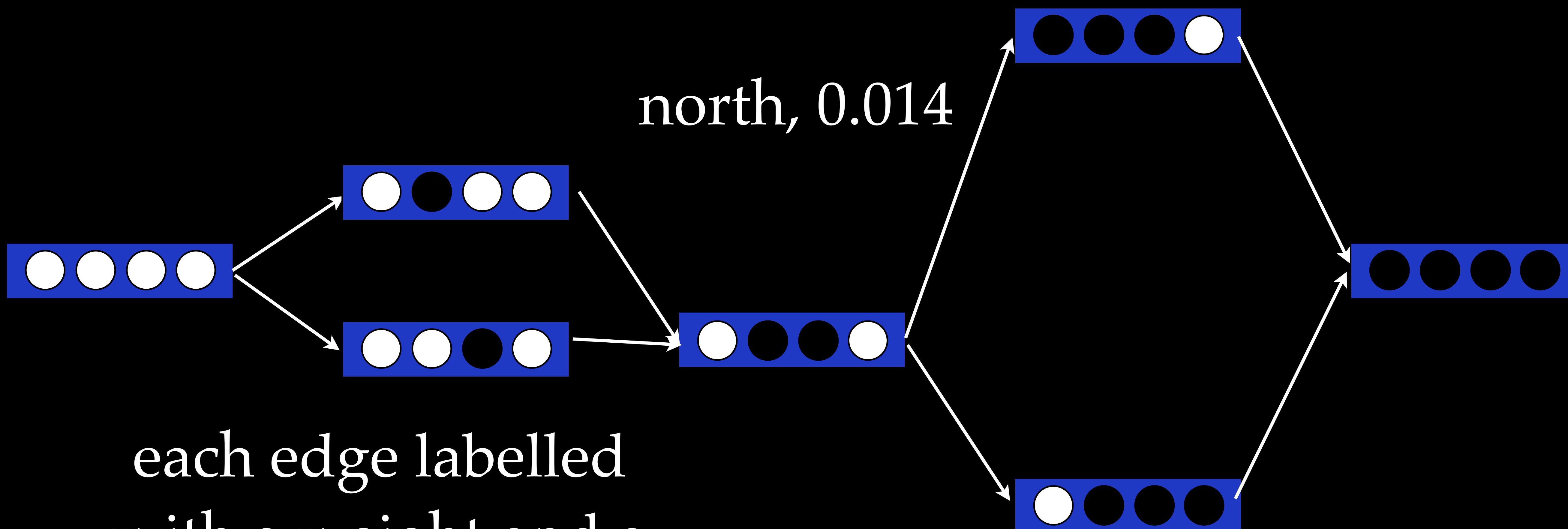
Dynamic Programming

Key Idea

amount of work: $O(5^n 2^n)$

bad, but much better than $O(5^n n!)$

weighted finite-state automata



each edge labelled with a weight and a word (or words)

Dynamic Programming

Weighted languages

- The lattice describing the set of all possible translations is a ***weighted finite state automaton***.
- So is the language model.
- Since regular languages are closed under intersection, we can intersect the devices and run shortest path graph algorithms.
- Taking their intersection is equivalent to computing the probability under Bayes' rule.

Practical Issues

$O(5^n 2^n)$ is still far too much work.

Practical Issues

$O(5^n 2^n)$ is still far too much work.

Can we do better?

Can we do better?

北 风 呼啸 。

Can we do better?

北 风 呼啸 。

north wind the strong .

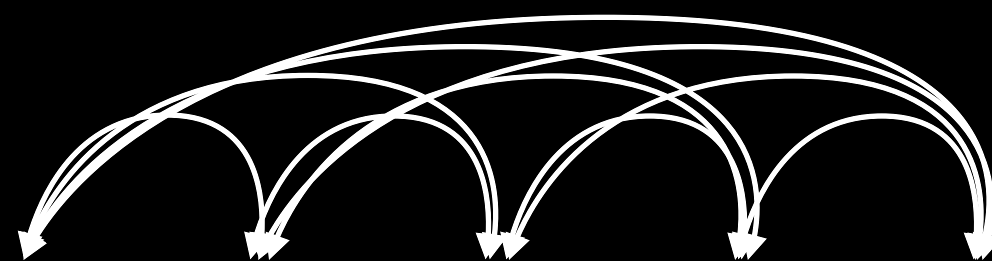
Can we do better?

北 风 呼啸 。

north wind the strong .

Can we do better?

北 风 呼啸 。

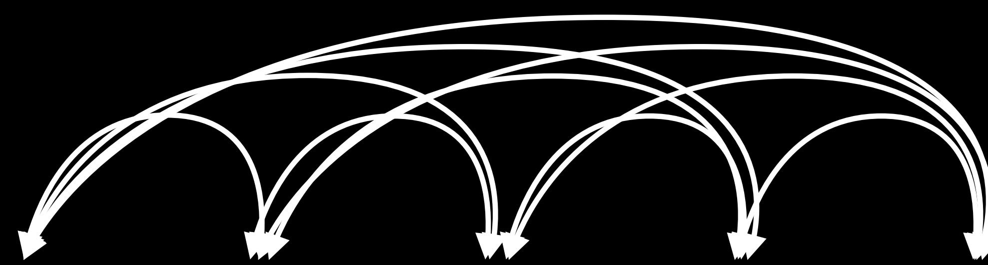


north wind the strong .

Each arc weighted by
translation probability +
bigram probability

Can we do better?

北 风 呼啸 。



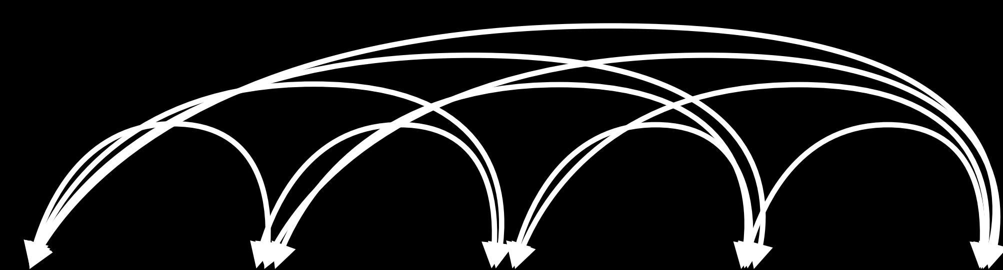
north wind the strong .

Each arc weighted by
translation probability +
bigram probability

Objective: find shortest path that visits each word once.

Can we do better?

北 风 呼啸 。



London Paris NY Tokyo .

Each arc weighted by
translation probability +
bigram probability

Objective: find shortest path that visits each word once.

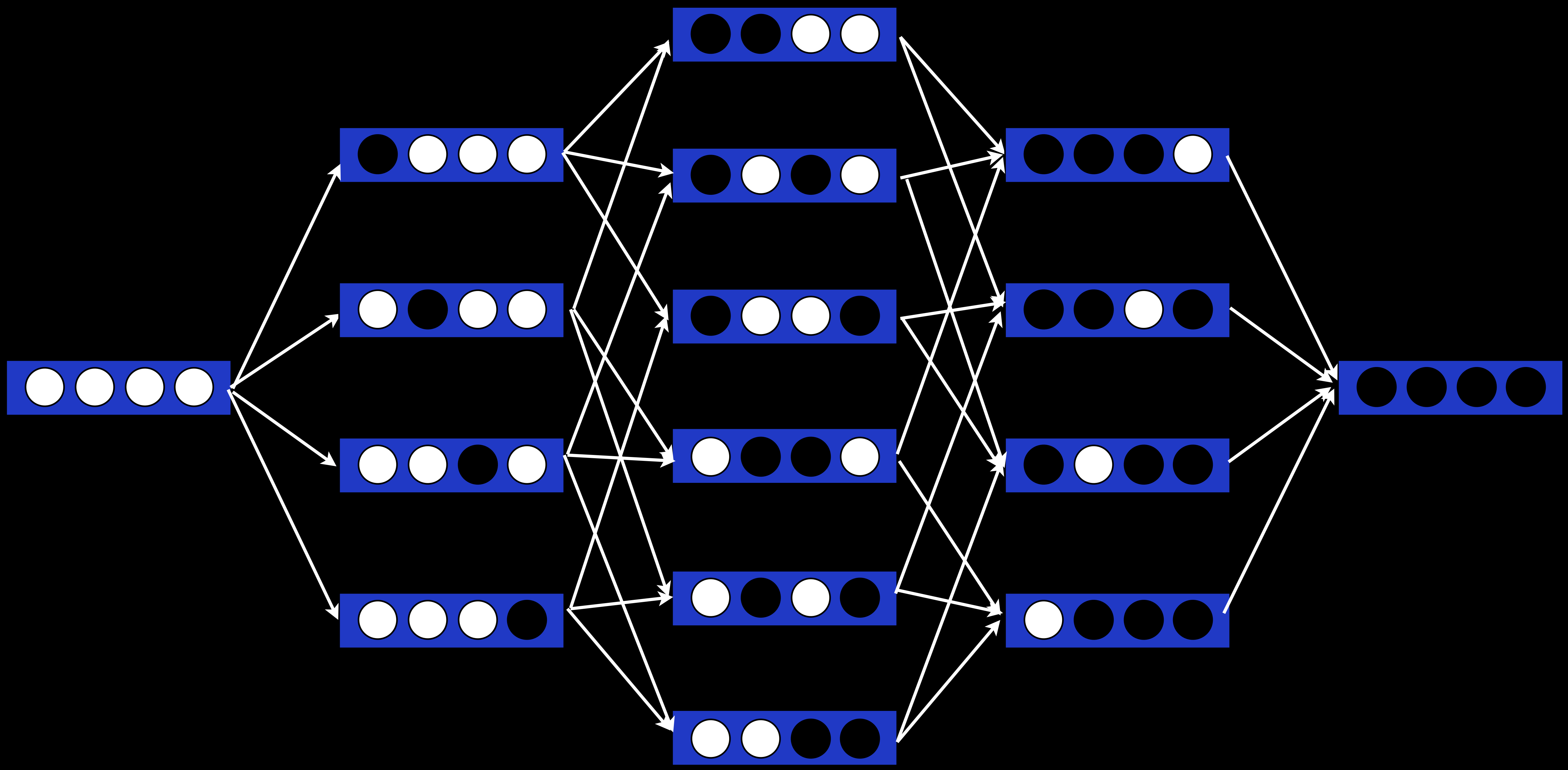
Can we do better?

Probably not: this is the traveling salesman problem.

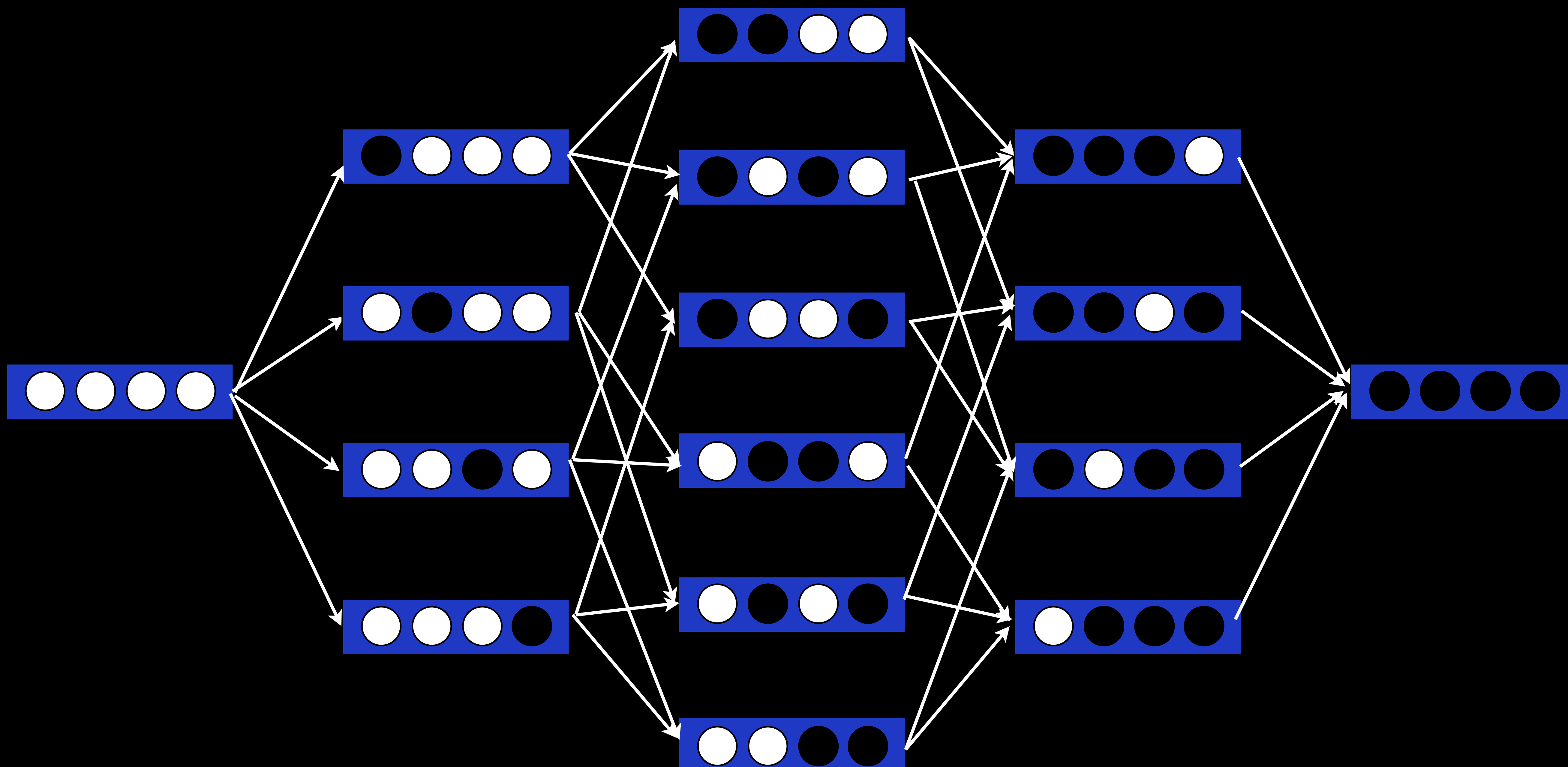
Each arc weighted by
translation probability +
bigram probability

Objective: find shortest path that visits each word once.

Approximation: Pruning

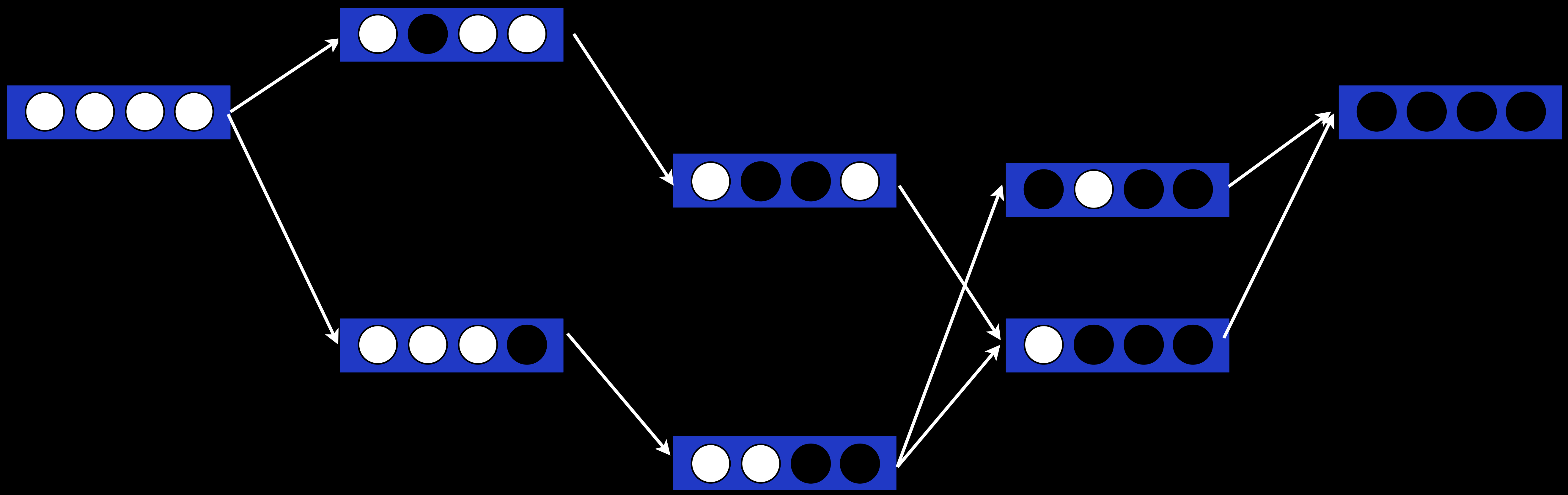


Approximation: Pruning



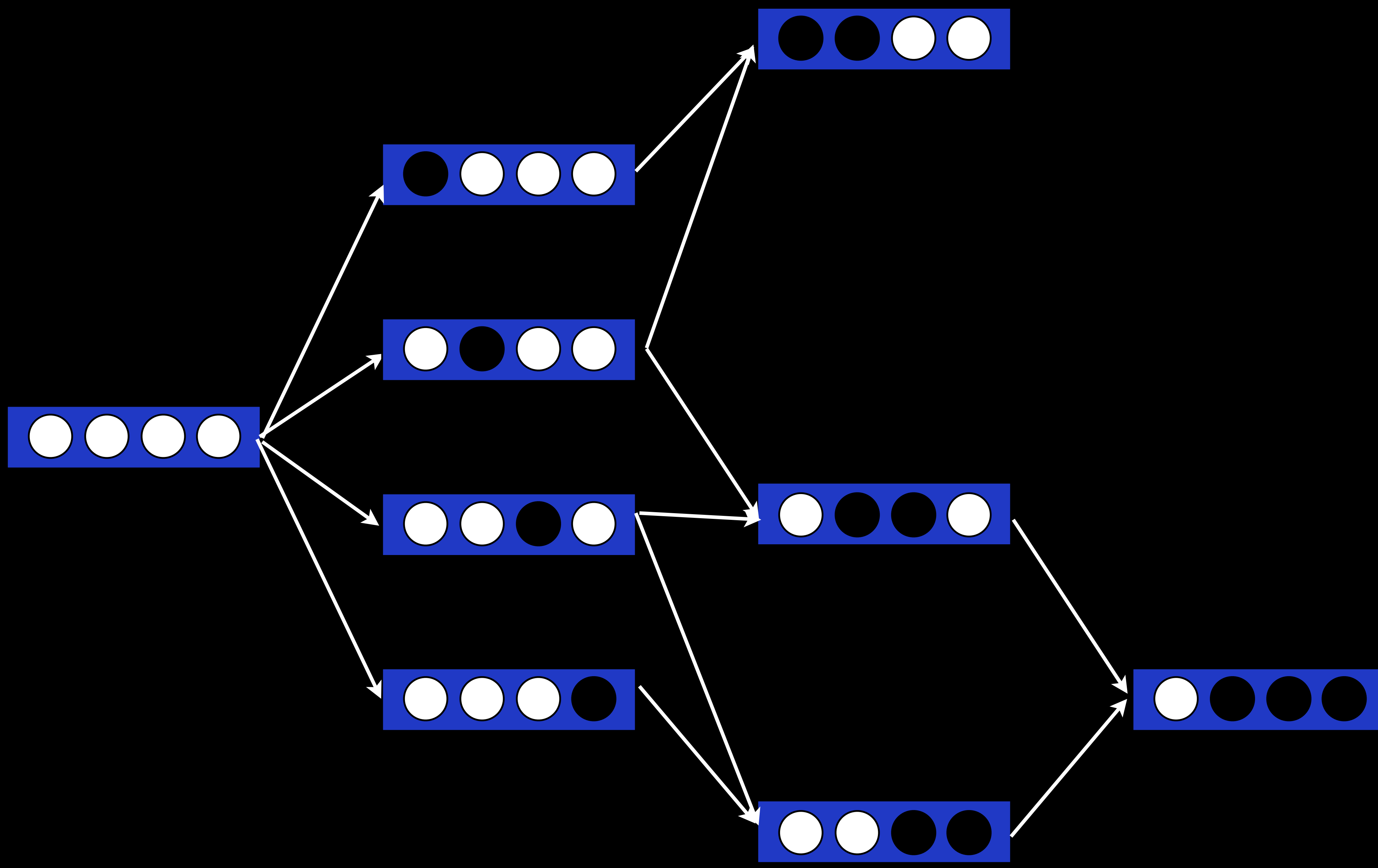
Idea: prune states by accumulated path length

Approximation: Pruning

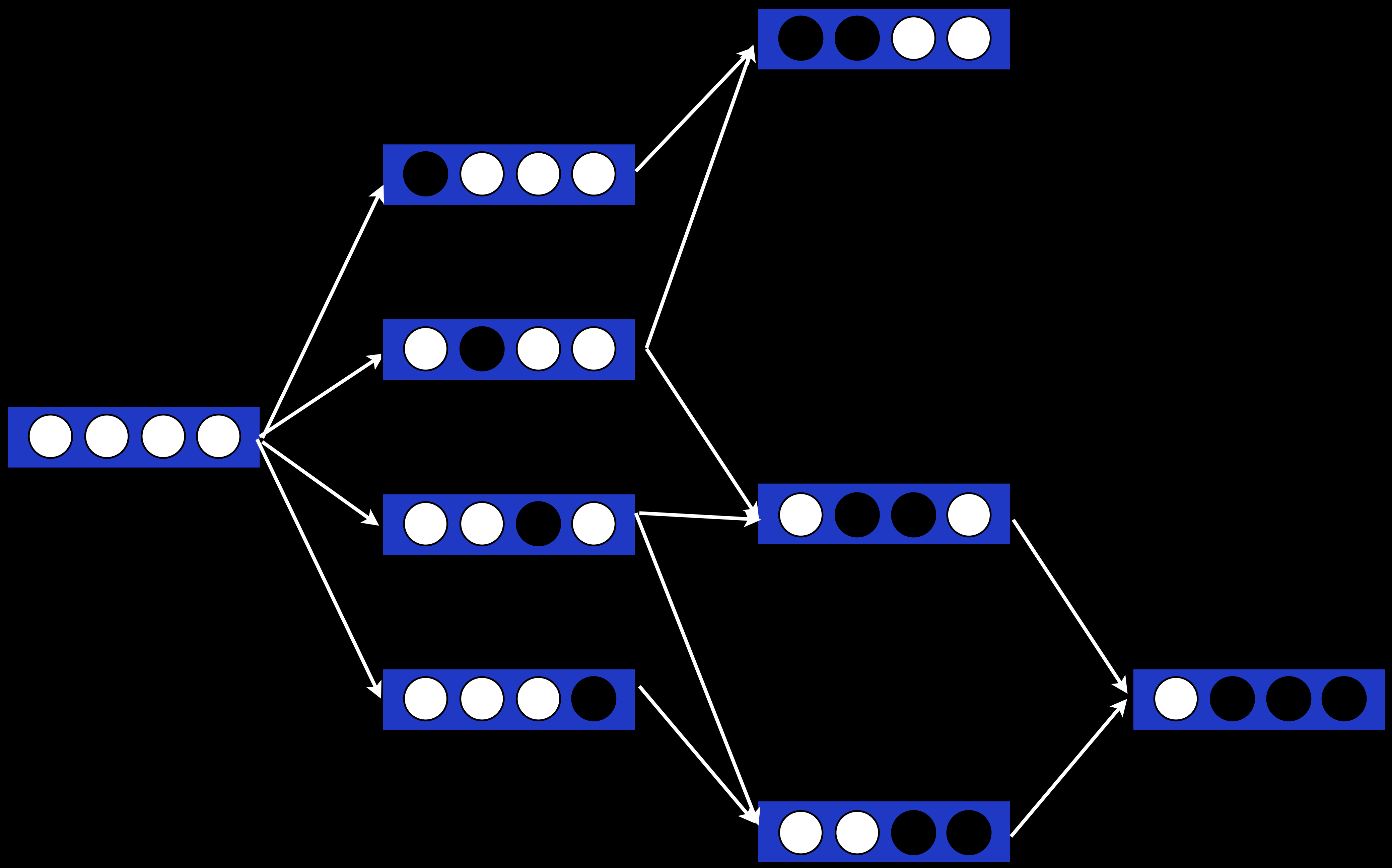


Idea: prune states by accumulated path length

Approximation: Pruning

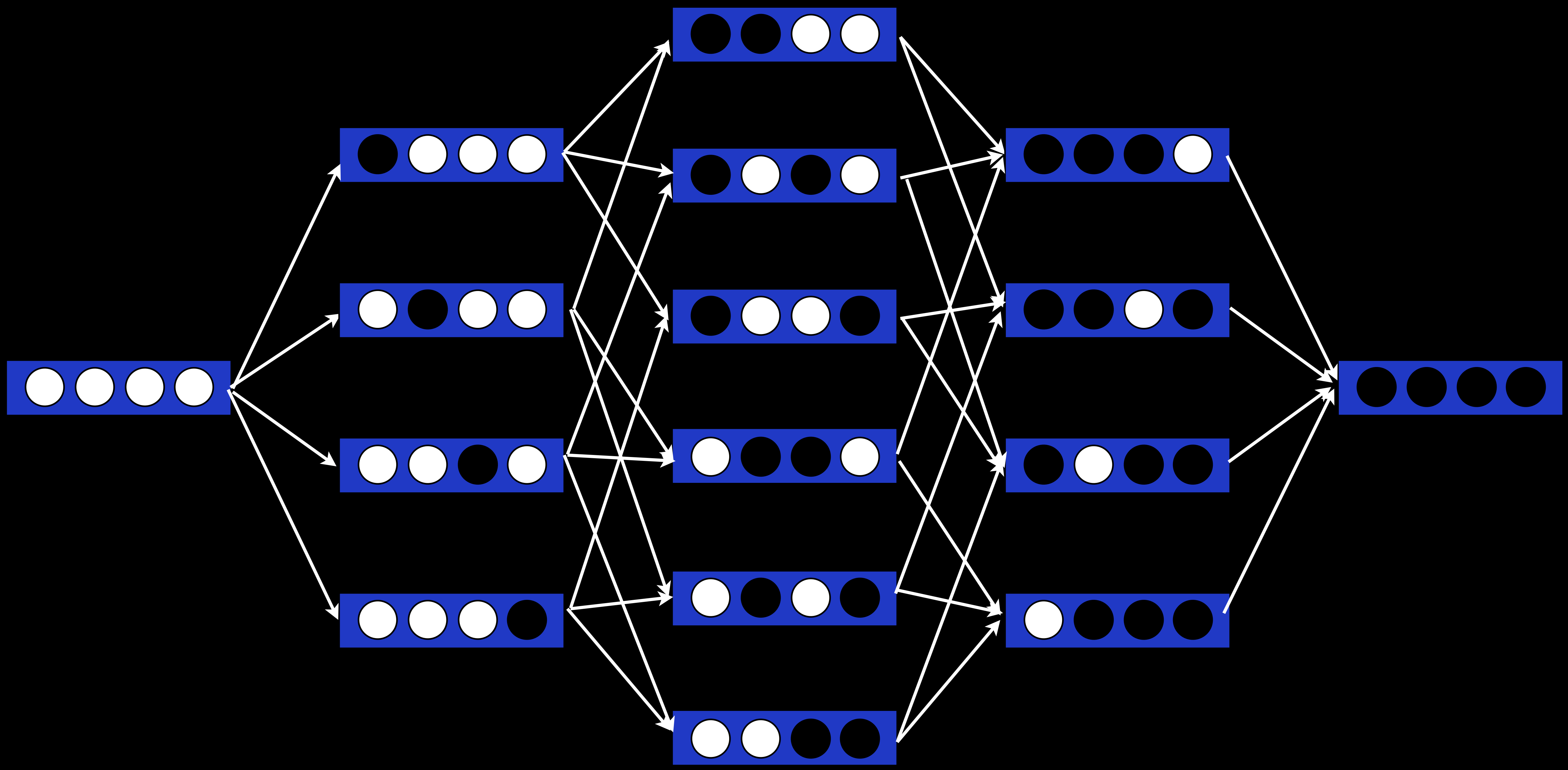


Approximation: Pruning

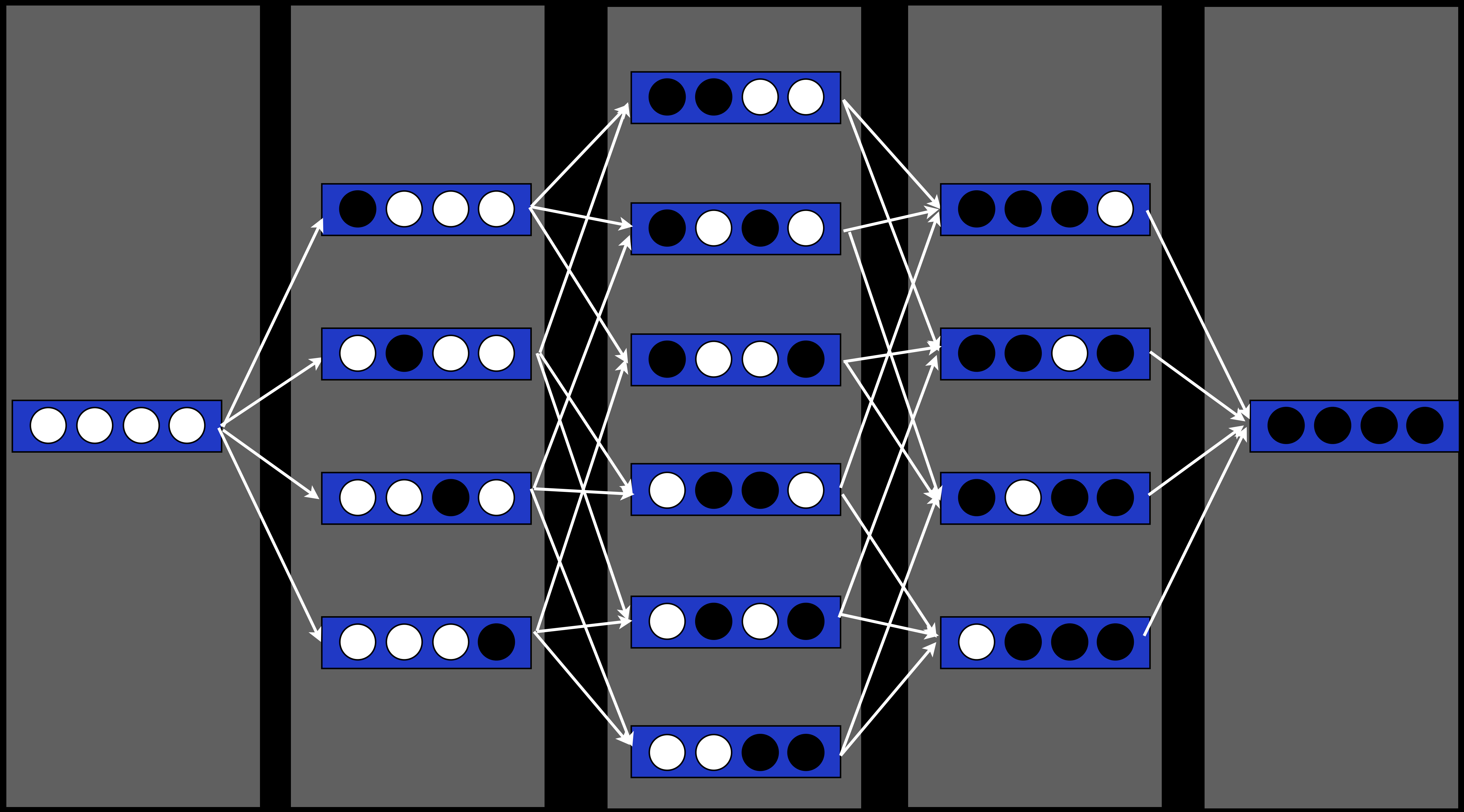


Reality: longer paths have lower probability!

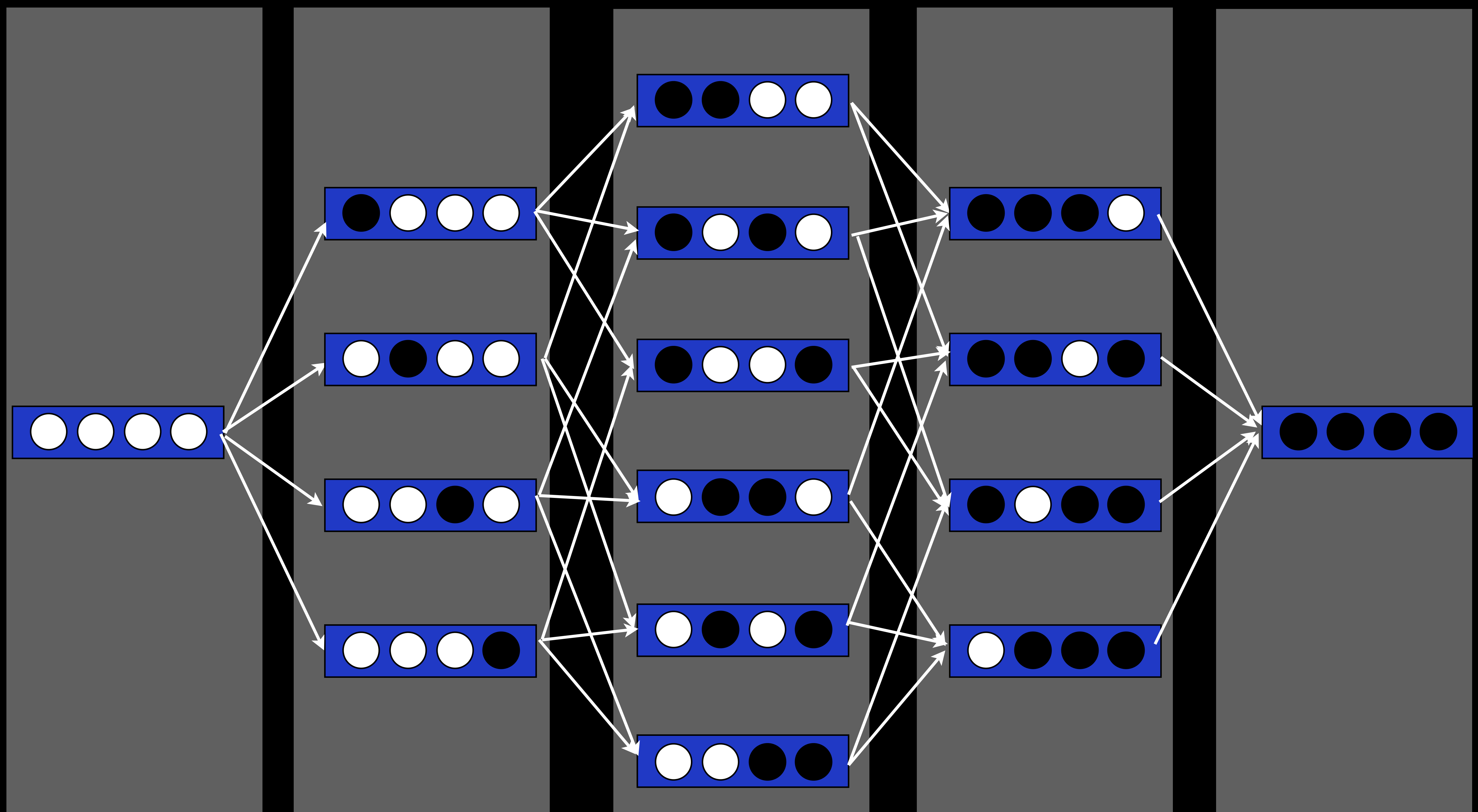
Approximation: Pruning



Approximation: Pruning

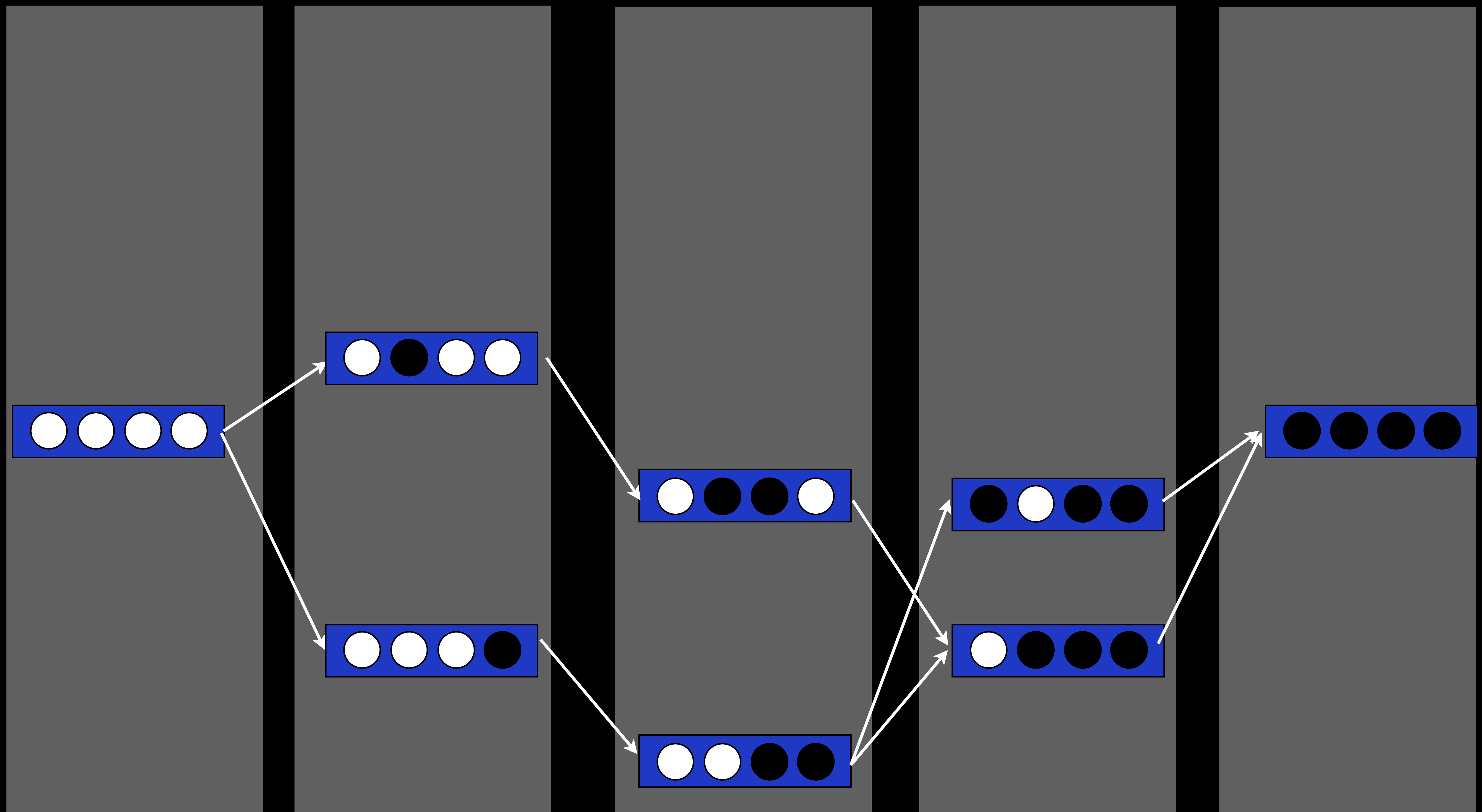


Approximation: Pruning



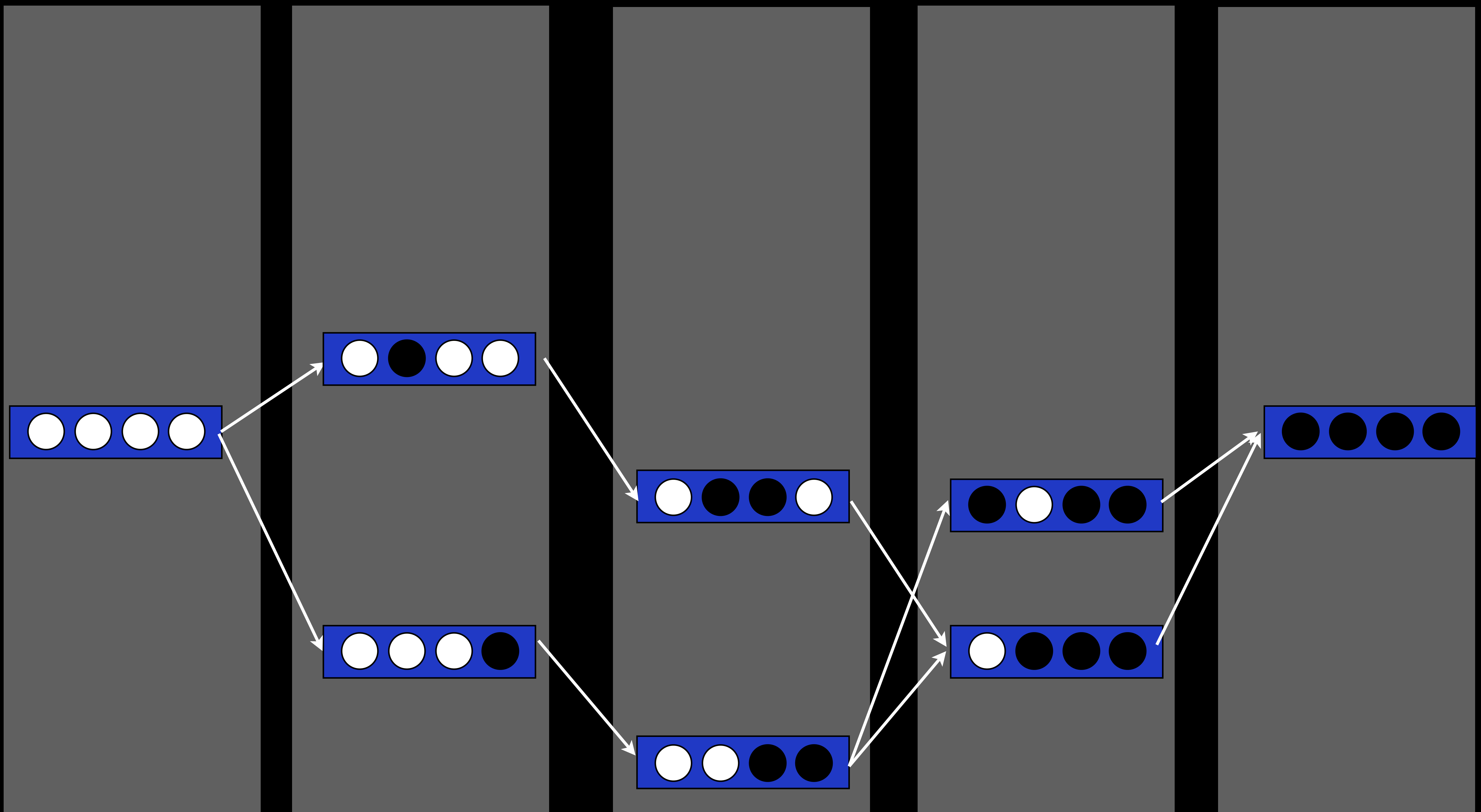
Solution: Group states by number of covered words.

Approximation: Pruning



Solution: Group states by number of covered words.

Approximation: Pruning



“Stack” decoding: a linear-time approximation

Approximation: Distortion Limits

the sky

虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

Approximation: Distortion Limits

number of vertices: $O(2^n)$

the sky

虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

Approximation: Distortion Limits

number of vertices: $O(2^n)$

the sky

$\boxed{\text{虽然 北 风 呼啸}}$, $\boxed{\text{但 天空}}$ 依然 十分 清澈 。

$d = 4$

window

Approximation: Distortion Limits

number of vertices: $O(2^n)$

the sky

虽然 北 风 呼啸 , 但 天空 依然 十分 清澈 。

outside window
to left: covered

$d = 4$

window

outside window
to right: uncovered

Approximation: Distortion Limits

number of vertices: $O(n2^d)$

the sky

$\boxed{\text{虽然 北 风 呼啸}}$, $\boxed{\text{但 天空}}$ 依然 十分 清澈 。

outside window to left: covered

$d = 4$

window

outside window to right: uncovered

Summary

Summary

- We need every possible trick to make decoding fast.

Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.

Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
- NP-Completeness means exact solutions unlikely.

Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
- NP-Completeness means exact solutions unlikely.
- Common approximations: stack decoding, distortion limits

Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
- NP-Completeness means exact solutions unlikely.
- Common approximations: stack decoding, distortion limits
- But, these approximations have a cost: we may not find the true argmax.

Modeling Translation

- Write down your model formally, e.g.

- Choose model parameters to optimize some objective, e.g.:

$$\hat{\theta} = \arg\ \max p_\theta(data)$$

- Search for translations that optimize some decision function, e.g.:

$$\mathrm{argmax}_{English}\, p(English|Chinese)$$

la empresa tiene enemigos fuertes en Europa .

Garcia and associates .
Garcia y asociados .

Carlos Garcia has three associates .
Carlos Garcia tiene tres asociados .

his associates are not strong .
sus asociados no son fuertes .

Garcia has a company also .
Garcia tambien tiene una empresa .

its clients are angry .
sus clientes estan enfadados .

the associates are also angry .
los asociados tambien estan enfadados .

the clients and the associates are enemies .
los clientes y los asociados son enemigos .

the company has three groups .
la empresa tiene tres grupos .

its groups are in Europe .
sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .
los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .
los grupos no venden zanzanina .

the small groups are not modern .
los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

Garcia and associates .
Garcia y asociados .

Carlos Garcia has three associates .
Carlos Garcia tiene tres asociados .

his associates are not strong .
sus asociados no son fuertes .

Garcia has a company also .
Garcia tambien tiene una empresa .

its clients are angry .
sus clientes estan enfadados .

the associates are also angry .
los asociados tambien estan enfadados .

the clients and the associates are enemies .
los clientes y los asociados son enemigos .

the company has three groups .
la empresa tiene tres grupos .

its groups are in Europe .
sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .
los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .
los grupos no venden zanzanina .

the small groups are not modern .
los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

the company has three groups .

la empresa tiene tres grupos .

his associates are not strong .

its groups are in Europe .

sus

pa .

Ga

maceuticals .

Garcia

cinas fuertes .

nine .

sus

nina .

the

dern .

los asociados tambien estan enfadados .

los grupos pequenos no son modernos .

Warning: these are not phrases in any linguistic sense.

la empresa tiene enemigos fuertes en Europa .

the company has strong enemies in Europe .

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus

Ga

Garcia

sus

the

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

pa .

maceuticals .

cinas fuertes .

nine .

nina .

dern .

los grupos pequenos no son modernos .

Warning: these are not phrases in any linguistic sense.

Restriction to linguistic phrases seems to hurt (Koehn et al. 2003)

Phrase-based Models

Although north wind howls , but sky still very clear .

虽然 北 风 呼啸 ，但 天空 依然 十分 清澈 。

Phrase-based Models

Although north wind howls , but sky still very clear .

虽然	北 风 呼啸	， 但	天空 依然 十分 清澈	。

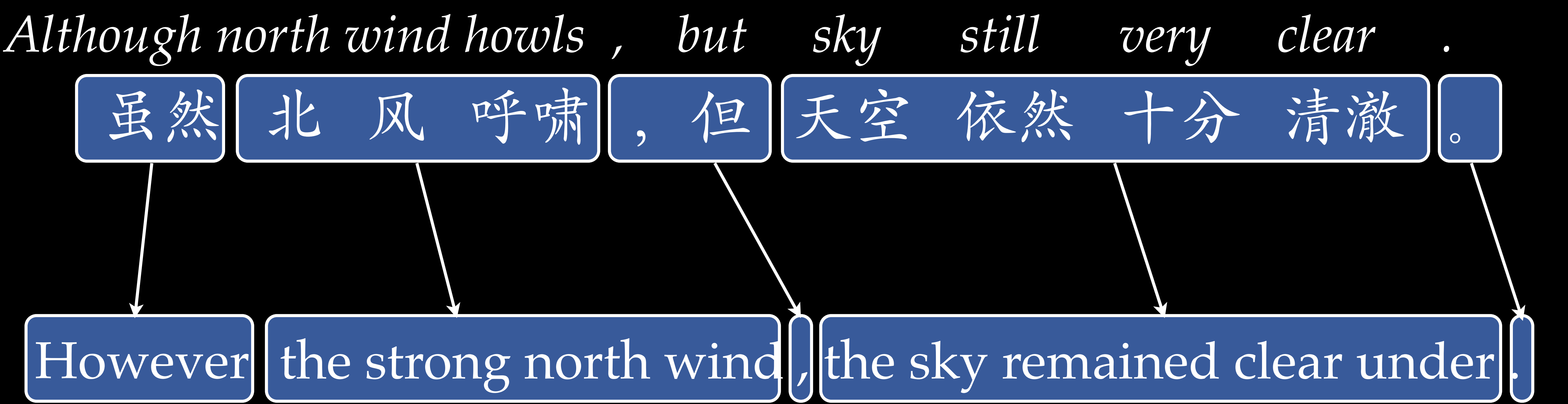
Phrase-based Models

Although north wind howls , but sky still very clear .

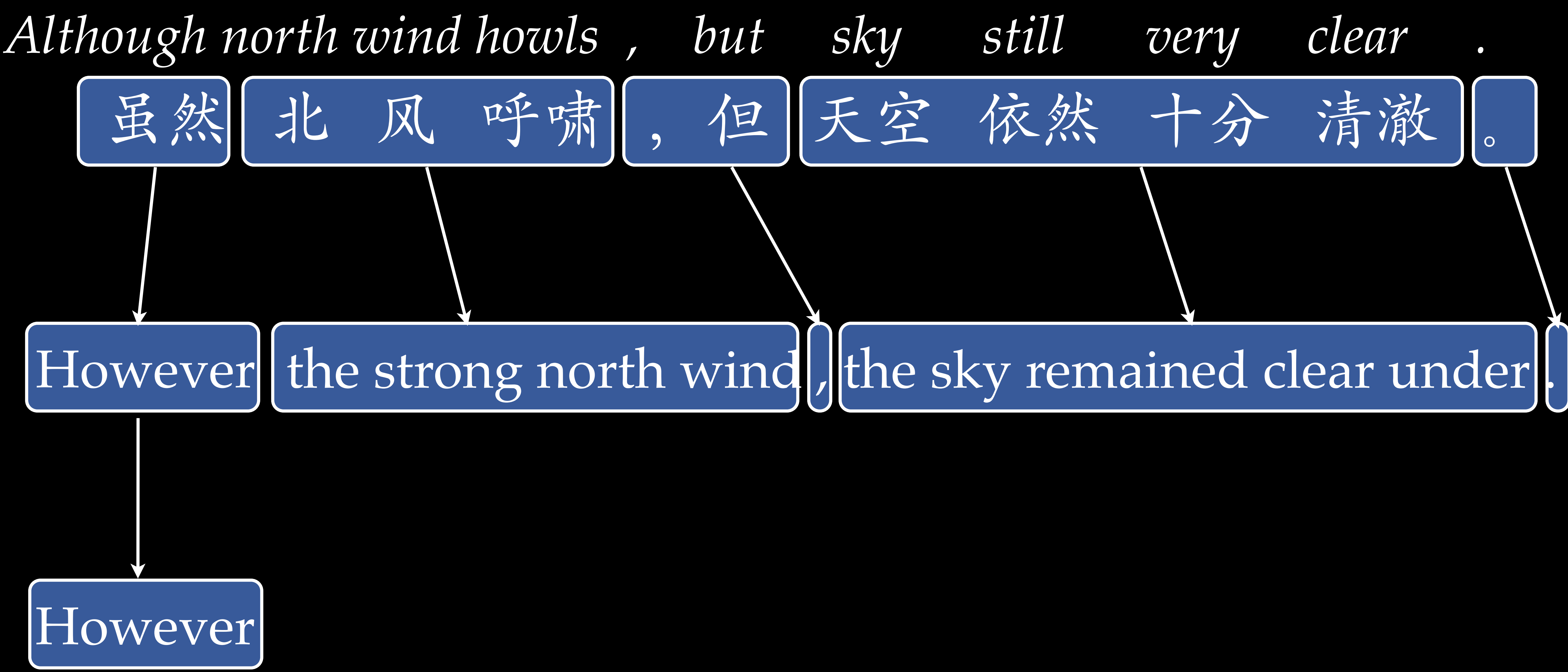
虽然 | 北 风 呼啸 | ，但 | 天空 依然 十分 清澈 | 。

However

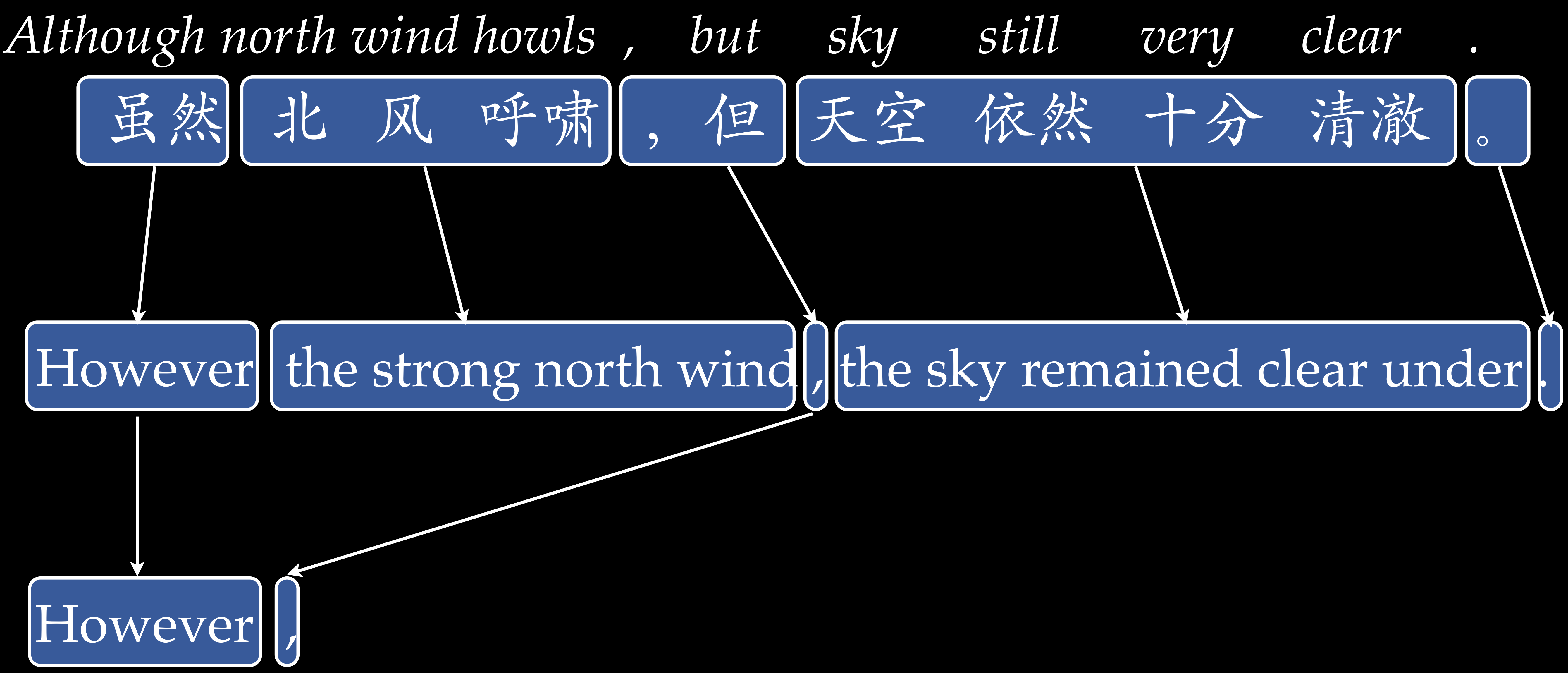
Phrase-based Models



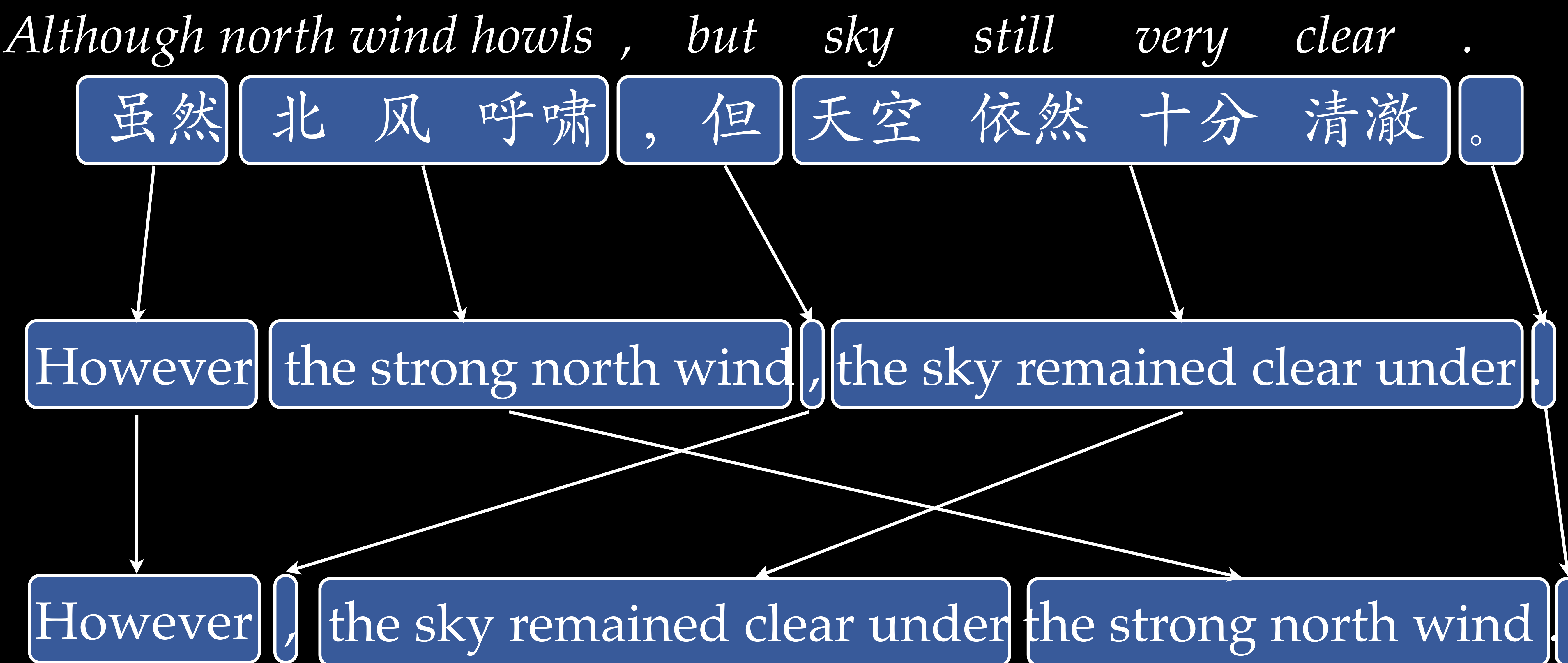
Phrase-based Models



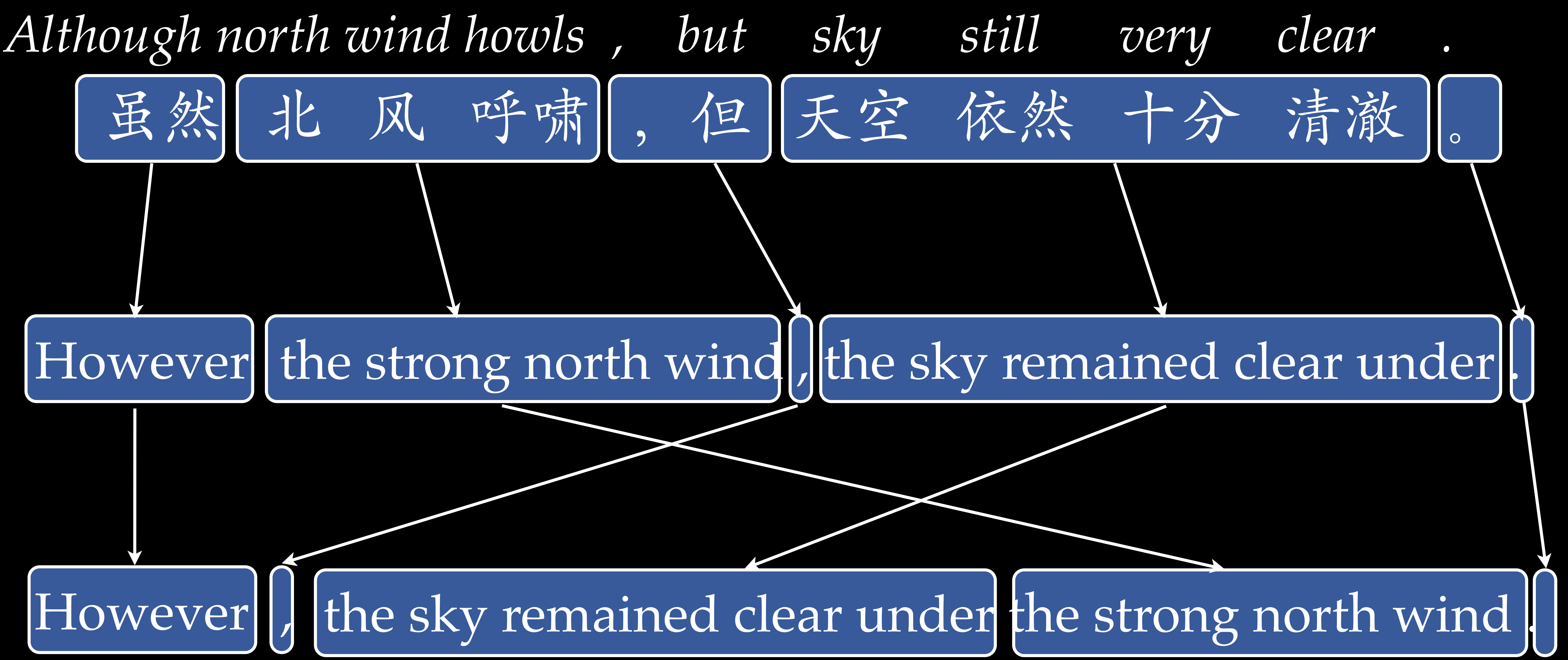
Phrase-based Models



Phrase-based Models



Phrase-based Models



$$p(English, alignment|Chinese) =$$

$$p(segmentation) \cdot p(translations) \cdot p(reorderings)$$

Phrase-based Models

Phrase-based Models

- Segmentation probabilities.

Phrase-based Models

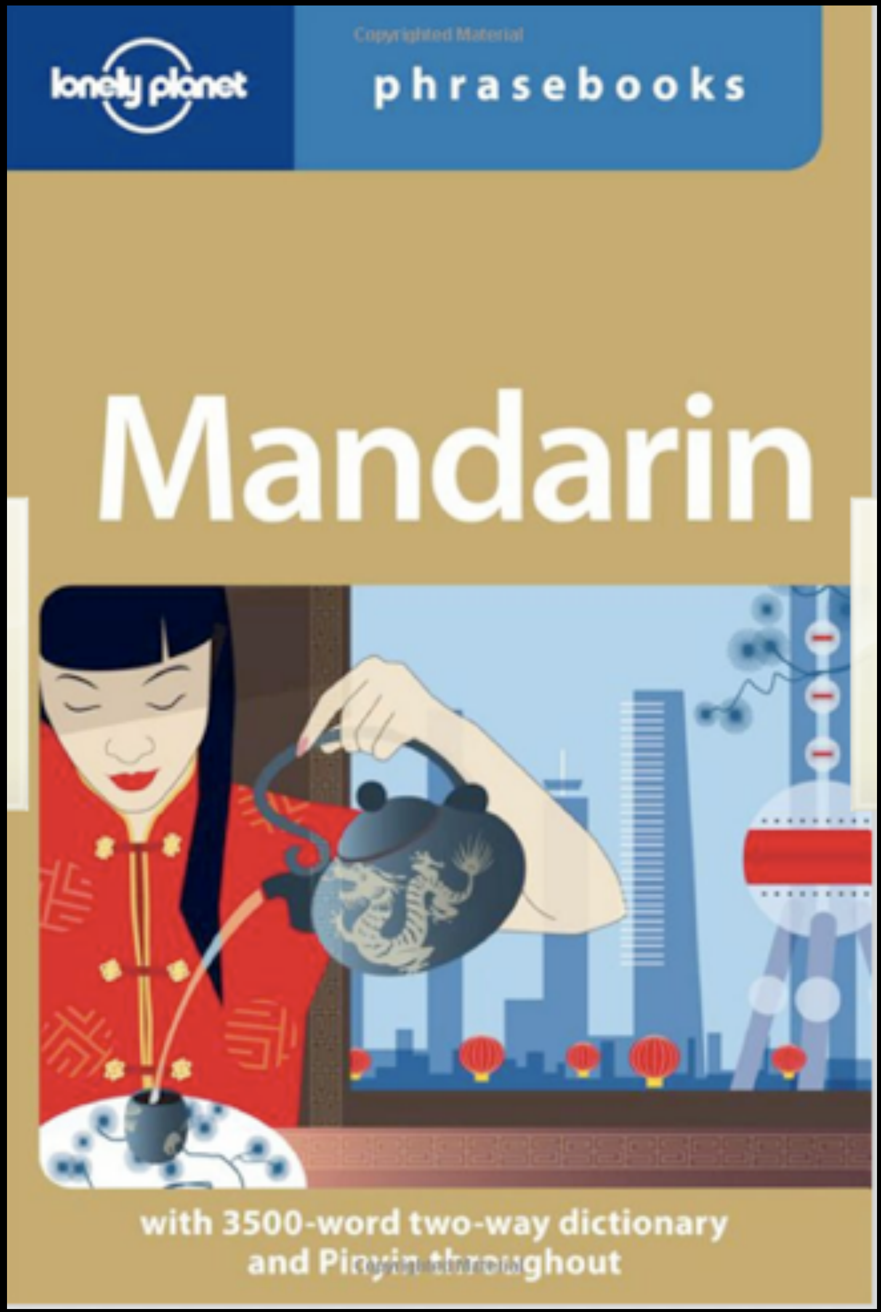
- Segmentation probabilities.
- Phrase translation probabilities.

Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.

Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.



Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.
- Some problems:
  - Weak reordering model -- output is not fluent.
  - ~~Many decisions -- many things can go wrong.~~

Learning

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.
- Guaranteed that likelihood is monotonically nondecreasing.

Learning

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, using expected counts as proxy for observed counts.
- Iterate.
- Guaranteed that likelihood is monotonically nondecreasing.

Learning

- Arbitrarily select a set of parameters (say, uniform).
- Calculate ***expected counts*** of the unseen events.
- Choose new parameters to maximize likelihood, us nts.
- It
- G
  n

Computing expectations from a phrase-based model, given a sentence pair, is #P-Complete (by reduction to counting perfect matchings; DeNero & Klein, 2008)

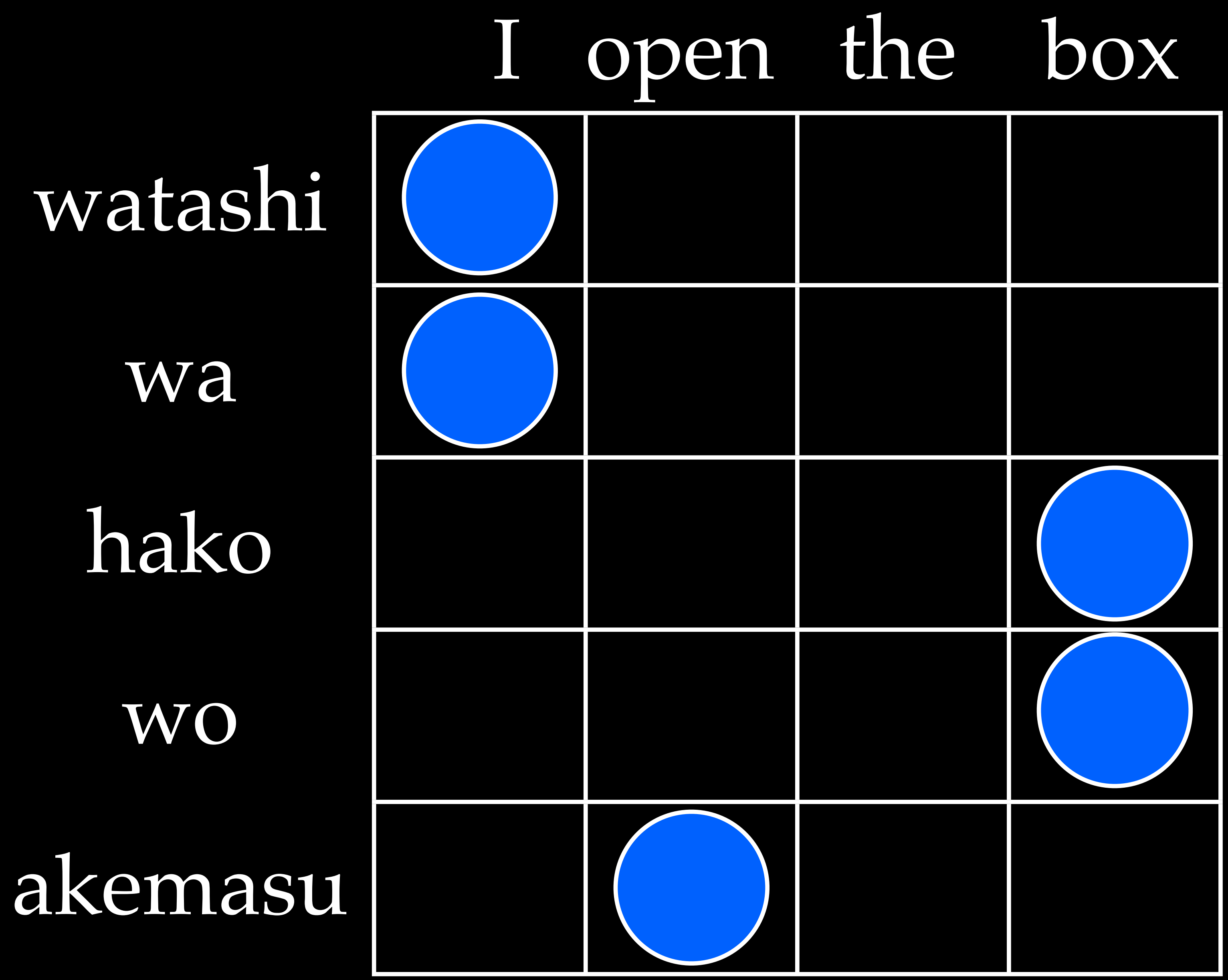
Now What?

- Option #1: approximate expectations
  - Restrict computation to some tractable subset of the alignment space (arbitrarily biased).
  - Markov chain Monte Carlo (very slow).

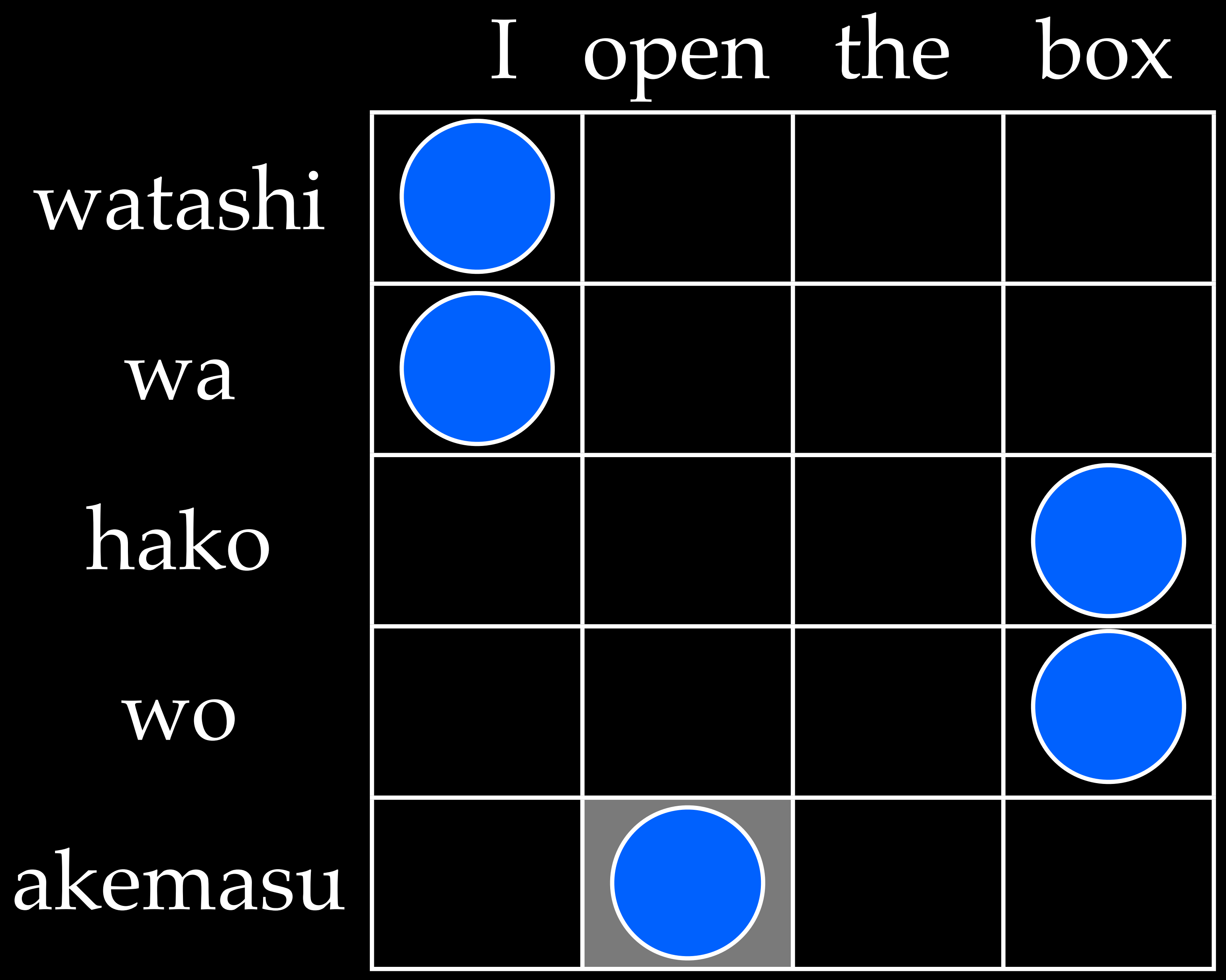
Now What?

- Change the problem definition
  - We already know how to learn word-to-word translation models efficiently.
  - Idea: learn word-to-word alignments, extract most probable alignment, then treat it as observed.
  - Learn phrase translations consistent with word alignments.
  - Decouples alignment from model learning -- is this a good thing?

Phrase Extraction

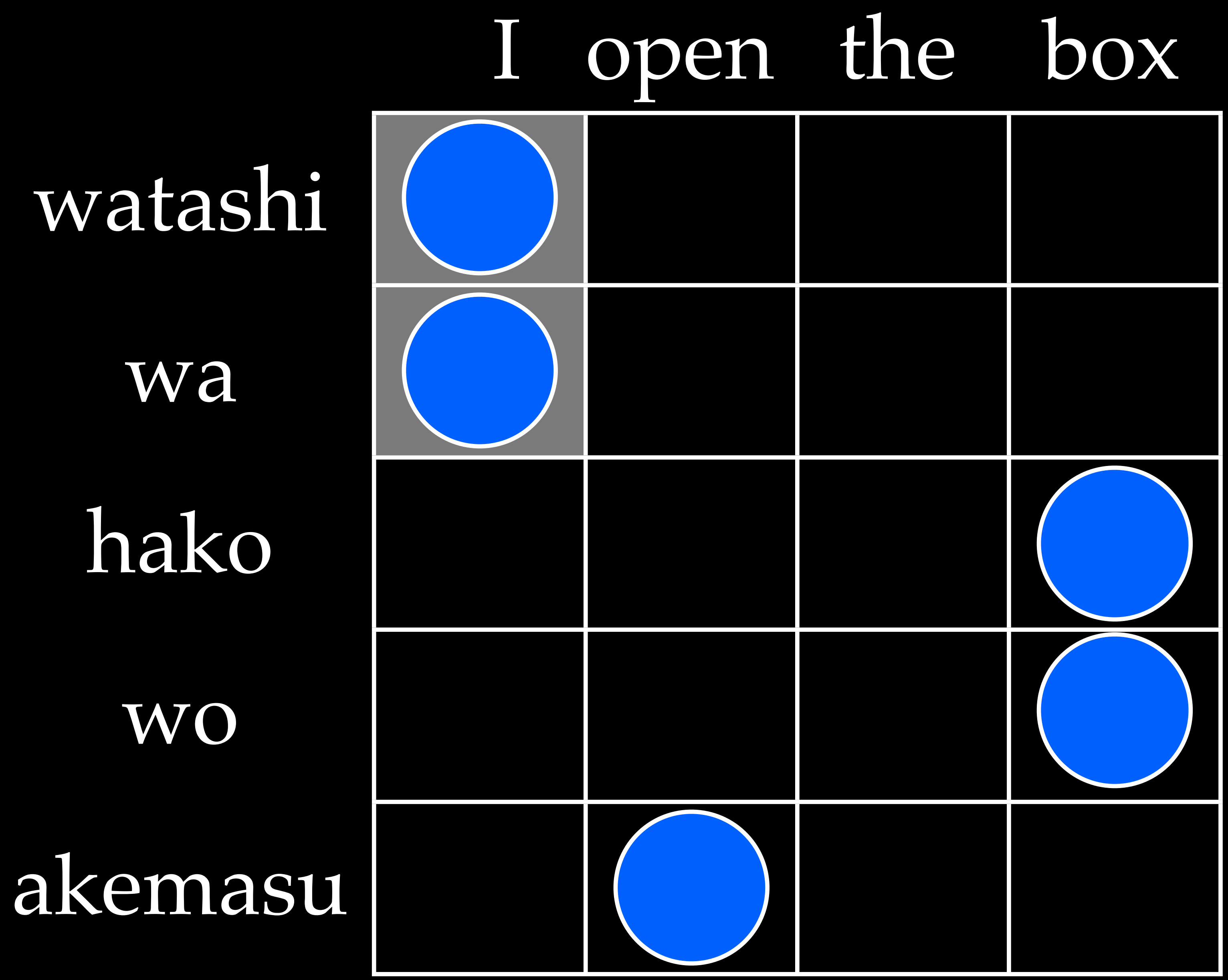


Phrase Extraction



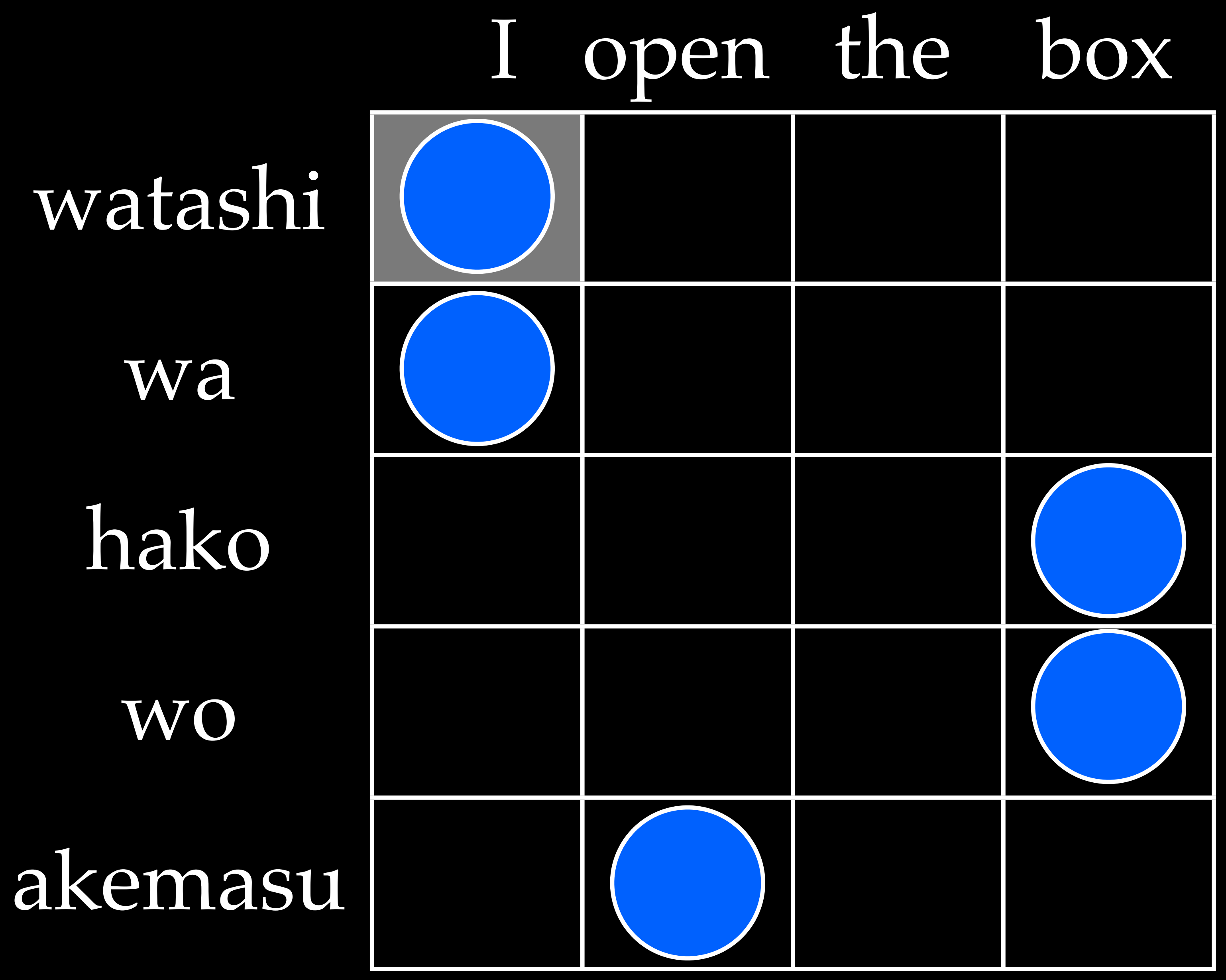
akemasu / open

Phrase Extraction



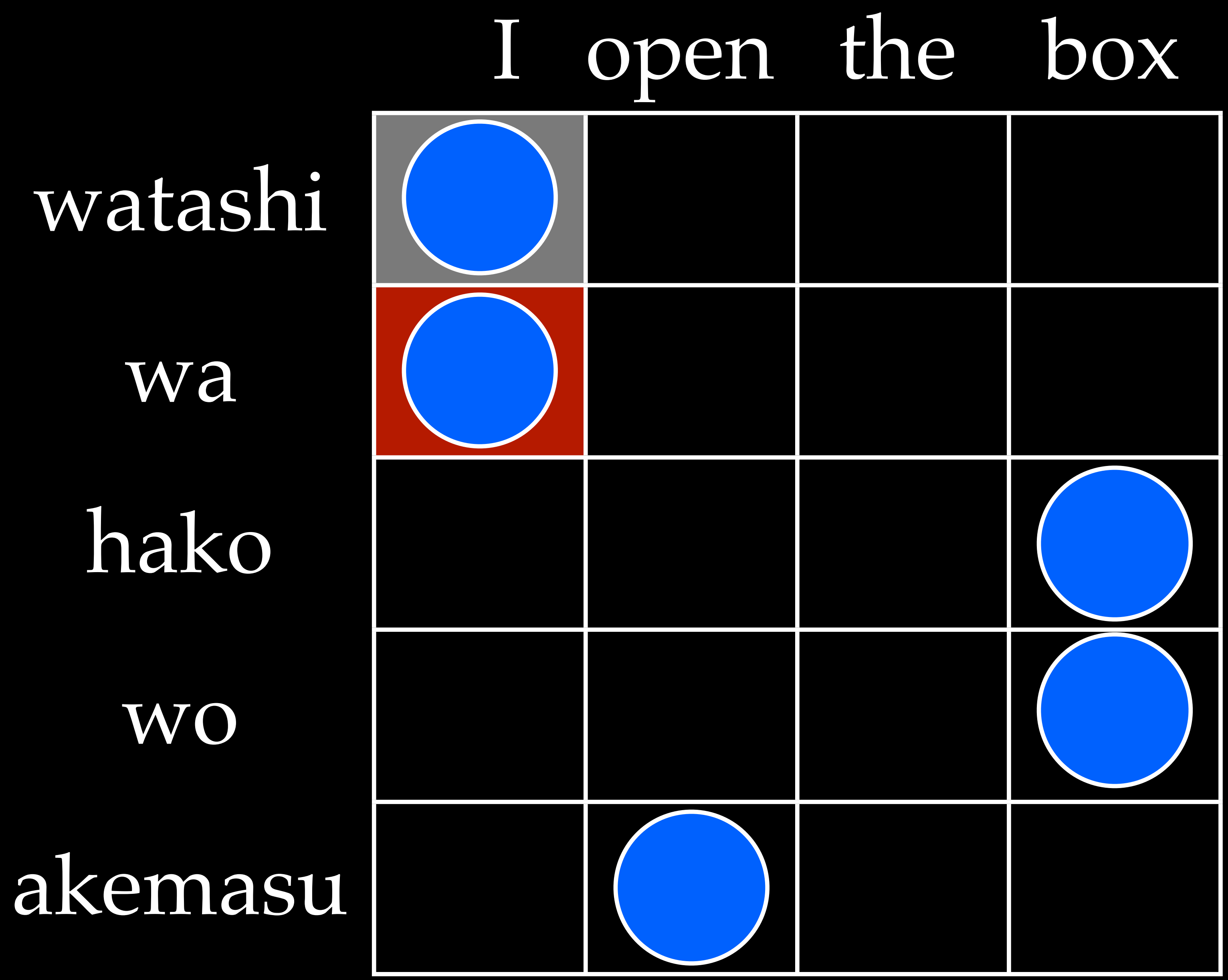
watashi wa / I

Phrase Extraction



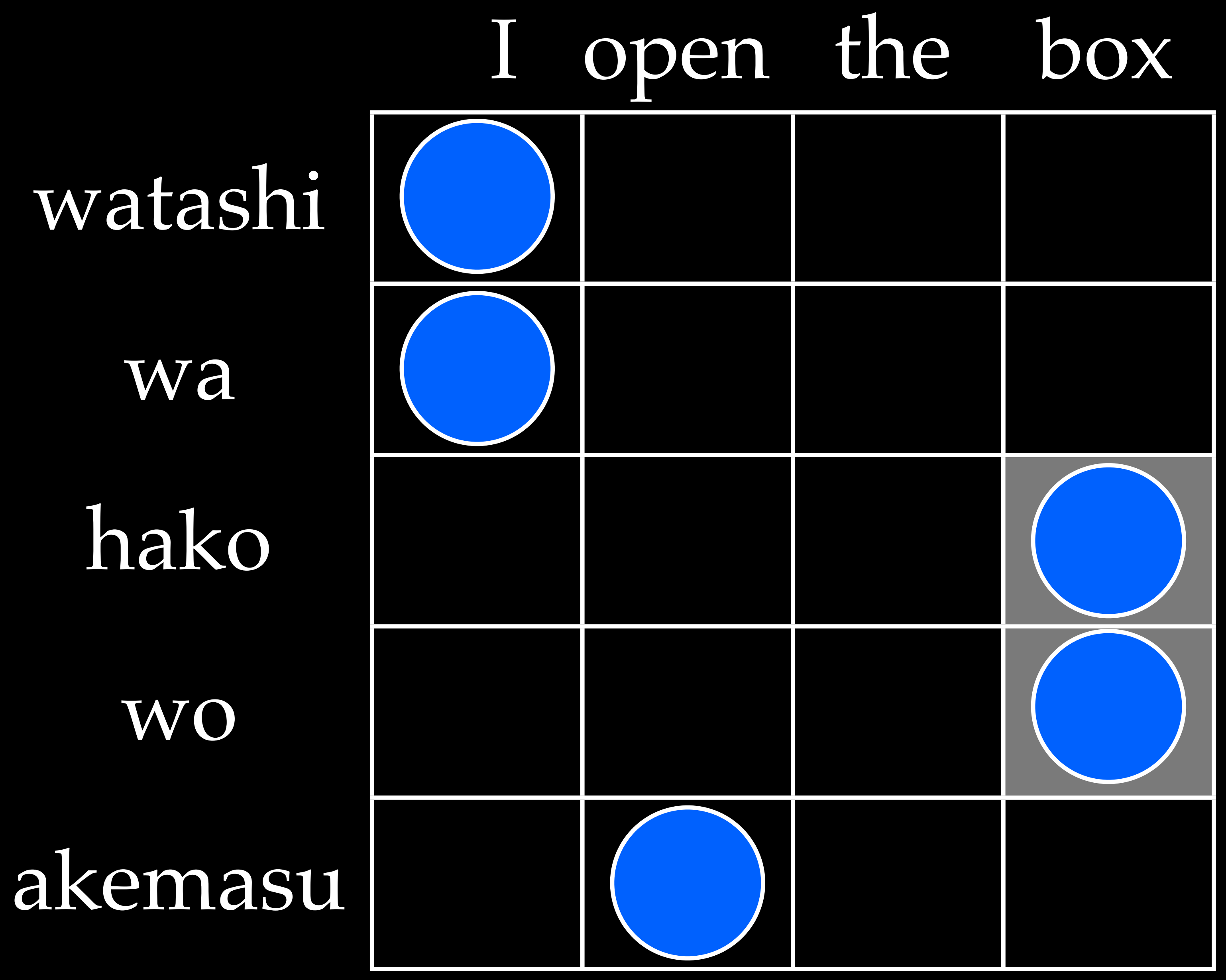
watashi / I

Phrase Extraction



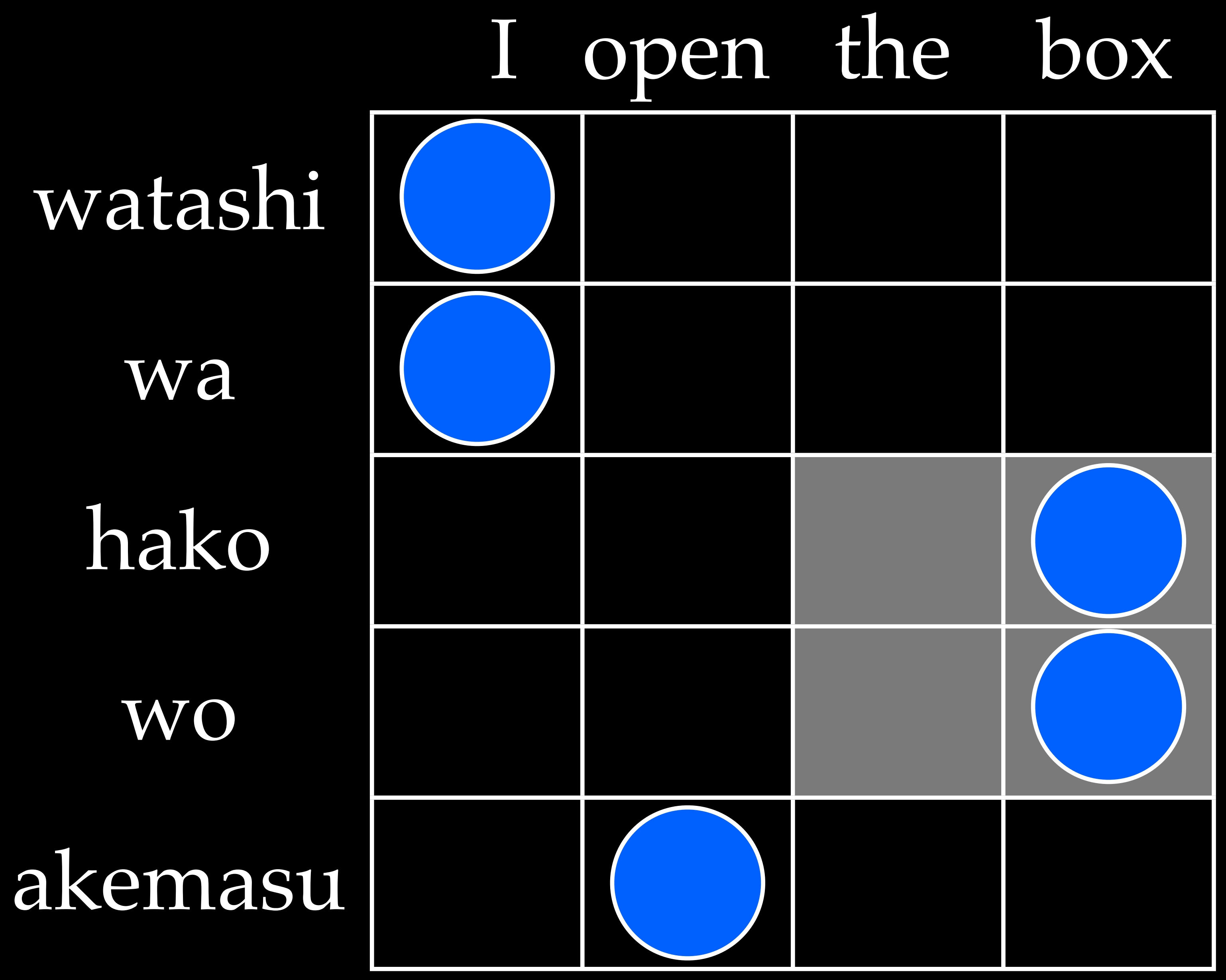
watashi wa / I

Phrase Extraction



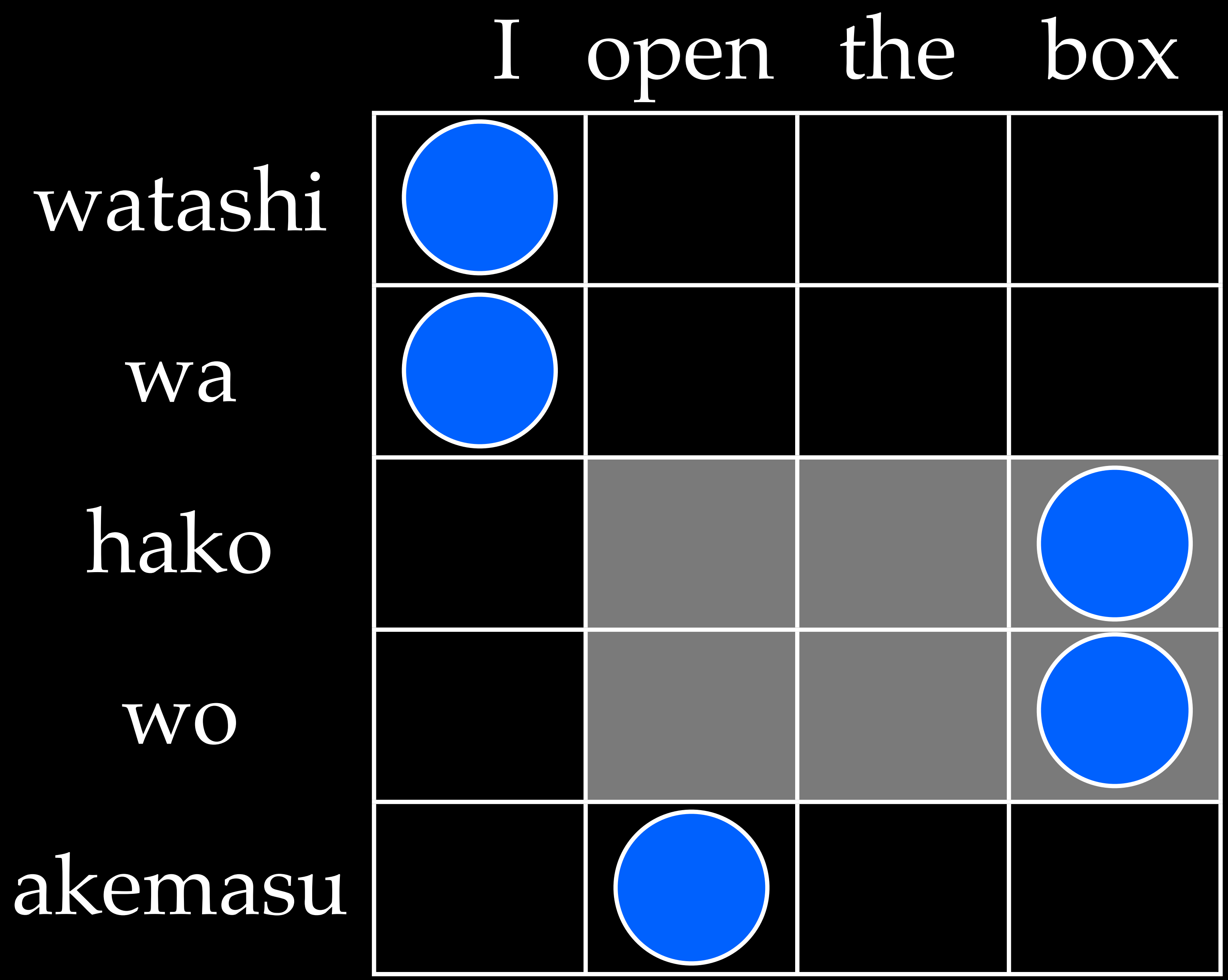
hako wo / box

Phrase Extraction



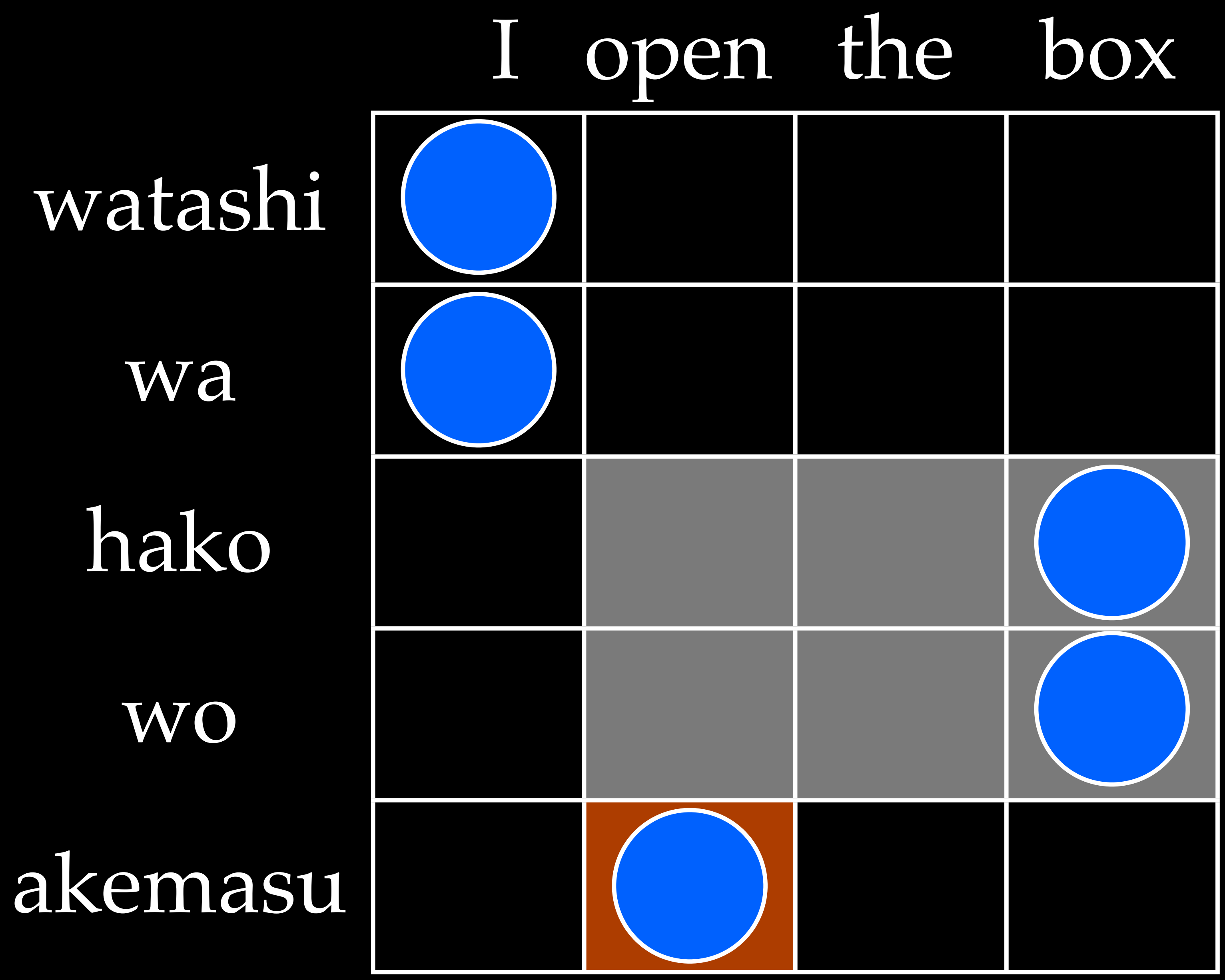
hako wo / the box

Phrase Extraction



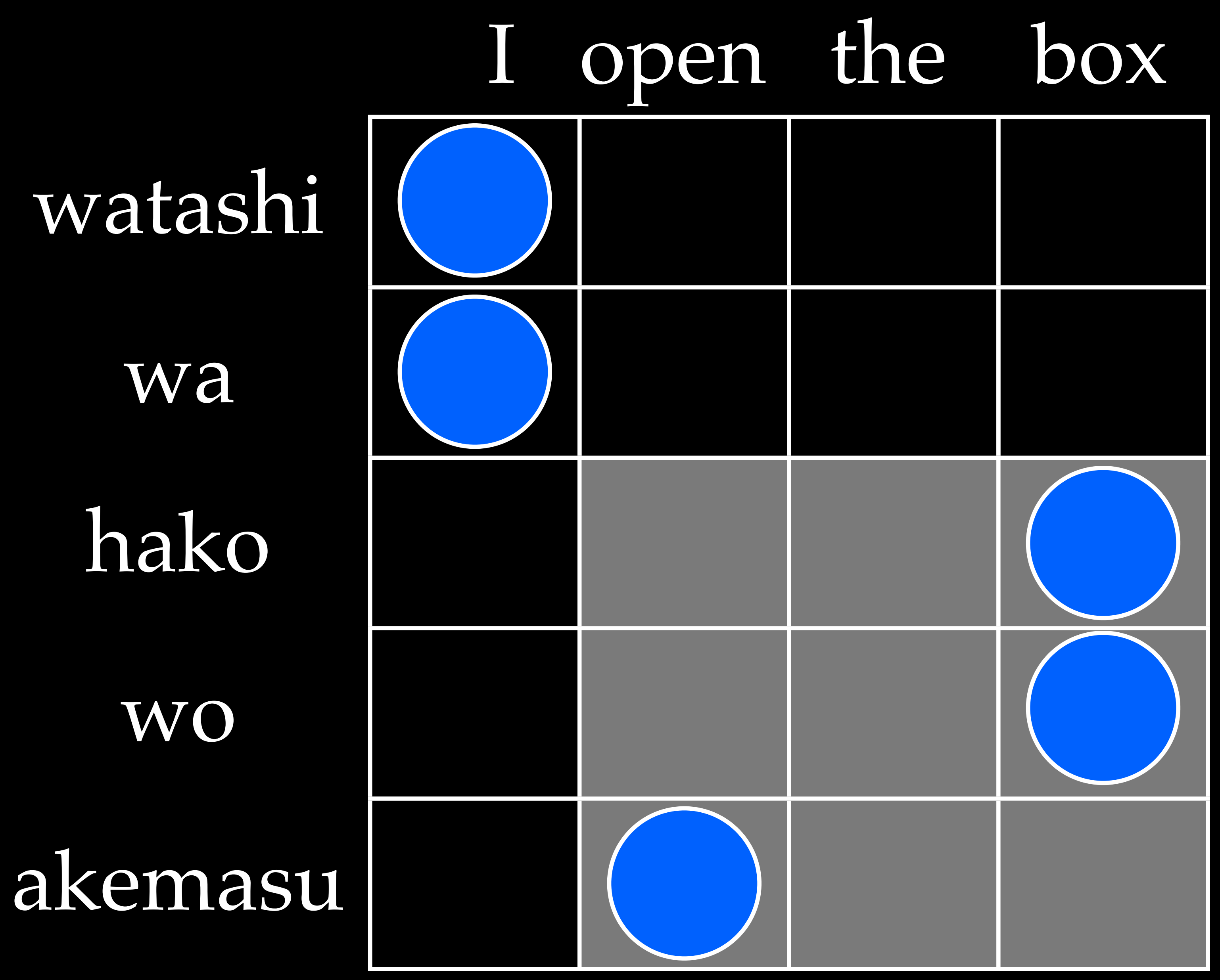
hako wo / open the box

Phrase Extraction



hako wo / open the box

Phrase Extraction



hako wo akemasu / open the box

Phrasal Translation Estimation

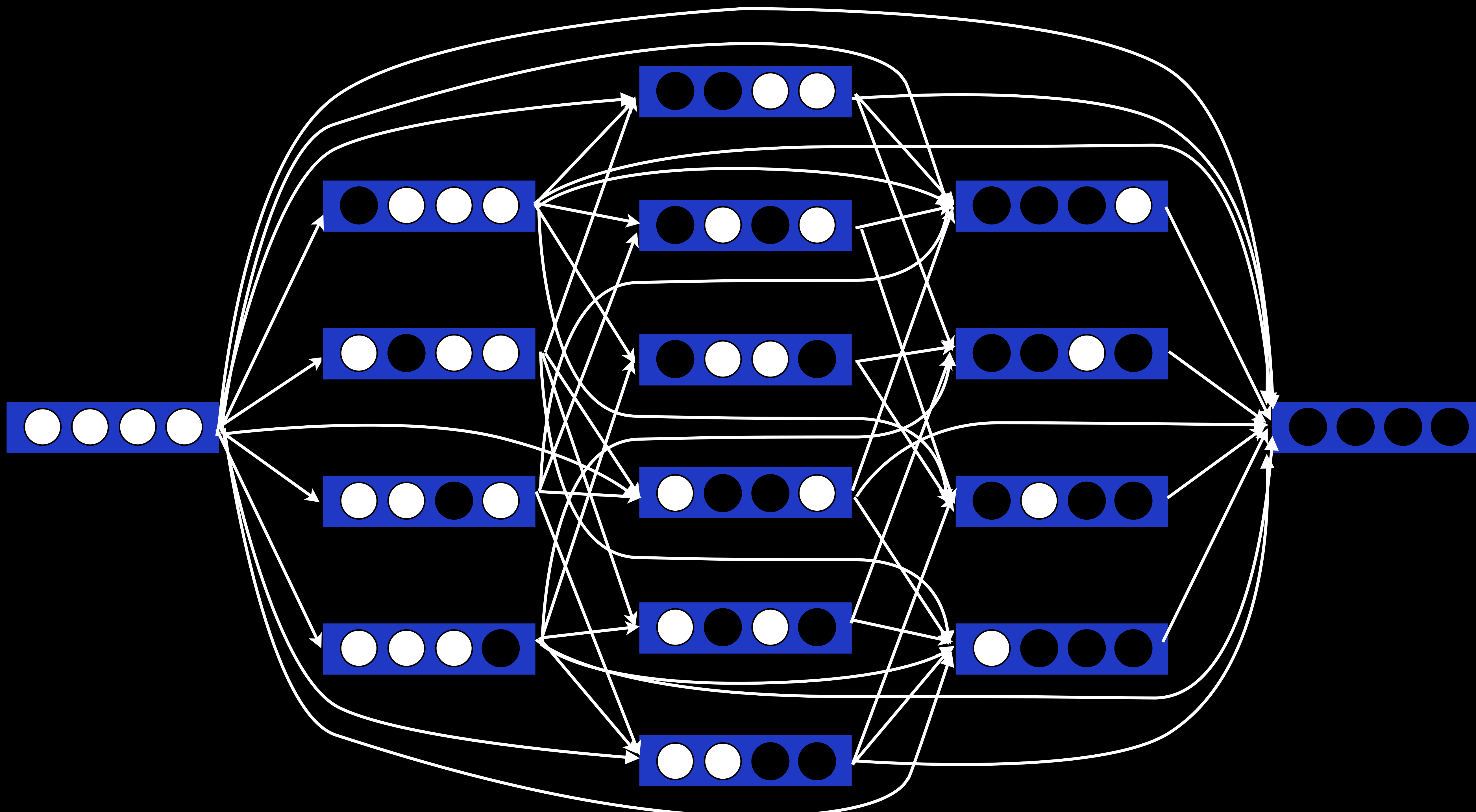
Phrasal Translation Estimation

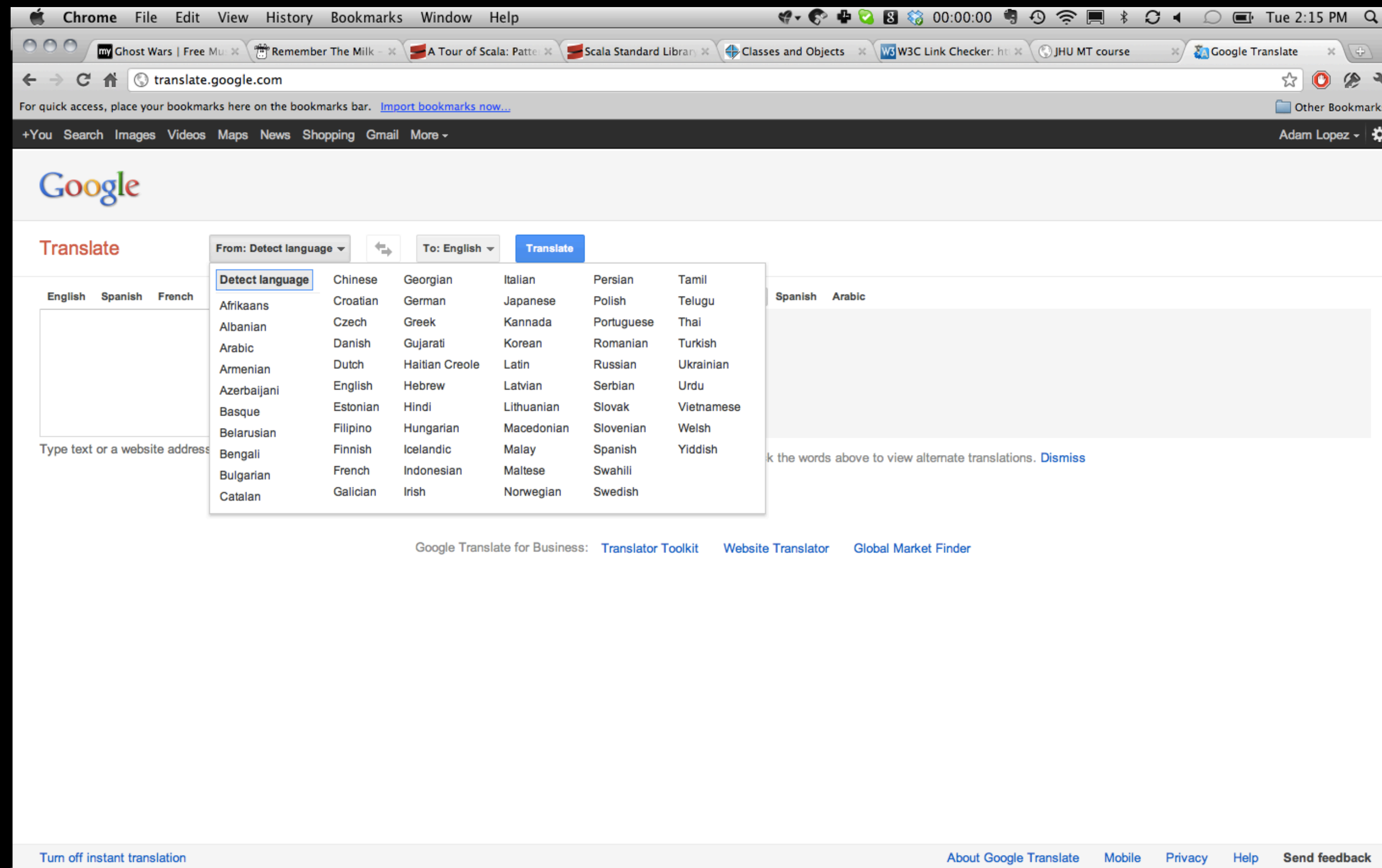
- Option #1 (EM over restricted space)
  - Align with a word-based model.
  - Compute expectations only over alignments consistent with the alignment grid.

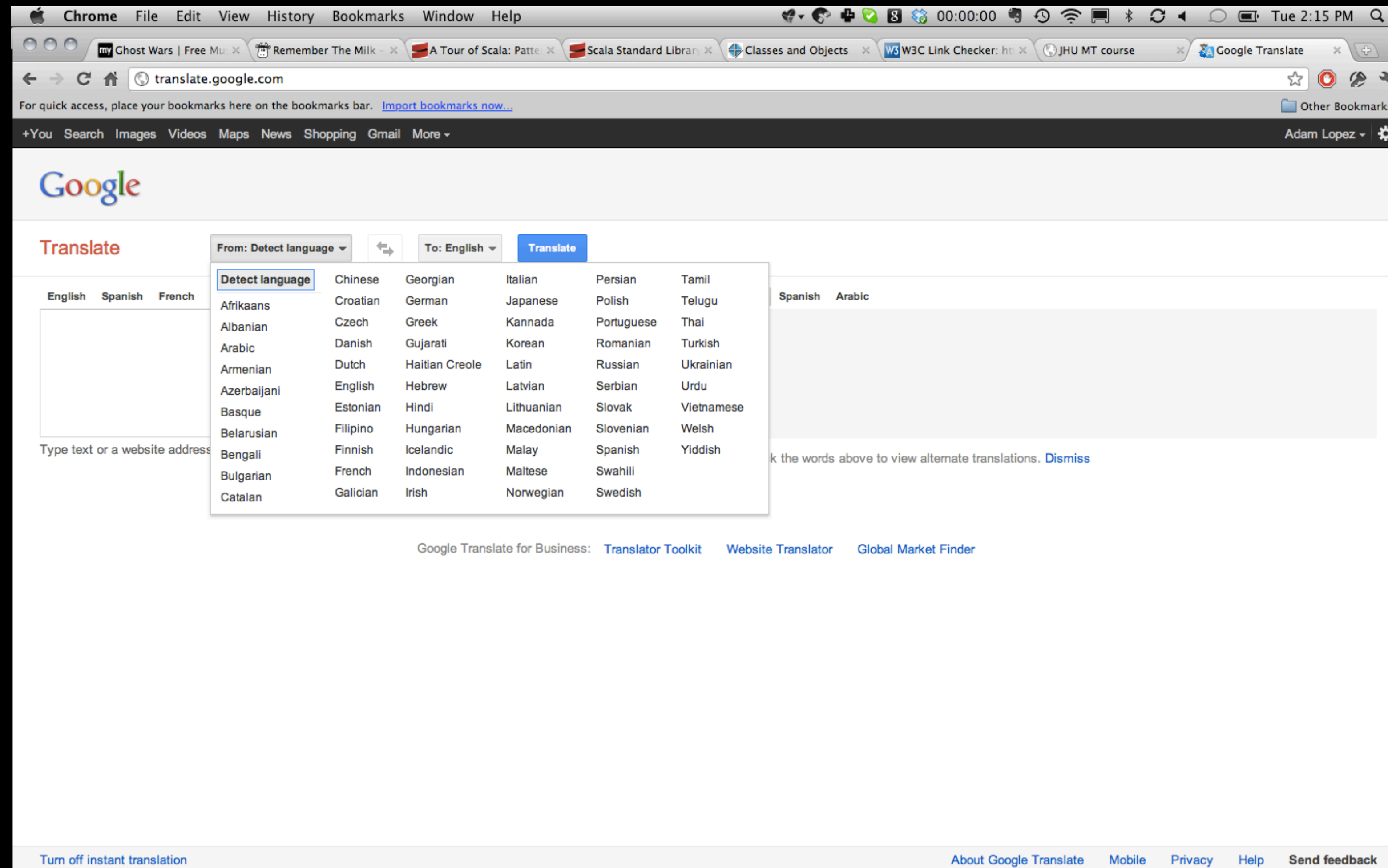
Phrasal Translation Estimation

- Option #1 (EM over restricted space)
  - Align with a word-based model.
  - Compute expectations only over alignments consistent with the alignment grid.
- Option #2 (Non-global estimation)
  - View phrase pairs as observed, irrespective of context or overlap.

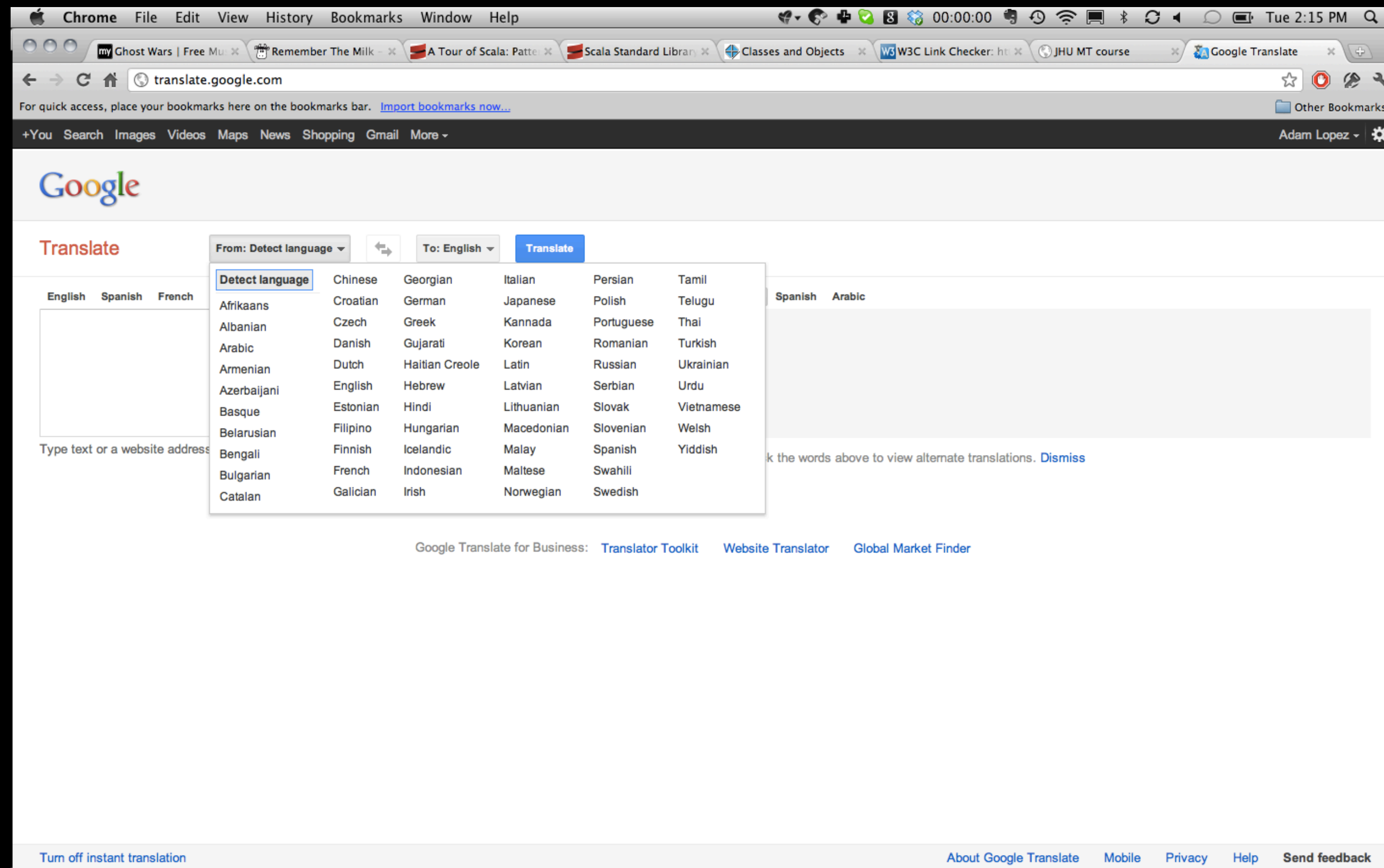
Search



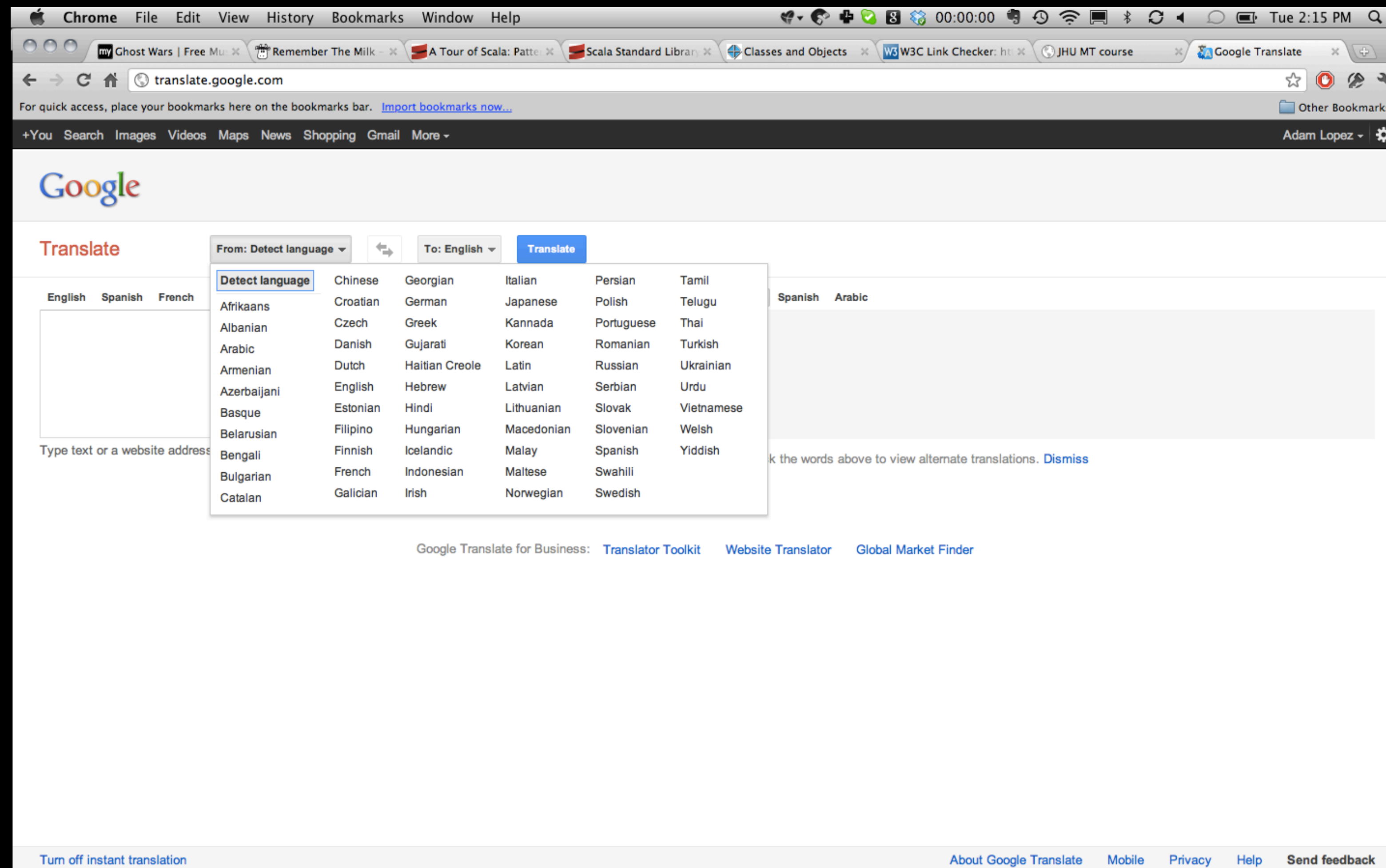

Chrome File Edit View History Bookmarks Window Help
Tue 2:15 PM
Ghost Wars | Free Mu
Remember The Milk
A Tour of Scala: Patte
Scala Standard Librar
Classes and Objects
W3C Link Checker: ht
JHU MT course
Google Translate
translate.google.com
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Other Bookmarks
+You Search Images Videos Maps News Shopping Gmail More
Adam Lopez
Google
Translate
From: Detect language
To: English
Translate
English Spanish French
Spanish Arabic
Detect language
Afrikaans
Albanian
Arabic
Armenian
Azerbaijani
Basque
Belarusian
Bengali
Bulgarian
Catalan
Chinese
Croatian
Czech
Danish
Dutch
English
Estonian
Filipino
Finnish
French
Galician
Georgian
German
Greek
Gujarati
Haitian Creole
Hebrew
Hindi
Hungarian
Icelandic
Indonesian
Irish
Italian
Japanese
Kannada
Korean
Latin
Latvian
Lithuanian
Macedonian
Malay
Maltese
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Slovak
Slovenian
Spanish
Swahili
Swedish
Tamil
Telugu
Thai
Turkish
Ukrainian
Urdu
Vietnamese
Welsh
Yiddish
Type text or a website address
k the words above to view alternate translations. Dismiss
Google Translate for Business: Translator Toolkit Website Translator Global Market Finder
Turn off instant translation
About Google Translate Mobile Privacy Help Send feedback



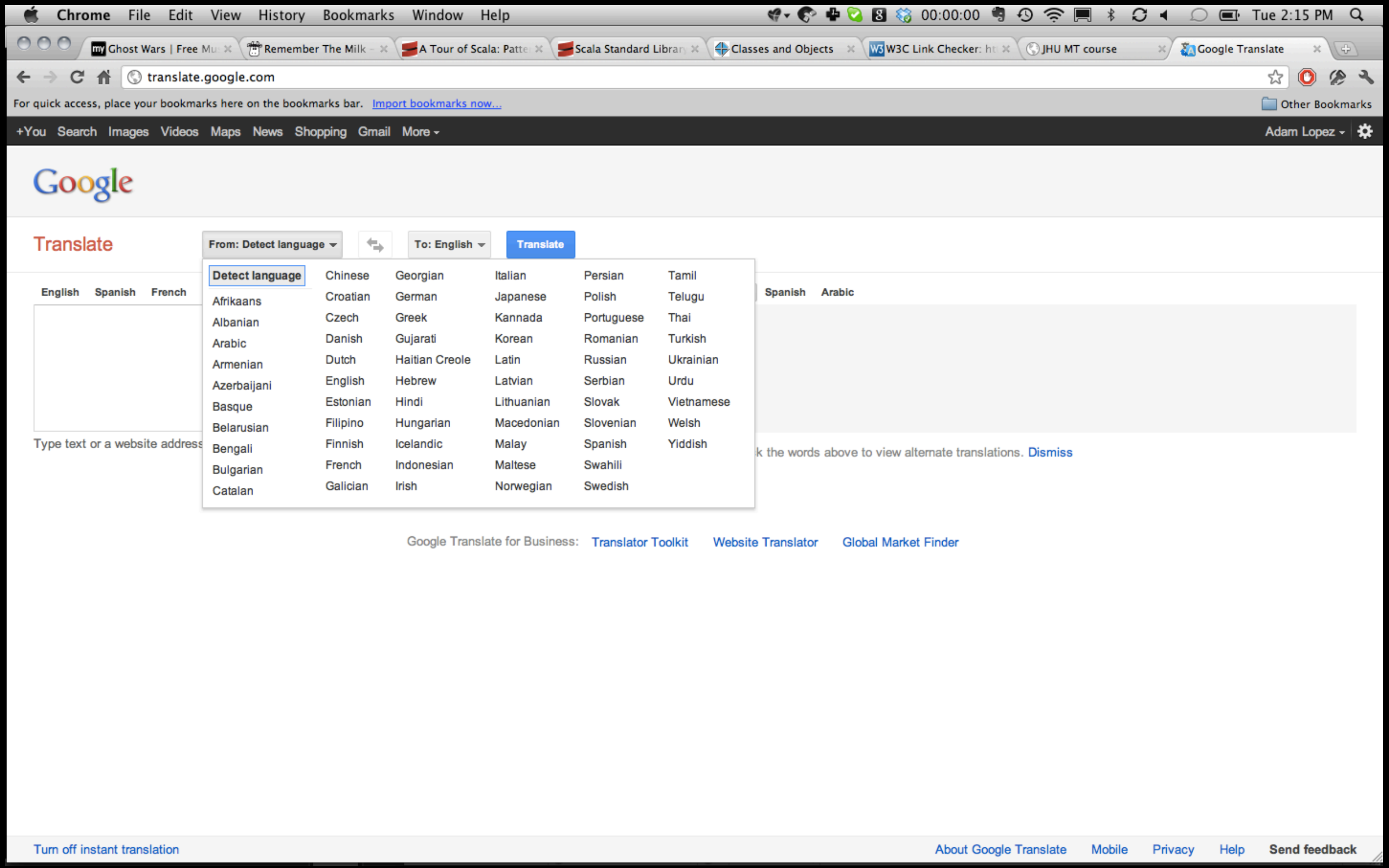
- Some (not all) key ingredients in Google Translate:



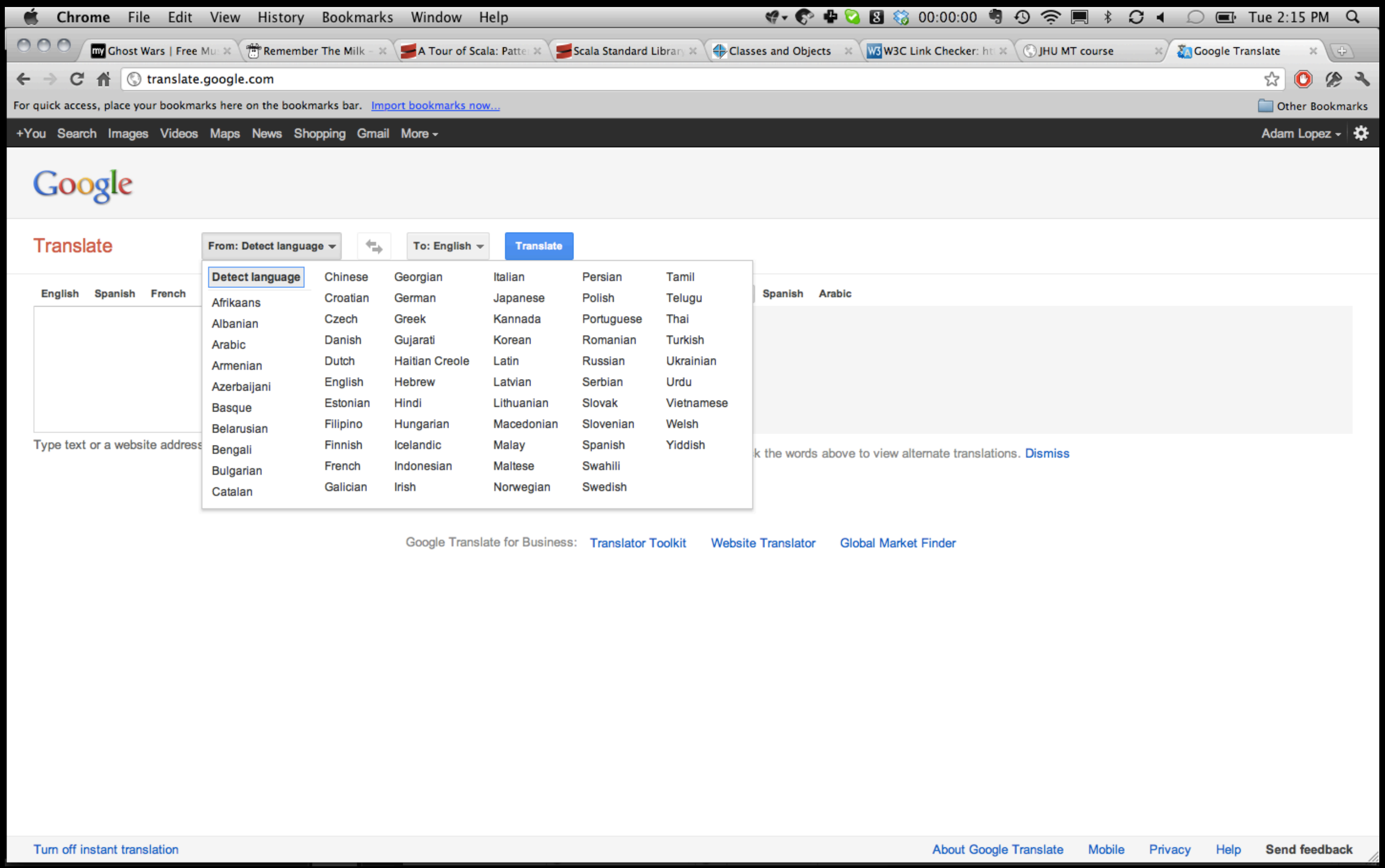
- Some (not all) key ingredients in Google Translate:
  - Phrase-based translation models



- Some (not all) key ingredients in Google Translate:
  - Phrase-based translation models
  - ... Learned heuristically from word alignments



- Some (not all) key ingredients in Google Translate:
  - Phrase-based translation models
  - ... Learned heuristically from word alignments
  - ... Coupled with a huge language model



- Some (not all) key ingredients in Google Translate:
  - Phrase-based translation models
  - ... Learned heuristically from word alignments
  - ... Coupled with a huge language model
  - ... And very tight pruning heuristics

Phrase-based Models

Phrase-based Models

- Phrase-based Models are dumb.

Phrase-based Models

- Phrase-based Models are dumb.
- But they are widely regarded as state-of-the-art.

Phrase-based Models

- Phrase-based Models are dumb.
- But they are widely regarded as state-of-the-art.
- Why? Simple models are easier to learn and deploy.

Phrase-based Models

- Phrase-based Models are dumb.
- But they are widely regarded as state-of-the-art.
- Why? Simple models are easier to learn and deploy.
- Need proof? Google uses a phrase-based model.

Implementations

- Phrase-based Translation
  - Moses -- www.statmt.org/moses/
  - cdec -- www.cdec-decoder.org
- Language models
  - KenLM -- http://kheafield.com/code/kenlm/
  - SRI-LM -- www.speech.sri.com/projects/srilm/

Recap: Finite-State Models

Recap: Finite-State Models

- Probability theory enables us to learn from data.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There's no data like more data.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
- *All* of these models are weighted regular languages.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There's no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
- *All* of these models are weighted regular languages.
- Need dynamic programming with approximations.

Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There's no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
- *All* of these models are weighted regular languages.
- Need dynamic programming with approximations.
- Is this the best we can do?