Neural Networks for Machine Translation

David Chiang
University of Southern California / University of Notre Dame

Joint work with
Victoria Fossum
Ashish Vaswani
Yinggong Zhao

Outline

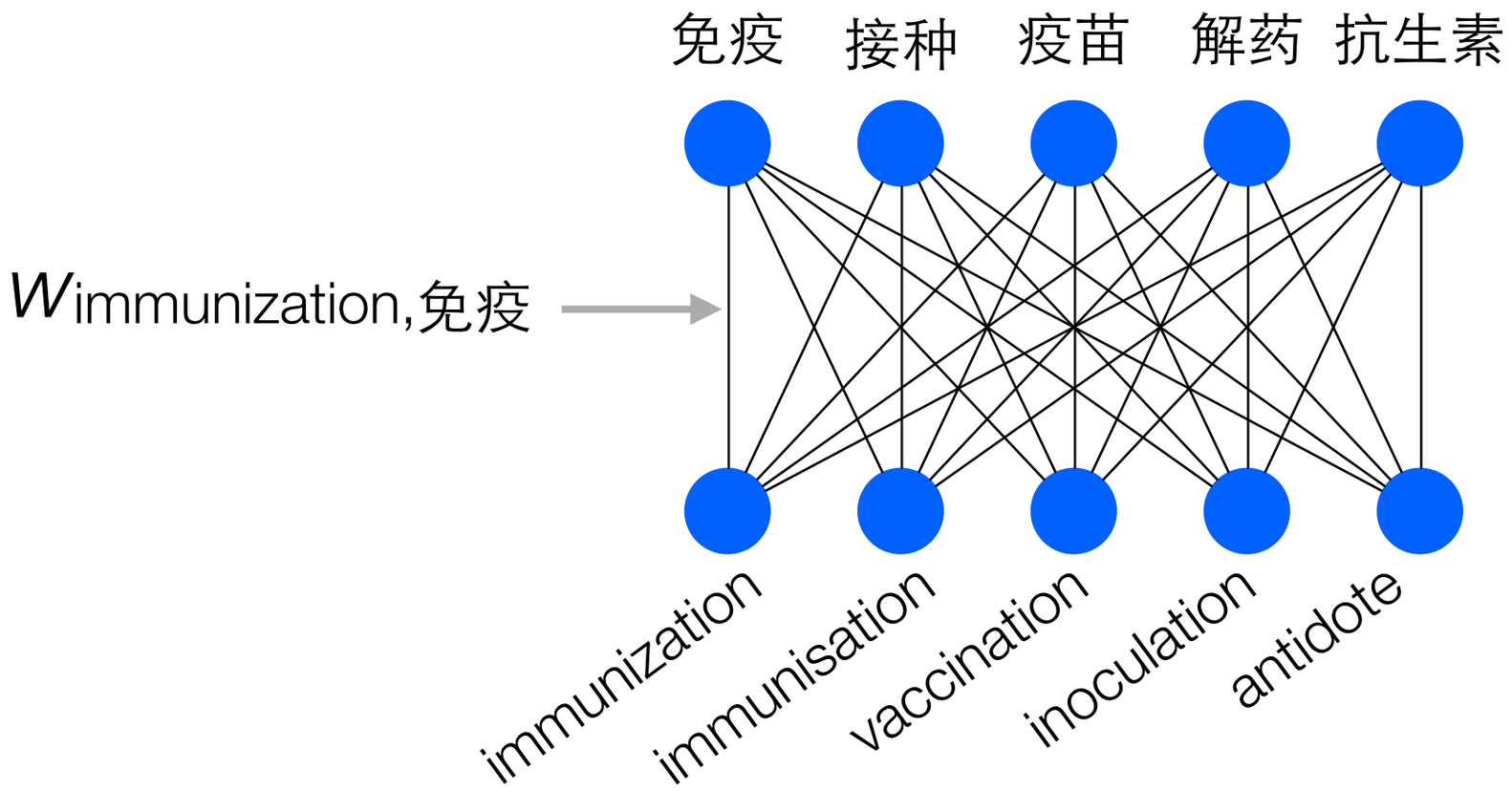
Introduction to neural networks

Fast training of neural networks

Neural networks for machine translation

Introduction to neural networks

Log-linear model



$$P(c|e) \propto \exp w_{e,c}$$

Word translation

English-to-Chinese

	免疫 miǎnyì	接种 jiēzhòng	疫苗 yìmiáo
immunization	0.04	0.1	0.04
immunisation	0.1	0.06	0.06
vaccination	0.06	0.2	0.2
inoculation	0	0.1	0.1

Features

h(*e*) = 1 iff…

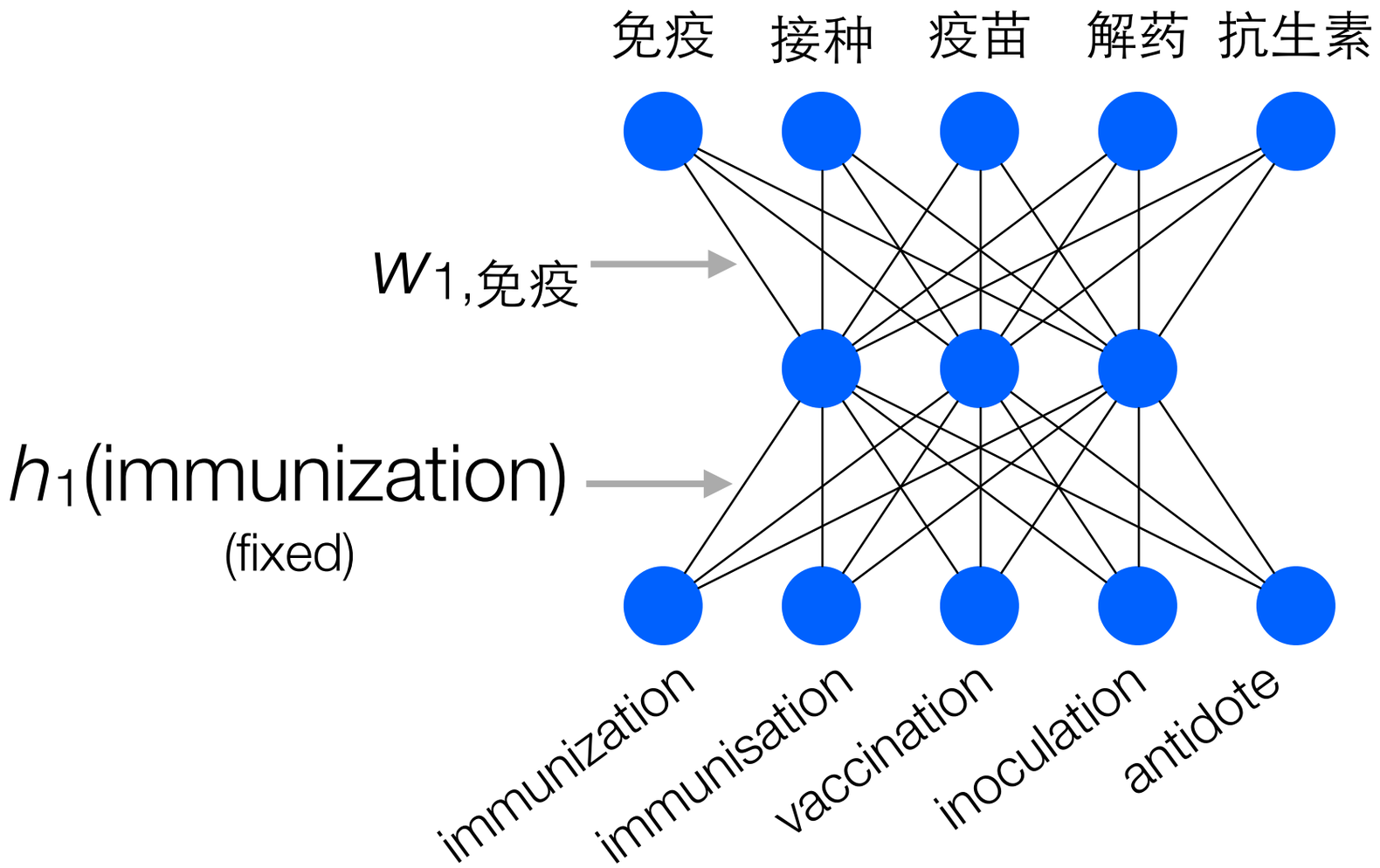
e = immunization

pos(*e*) = NN

e[:5] = immun-

|*e*| < 5

Log-linear model



$$P(c|e) \propto \exp \Sigma_i \, w_{i,c} \, h_i(e)$$

Conjunctions of features

h(*e*) = 1 iff…

e = immunization

pos(*e*) = NN

e[:5] = immun-

|*e*| < *5*

e = immunization and pos(*e*) = NN

e = immunization and *e*[–7:] = -ization

e = immunization and |*e*| < 5

pos(*e*) = NN and *e*[–7:] = -ization

pos(*e*) = NN and |*e*| < *5*

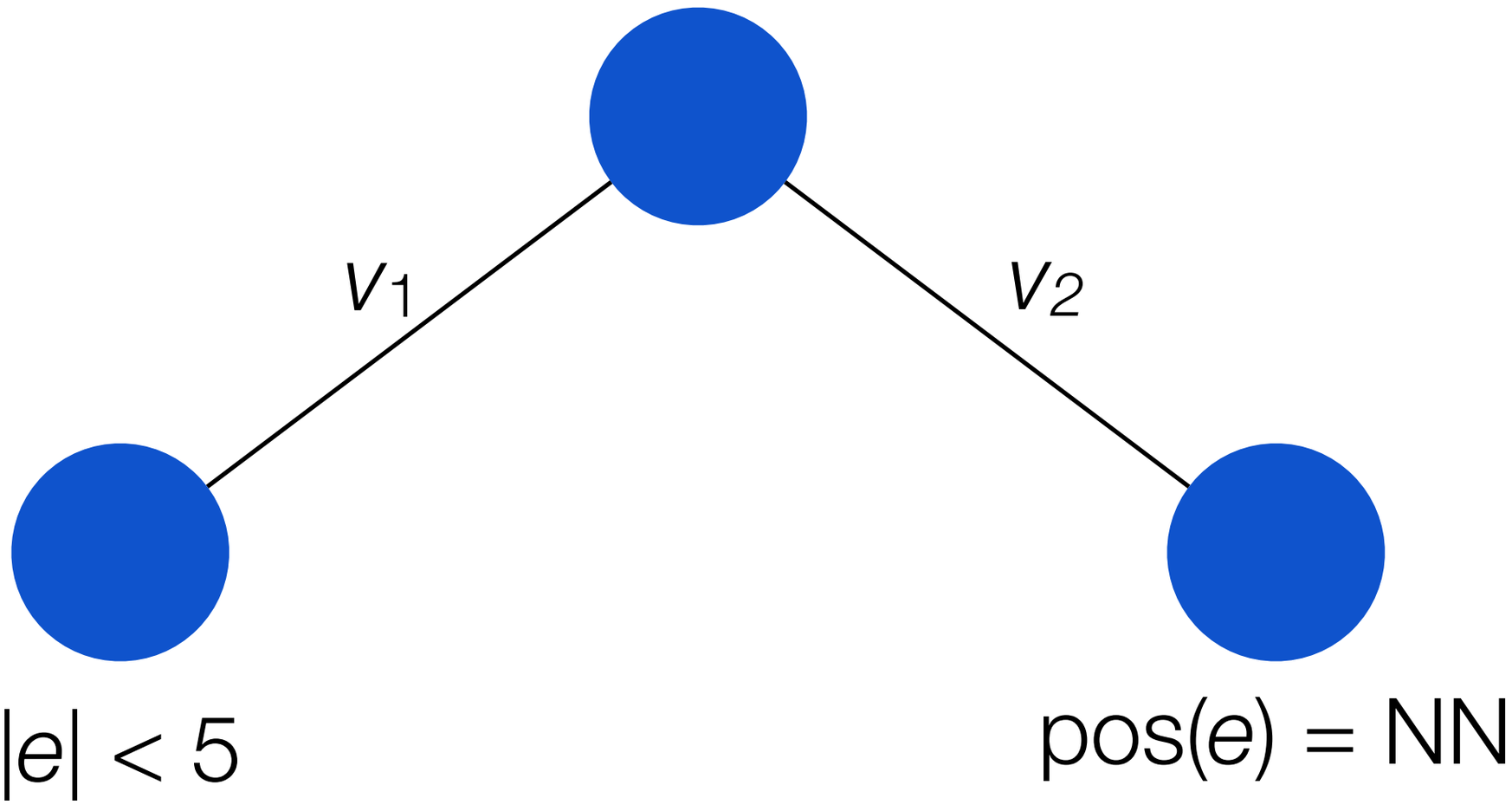
e[–7:] = -ization and |*e*| < *5*

“Programmable” features

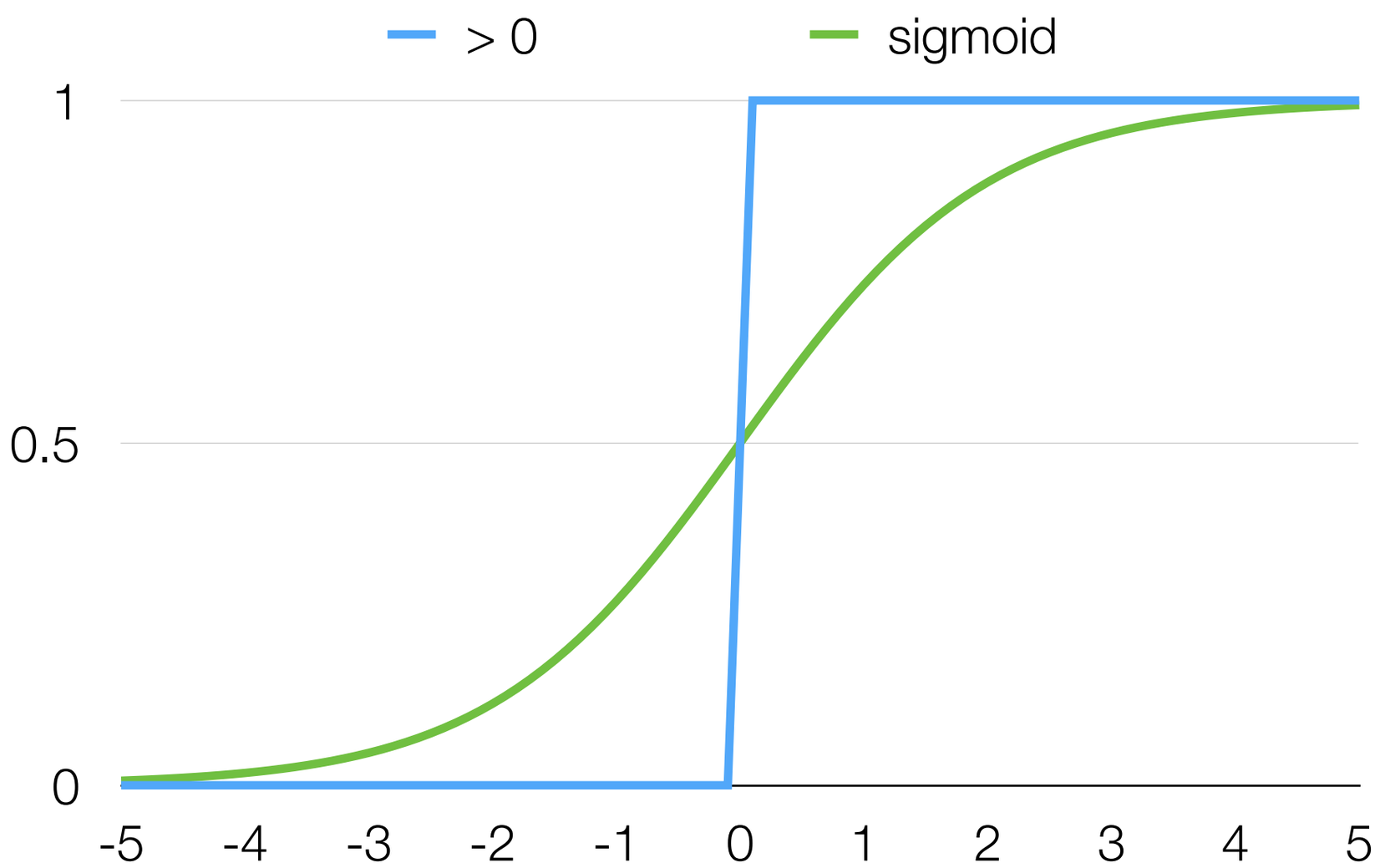
½ – (pos(*e*) = NN) > 0	not (pos(*e*) = NN)
(|*e*| < 5) + (pos(*e*) = NN) – 1½ > 0	|*e*| < 5 and pos(*e*) = NN
(|*e*| < 5) + (pos(*e*) = NN) – ½ > 0	|*e*| < 5 or pos(*e*) = NN

“Programmable” features

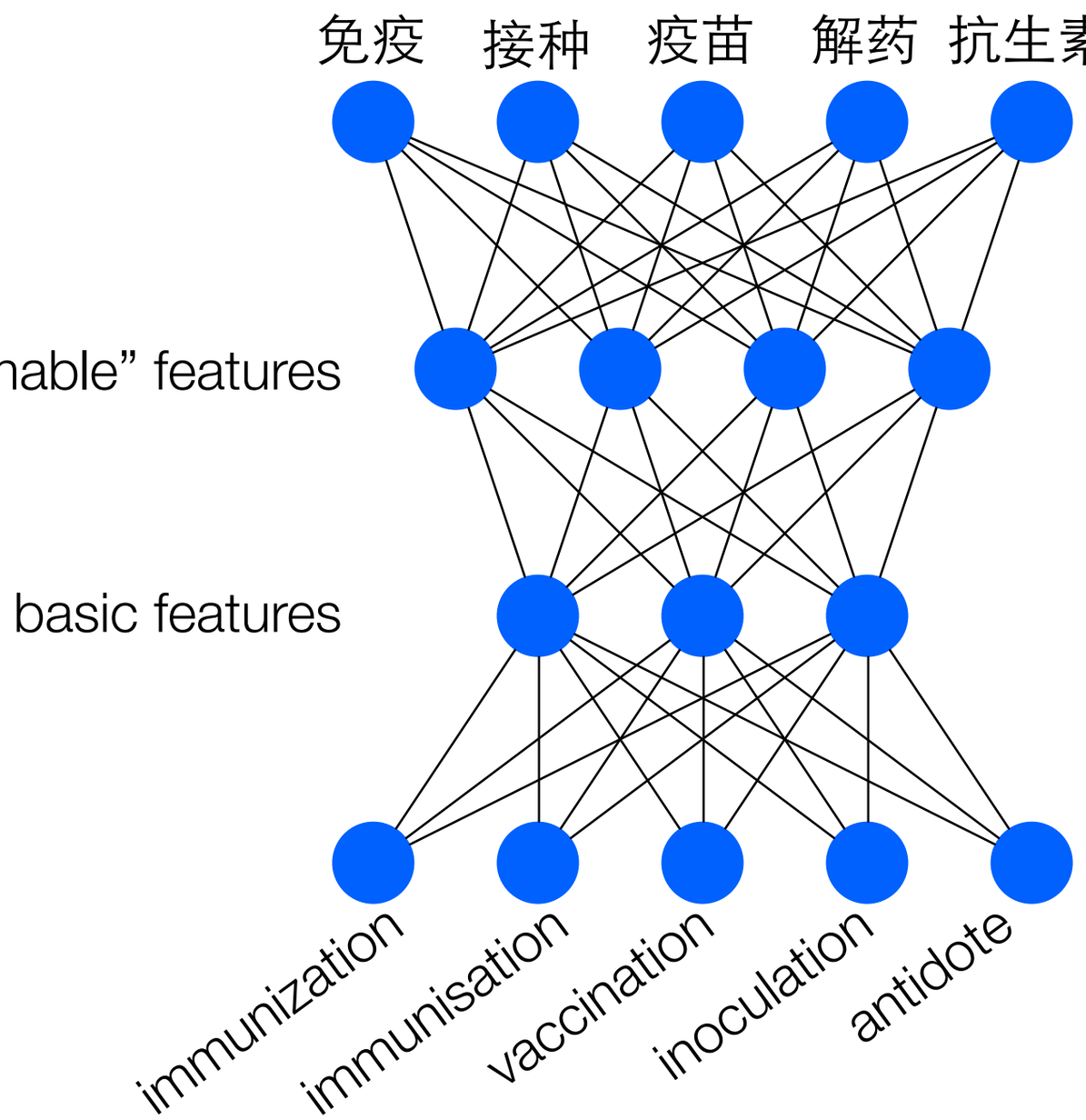
$$v_1(|e| < 5) + v_2(\text{pos}(e) = \text{NN}) + b > 0$$



Activation functions

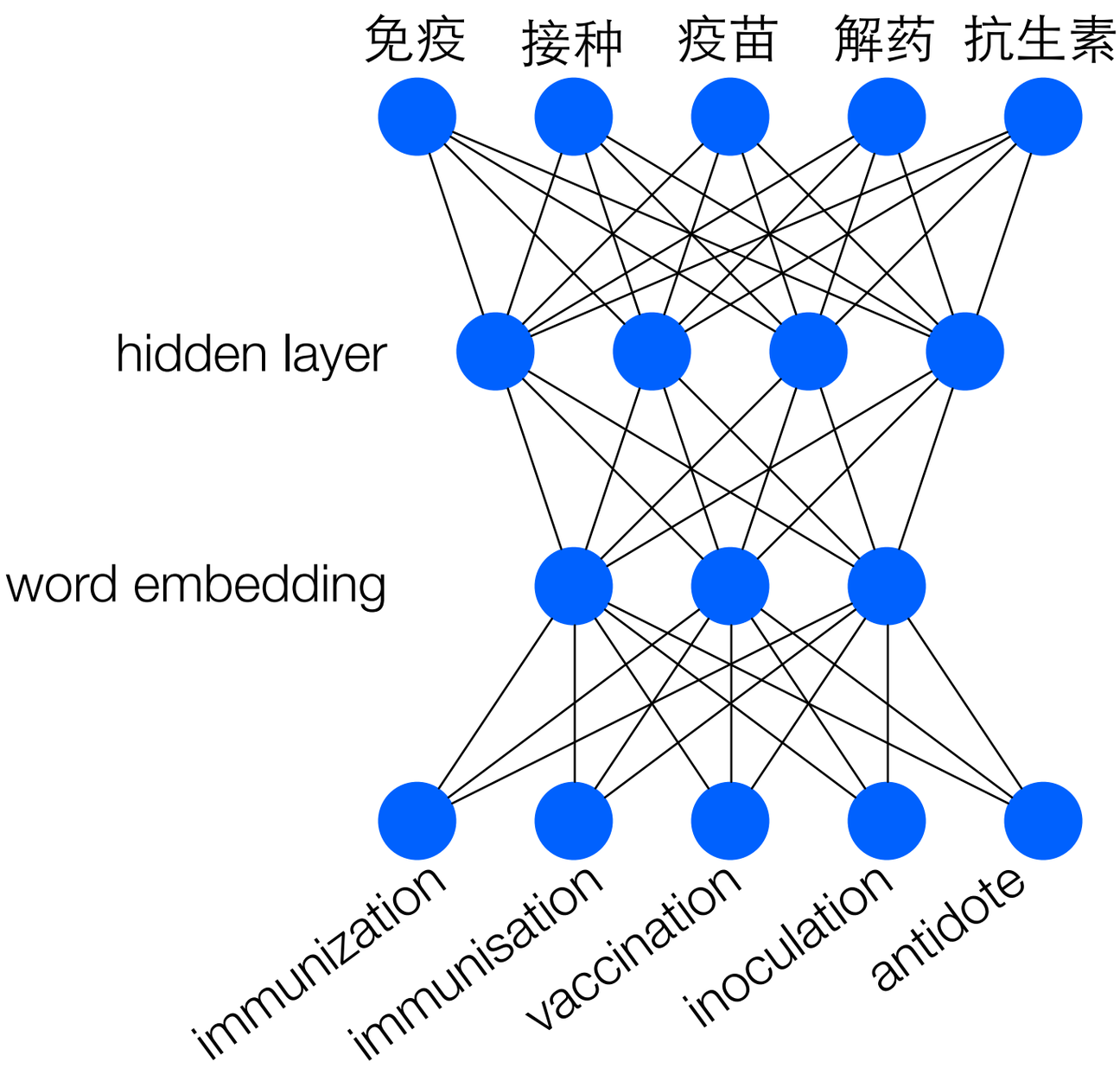


“Programmable” features as hidden layer



$$P(c|e) \propto \exp \sum_j w_{j,c}\, \text{sigmoid}(\sum_i v_{i,j}\, h_i(e) + b)$$

Feedforward neural network



$$P(c|e) \propto \exp \Sigma_j w_{j,c} \, \text{sigmoid}(\Sigma_i v_{i,j} u_{e,i} + b)$$


cognition
orientation
experimentation
hibernation
vacation
affliction
desertion
defamation
extortion
immolation
matriculation
evolution
visitation
addiction
perspiration
laceration
molestation
defection
assassination
lotion
indigestion
dislocation
mutilation
insemination
urination
disorientation
reeducation
procreation
maturation
assassination
masturbation
dissection
fibrillation
menstruation
infarction
infliction
domestication
ovulation
strangulation
abduction
sterilisation
stimulation
supplementation
nonprescription
constipation
malformation
amputation
castration
conviction
fertilisation
chelation
catheterization
incarceration
extradition
liposuction
intoxication
execution
detention
deportation
perforation
suffocation
propagation
deformation
asphyxiation
electrocution
eviction
transplantation
prescription
probation
replication
fertilization
gestation
hydration
sedation
respiration
decomposition
inhalation
repatriation
implantation
interrogation
detection
digestion
rehydration
reproduction
incubation
investigation
medication
hospitalization
retardation
exhaustion
hospitalisation
extinction
secretion
dysfunction
prosecution
detoxification
dehydration
pollination
adulteration
ingestion
degeneration
infection
recuperation
radiation
inflammation
precaution
litigation
mutation
irradiation
resuscitation
incineration
malnutrition
oxidation
starvation
disinfection
fermentation
contamination
combustion
subduction
acidification
malfunction
purification
ration
siltation
sedimentation
infestation
inundation
filtration
traction
lubrication
desalinization
suction
pressurization
devastation
prostitution
illumination
vegetation
abortion
gasification
condensation
contraception
evaporation
munition
precipitation
liquefaction
sterilization
desalination
ammunition
ignition
vibration
tuition
congestion
insulation
refrigeration
ventilation
inoculation
immunisation
vaccination
immunization
detonation
erection
navigation
naturalization
immigration
arbitration
irrigation
migration
emigration
sanitation
nutrition
education

cognition
orientation
experimentation
hibernation
vacation
affliction
desertion
defamation
extortion
immolation
matriculation
evolution
visitation
addiction
perspiration
laceration
molestation
defection
lotion
indigestion
dislocation
mutilation
assassination
insemination
urination
disorientation
reeducation
procreation
maturation
masturbation
assassination
domestication
dissection
menstruation
infarction
fibrillation
infliction
strangulation
ovulation
sterilisation
stimulation
constipation
malformation
amputation
abduction
supplementation
nonprescription
fertilisation
chelation
catheterization
liposuction
castration
conviction
incarceration
perforation
intoxication
extradition
execution
detention
deportation
suffocation
deformation
asphyxiation
electrocution
eviction
propagation
transplantation
prescription
probation
evacuati
replication
fertilization
gestation
hydration
sedation
respiration
decomposition
inhalation
repatriation
reloca
implantation
interrogation
detection
digestion
rehydration
reproduction
incubation
medication
hospitalization
investigation
retardation
exhaustion
hospitalisation
extinction
secretion
dysfunction
prosecution
dehydration
detoxification
pollination
adulteration
ingestion
degeneration
infection
inflammation
recuperation
radiation
precaution
litigation
mutation
irradiation
resuscitation
incineration
malnutrition
oxidation
starvation
disinfection
fermentation
contamination
combustion
subduction
acidification
malfunction
purification
ration
siltation
sedimentation
infestation
inundation
traction
filtration
lubrication
desalinization
cogeneration
suction
pressurization
devastation
prostitution
abortion
illumination
vegetation
condensation
evaporation
contraception
gasification
munition
precipitation
liquefaction
sterilization
ammunition
ignition
vibration
desalination
congestion
tuition
insulation
refrigeration
ventilation
detonation
erection
navigation
inoculation
immunisation
vaccination
immunization
naturalization
telecommunication
aviation
transportation
immigration
arbitration
irrigation
migration
emigration
sanitation
nutrition
education

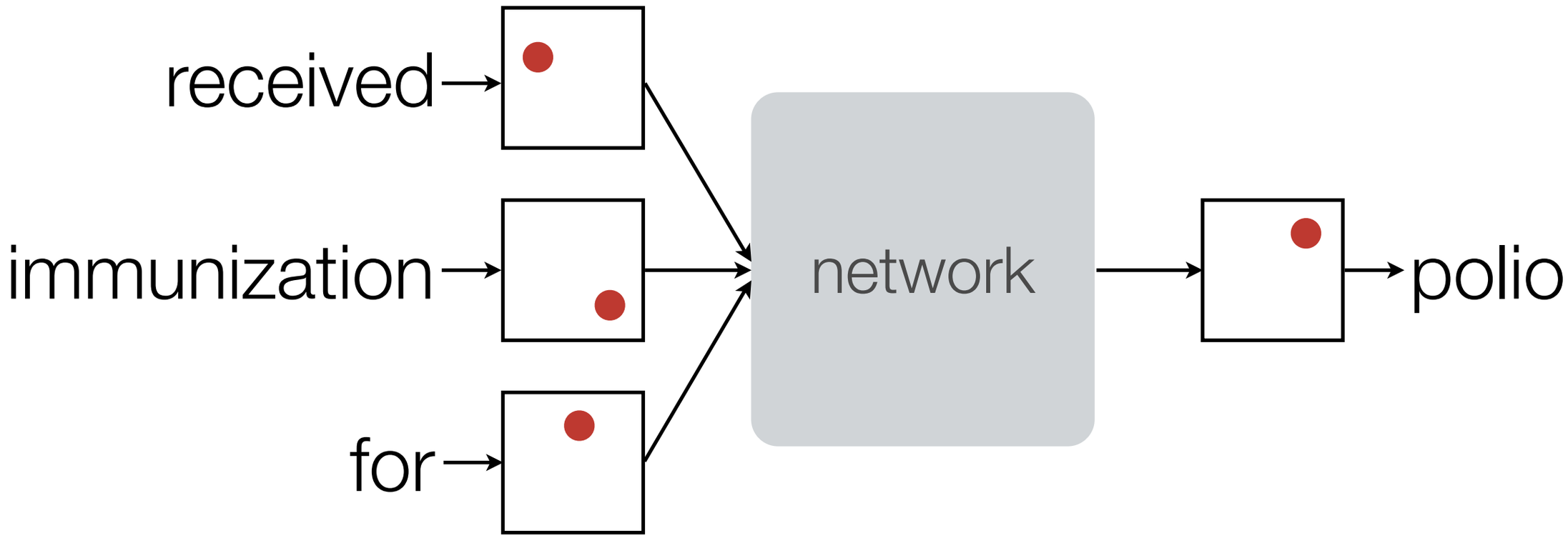
Fast training of neural networks

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. Decoding with large-scale neural language models improves translation. In *Proc. EMNLP 2013*.

Neural network language model

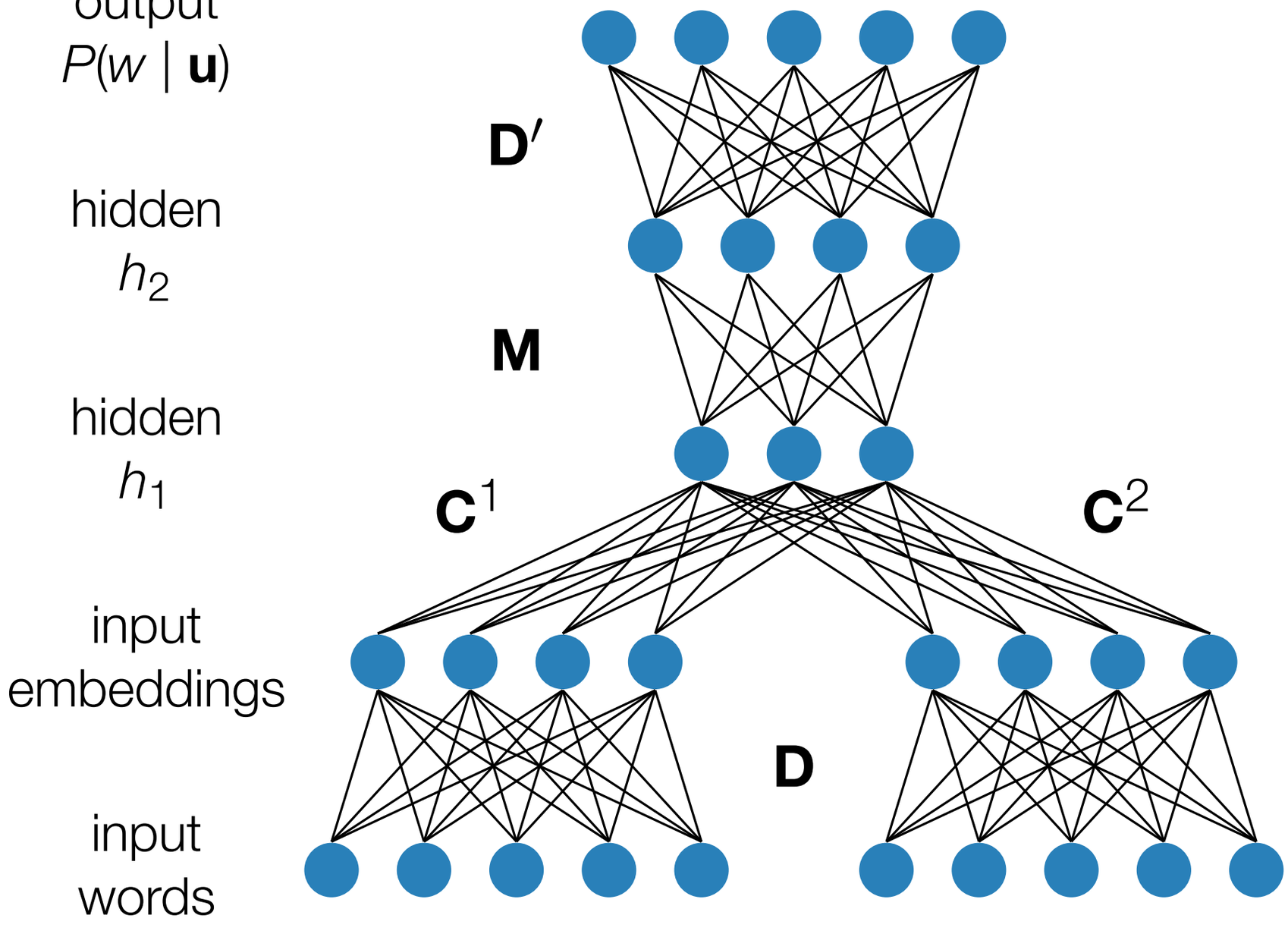
Bengio, 2003

Predict next English word *w* given previous English words **u**



Neural network language model

Bengio, 2003



Neural network language model

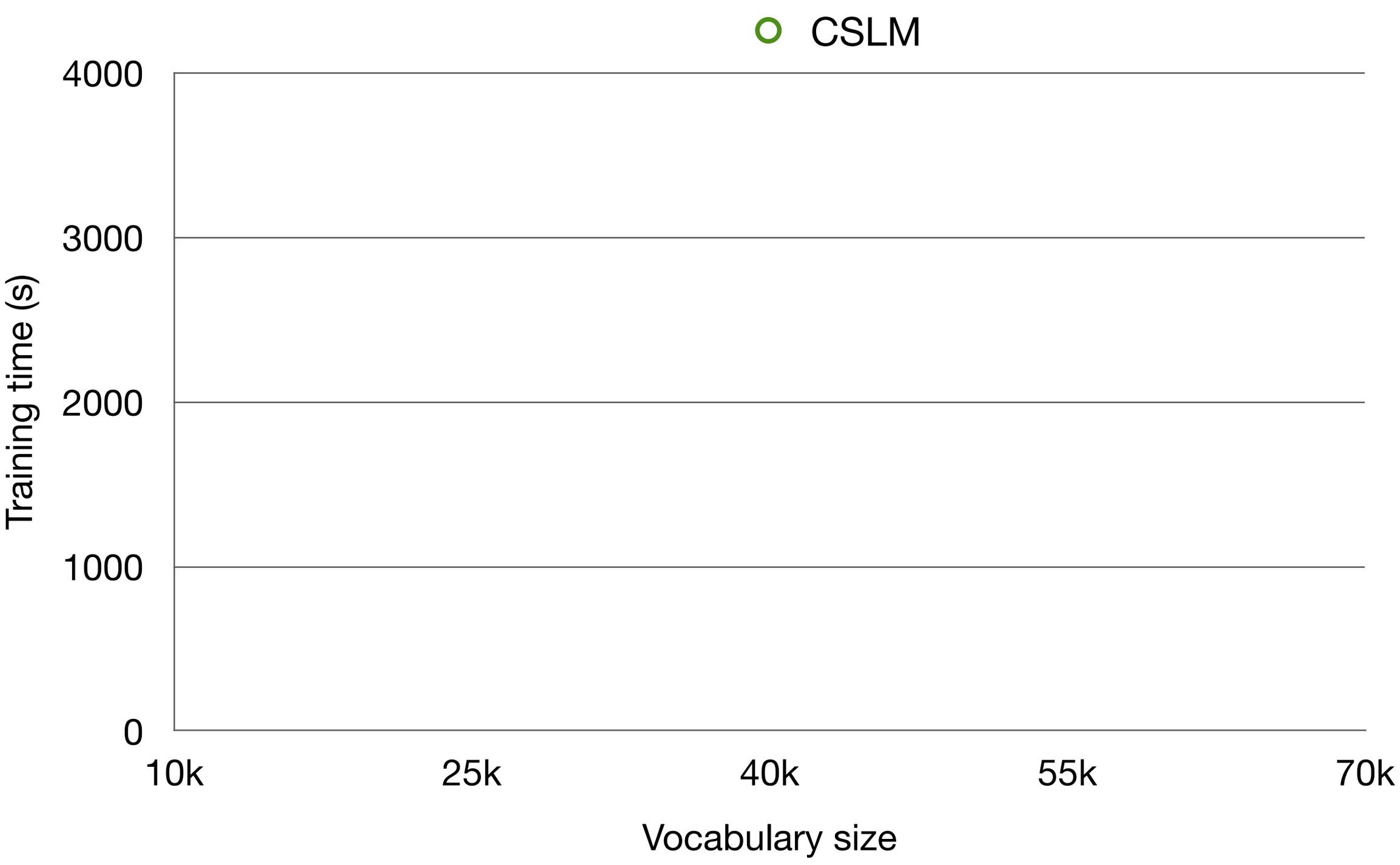
Bengio, 2003

Problem: normalization factor Z makes both training and decoding slow

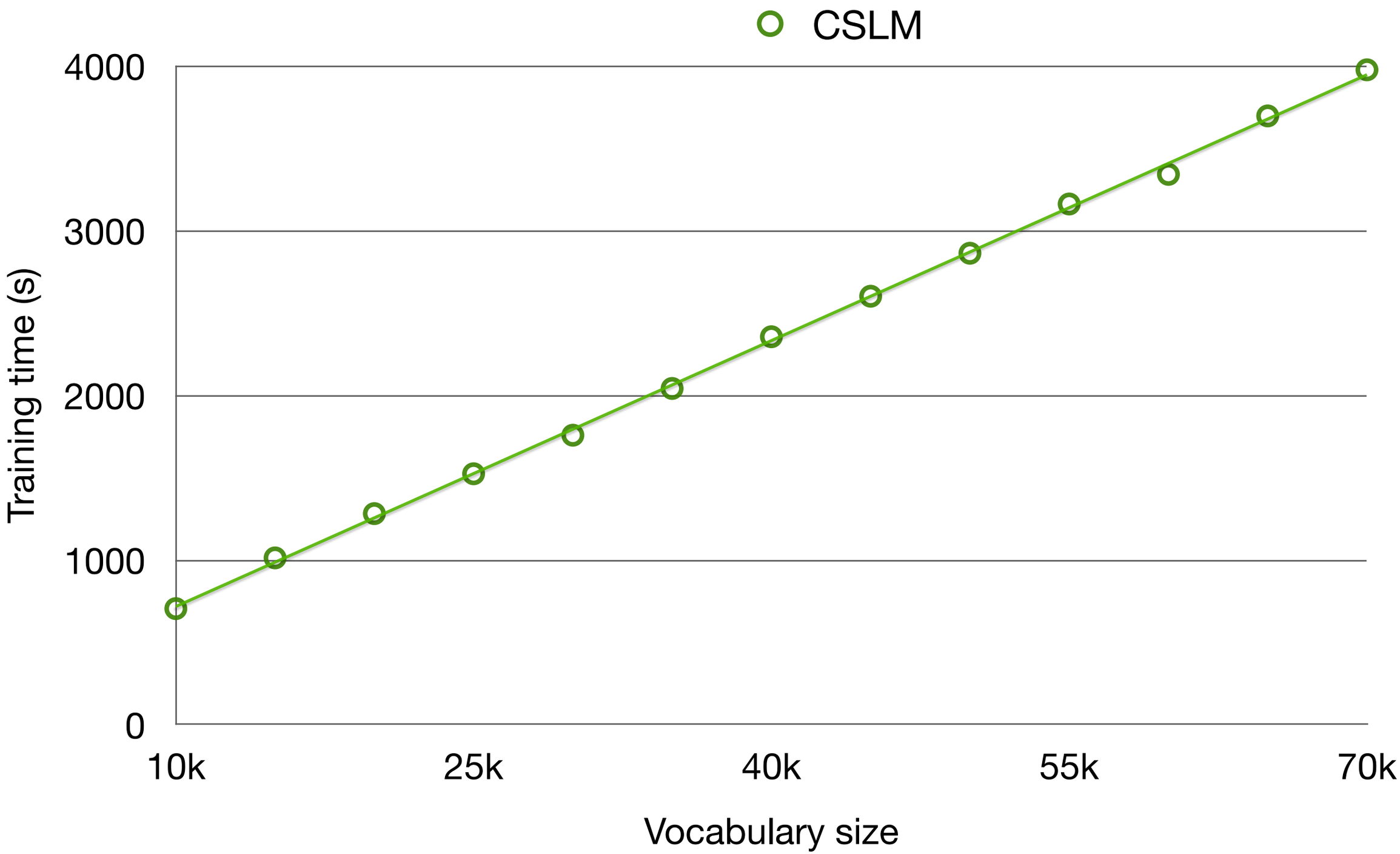
$$P(w \mid \mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \cdots$$

$$Z(\mathbf{u}) = \sum \exp \cdots$$

Speed vs. vocabulary size

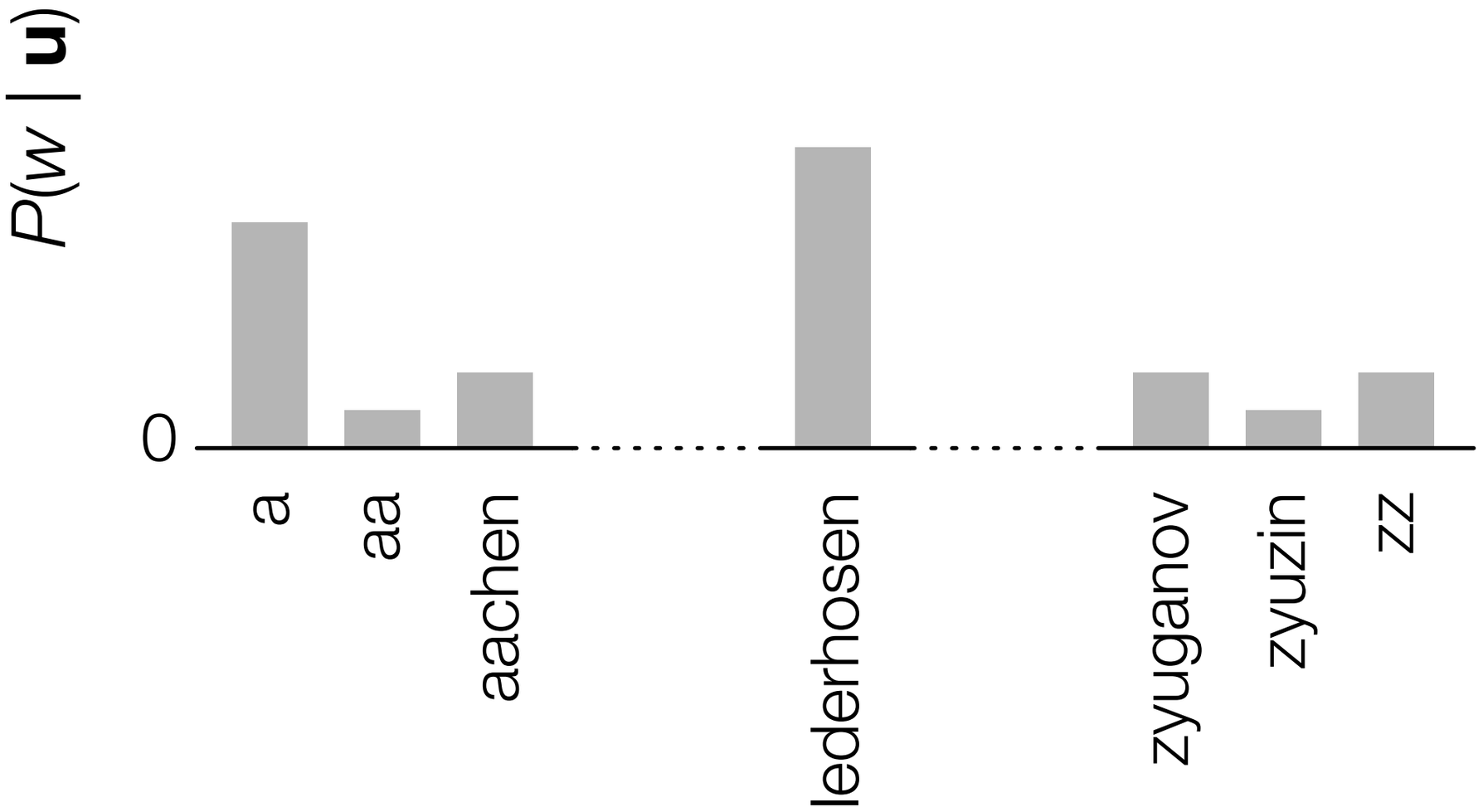


Speed vs. vocabulary size



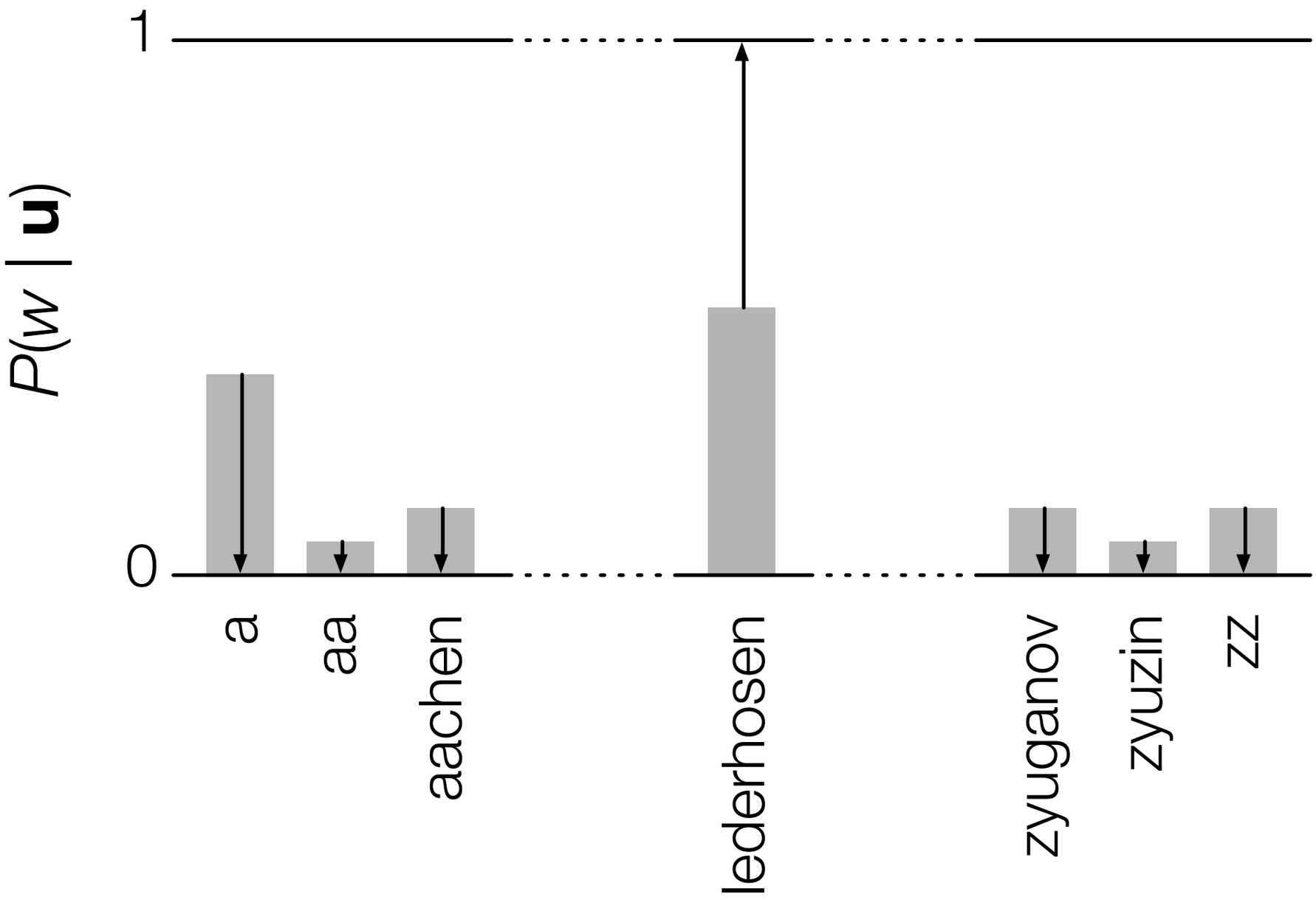
Maximum likelihood estimation

For each training example (***u***, *w*):



Maximum likelihood estimation

For each training example (**u**, *w*):



Noise contrastive estimation

Gutmann and Hyvarinen, 2010

For each training example (***u***, *w*):

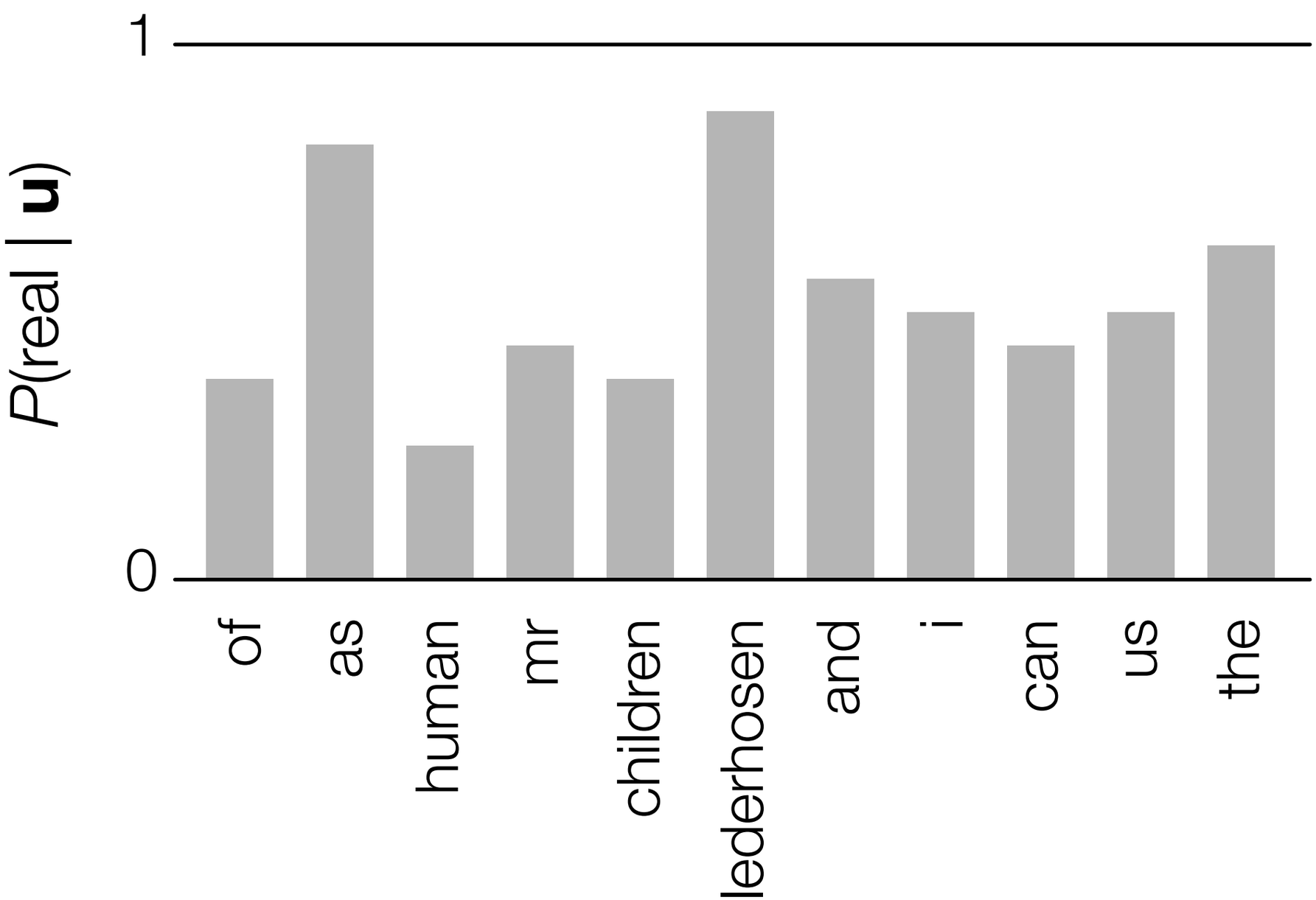
generate *k* noise samples

train model to classify real examples and noise samples

Noise contrastive estimation

Gutmann and Hyvarinen, 2010

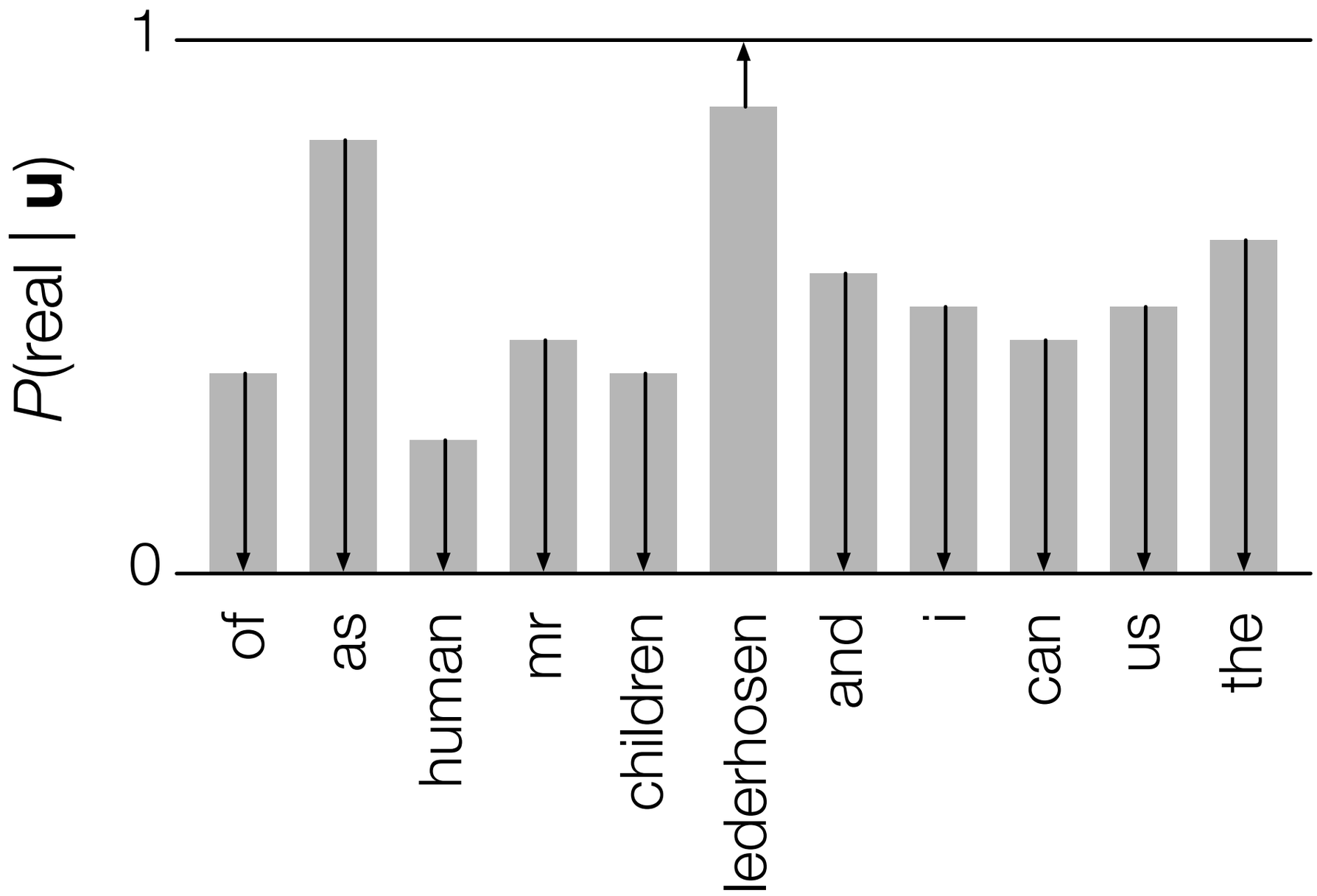
For each training example (**u**, *w*):



Noise contrastive estimation

Gutmann and Hyvarinen, 2010

For each training example (***u***, *w*):



Noise contrastive estimation

Gutmann and Hyvarinen, 2010

Bonus: this estimates Z also

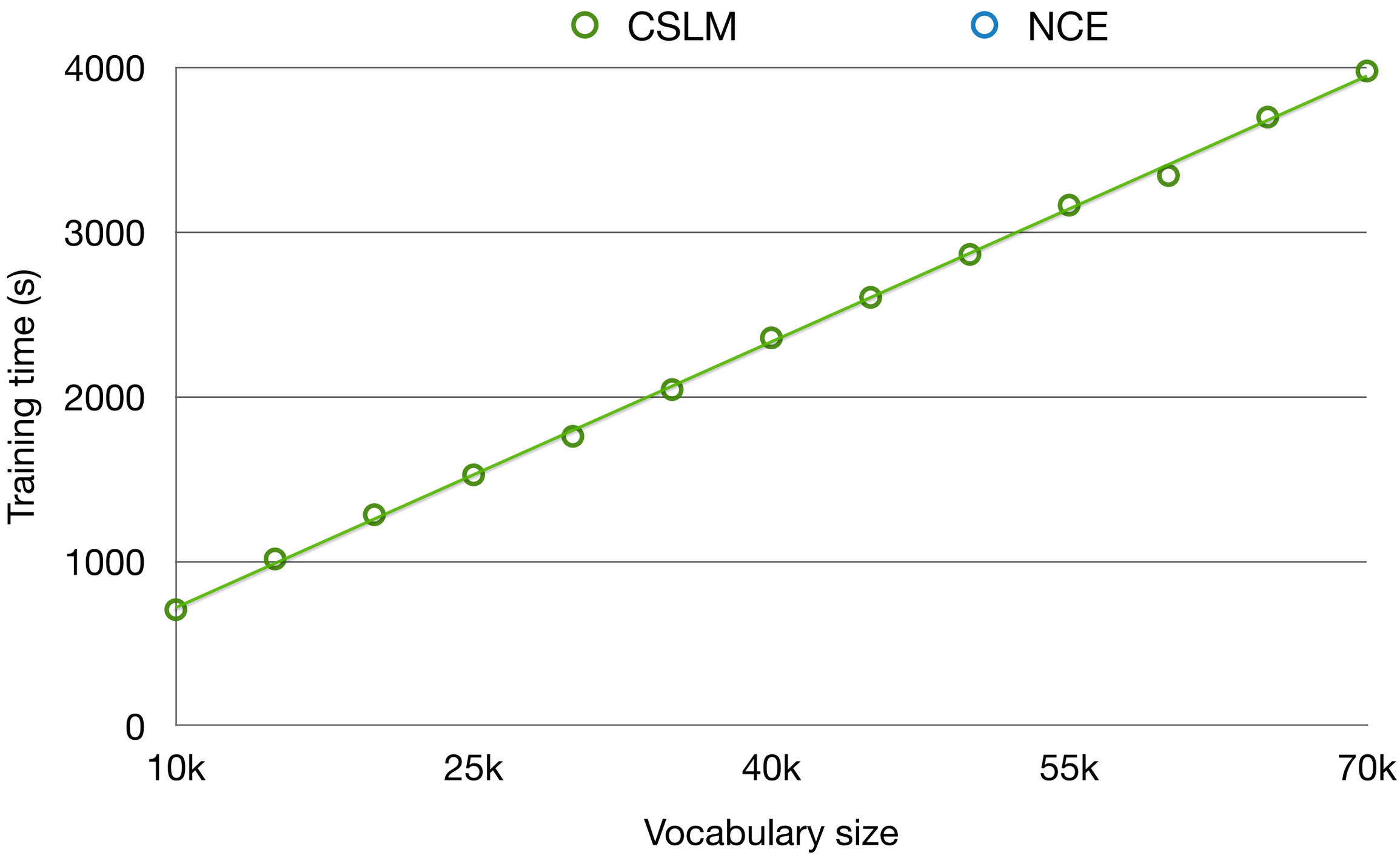
$$P(w \mid \mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \cdots$$

$$Z(\mathbf{u}) = \sum \exp \cdots$$

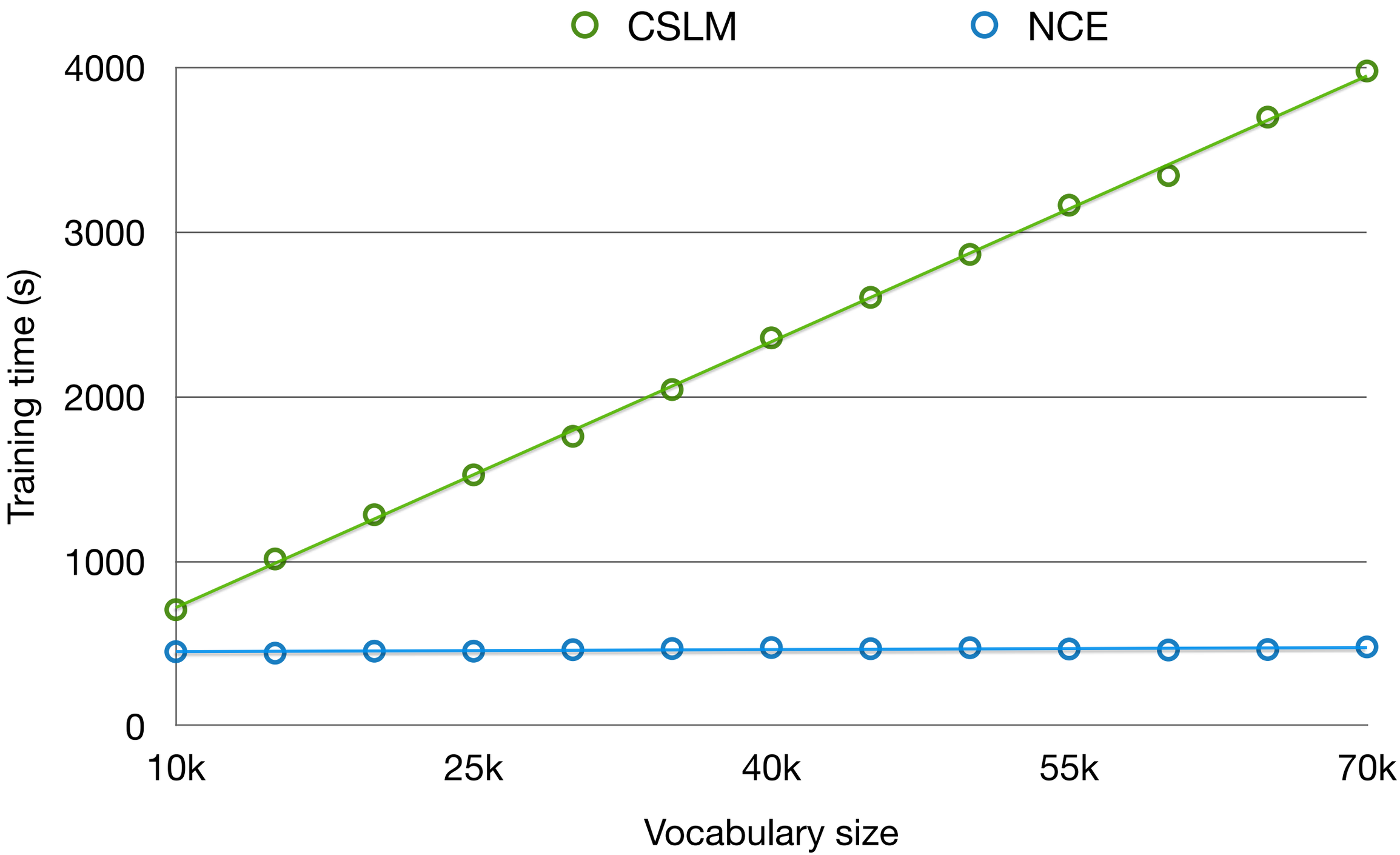
Mnih and Teh: just freeze $Z = 1$

⇒ During decoding, we can skip expensive calculation of Z

Speed vs. vocabulary size



Speed vs. vocabulary size

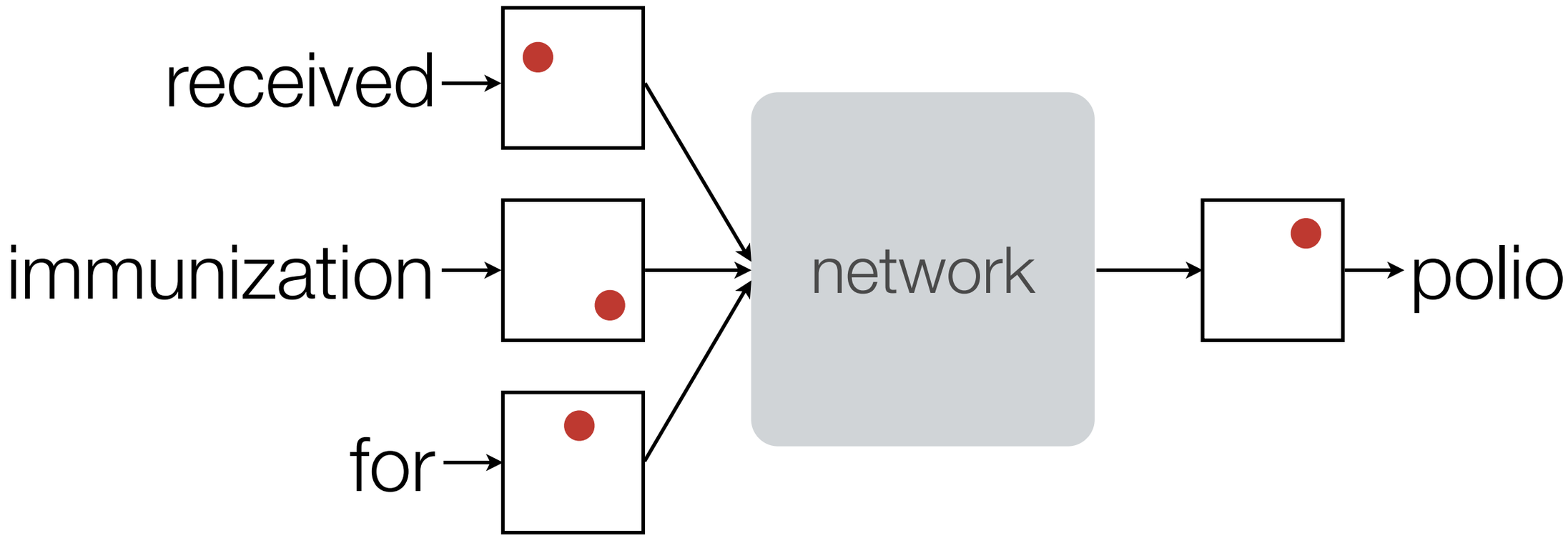


Neural networks for machine translation

Neural network language model

Bengio, 2003

Predict next English word *w* given previous English words **u**



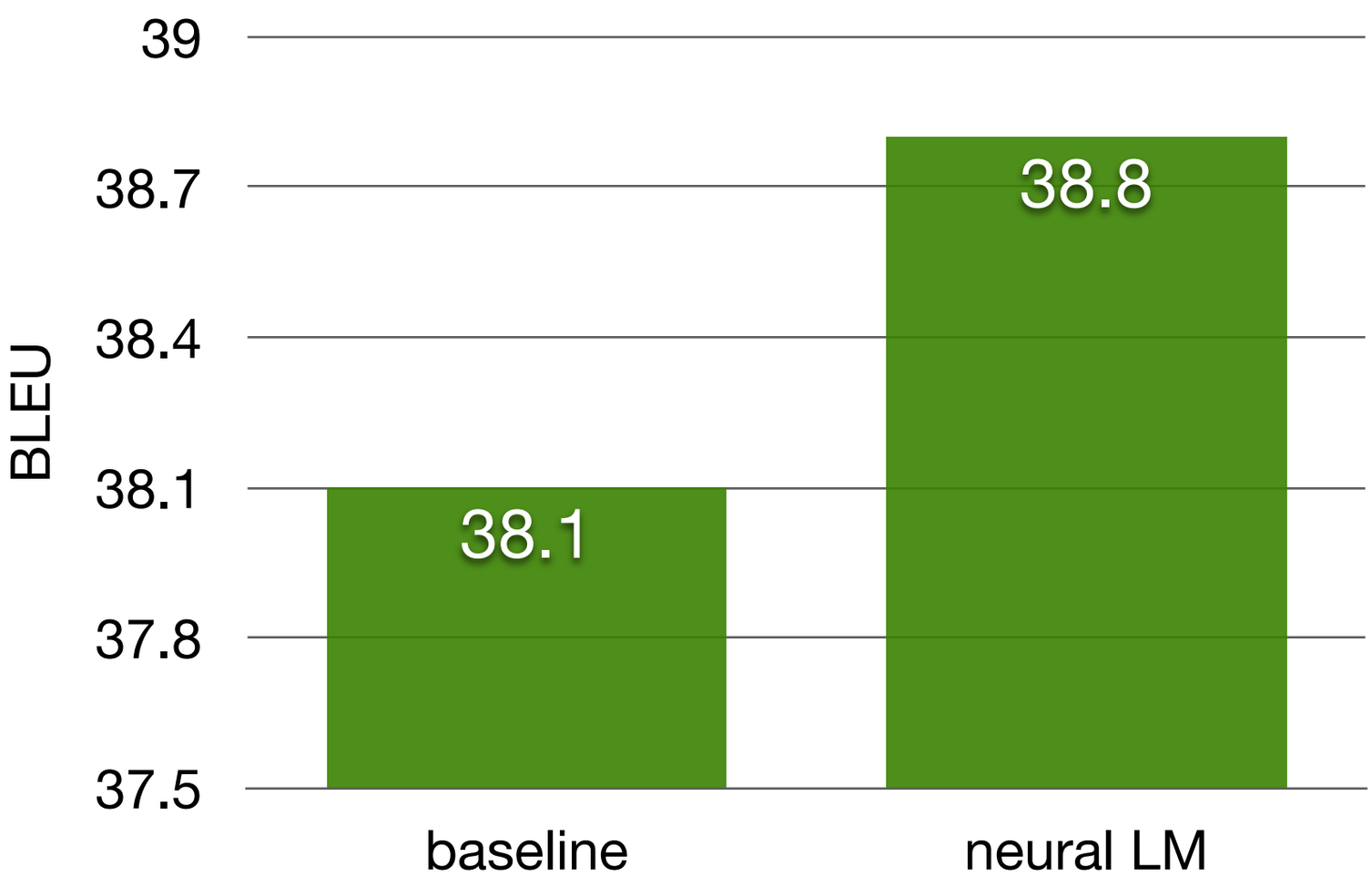
Translation results

Chinese-English
BLEU

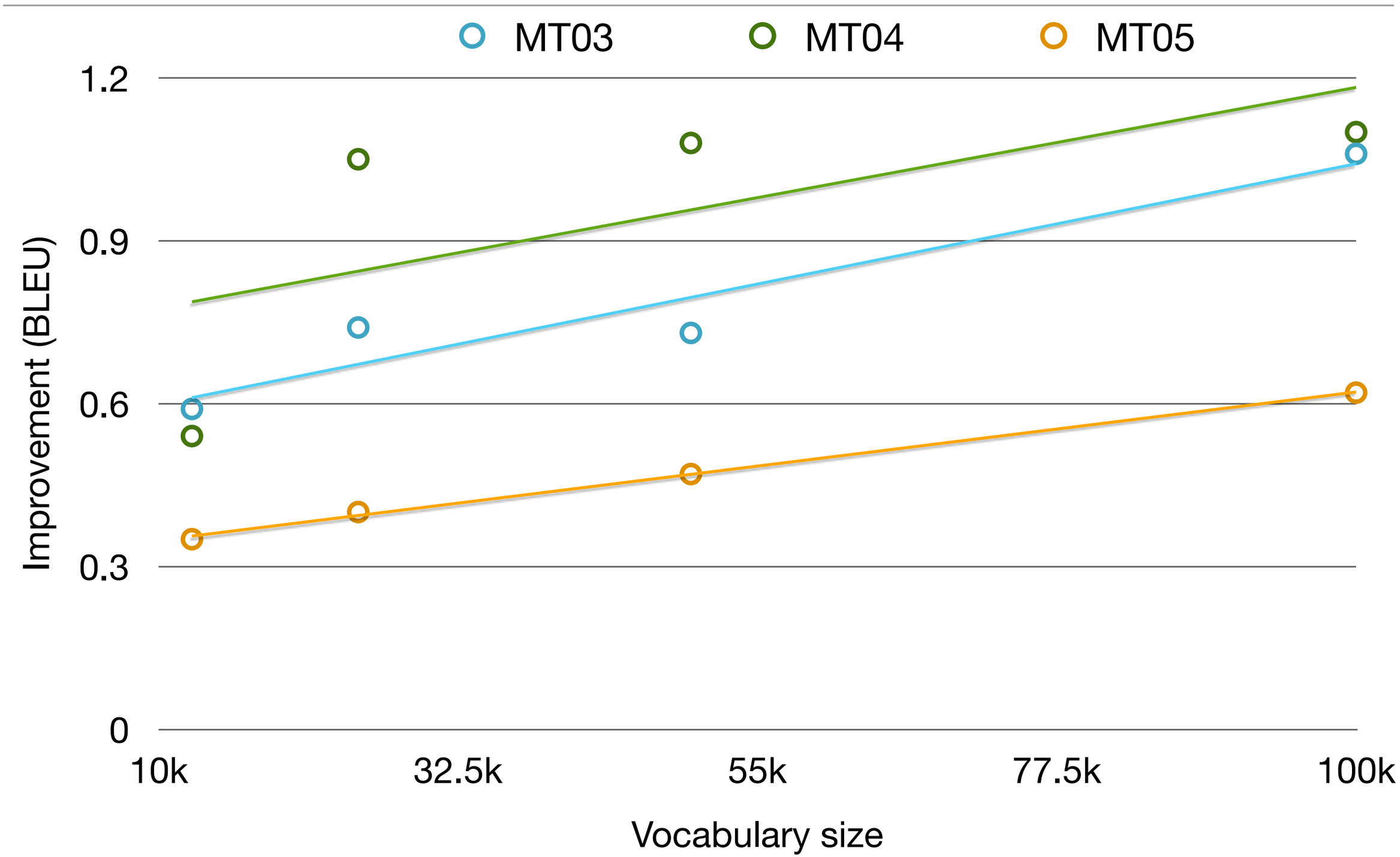
	2004	2005	2006
baseline	38.4	37.7	34.3
neural LM *n*-best reranking	38.6	37.8	34.7
neural LM during decoding	39.5	38.8	34.9

Translation results

BOLT Chinese-English forum



BLEU vs. Vocabulary size



Neural network translation models

Devlin et al., 2014

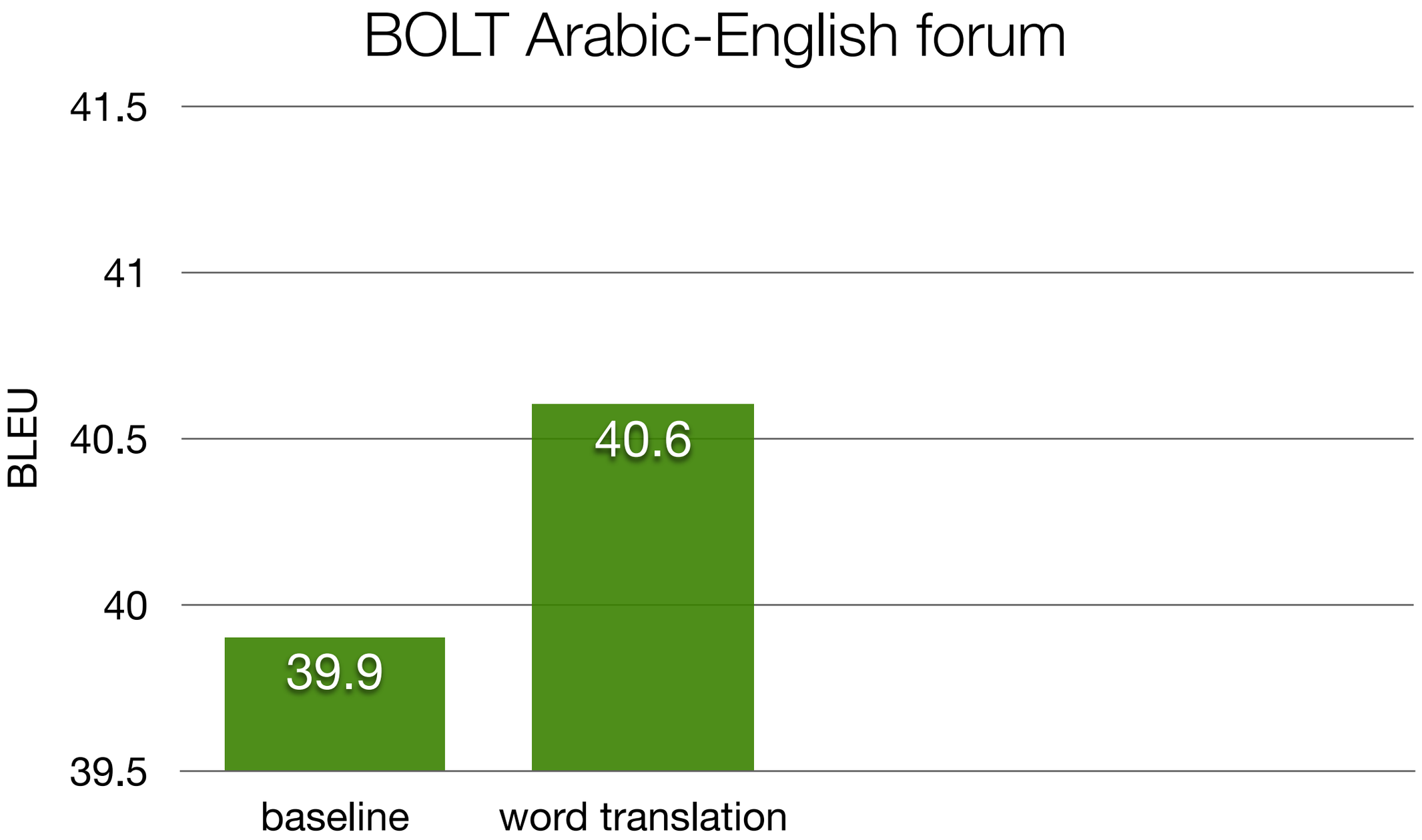
Predict next English word given Chinese words and previous English words

…that have diplomatic relations with North…

P(relations | 帮教, …)

澳洲 是 与 北韩 有 帮教 的 少数 国家 之一

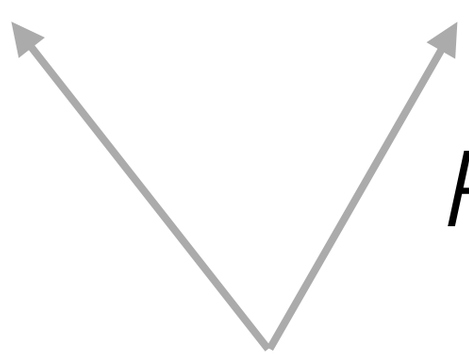
Translation results



Neural network translation models

Fertility: Predict number of English words given Chinese words

…that have diplomatic relations with North…

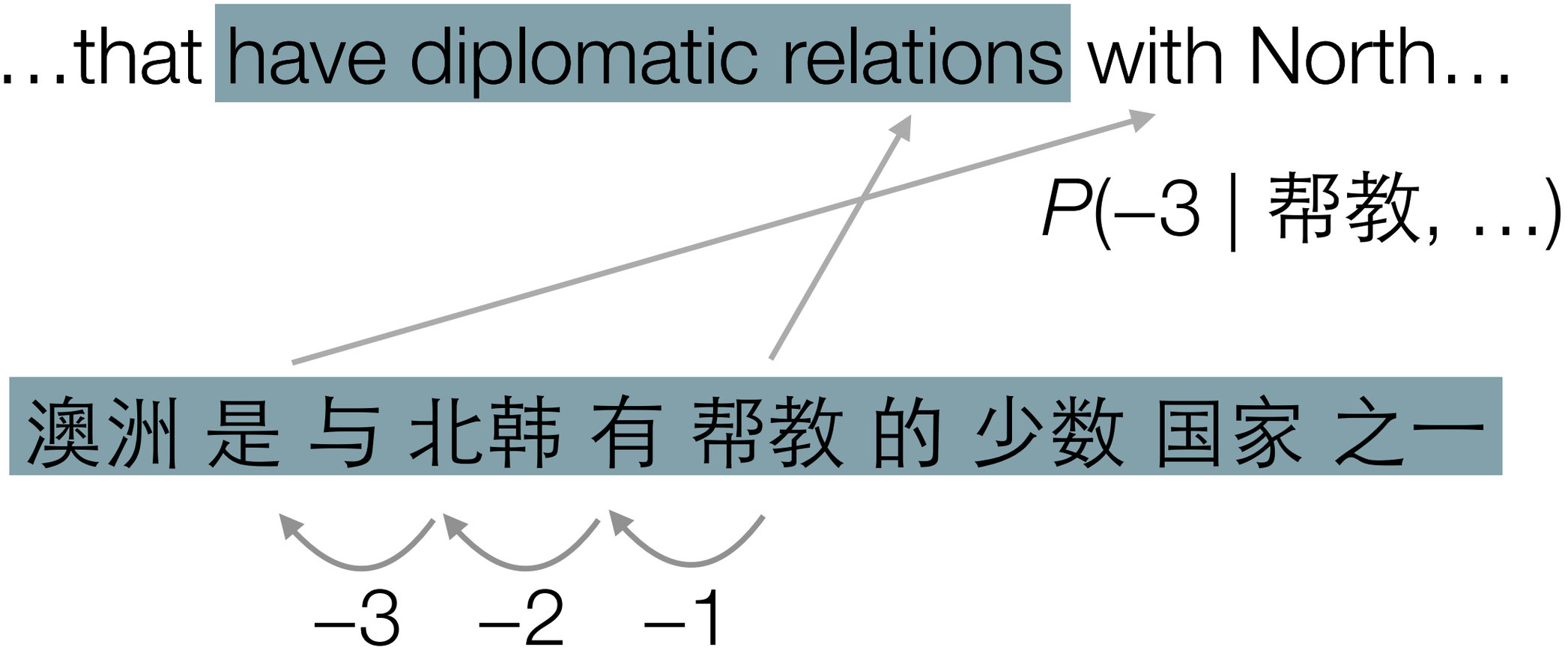


P(2 | 帮教, …)

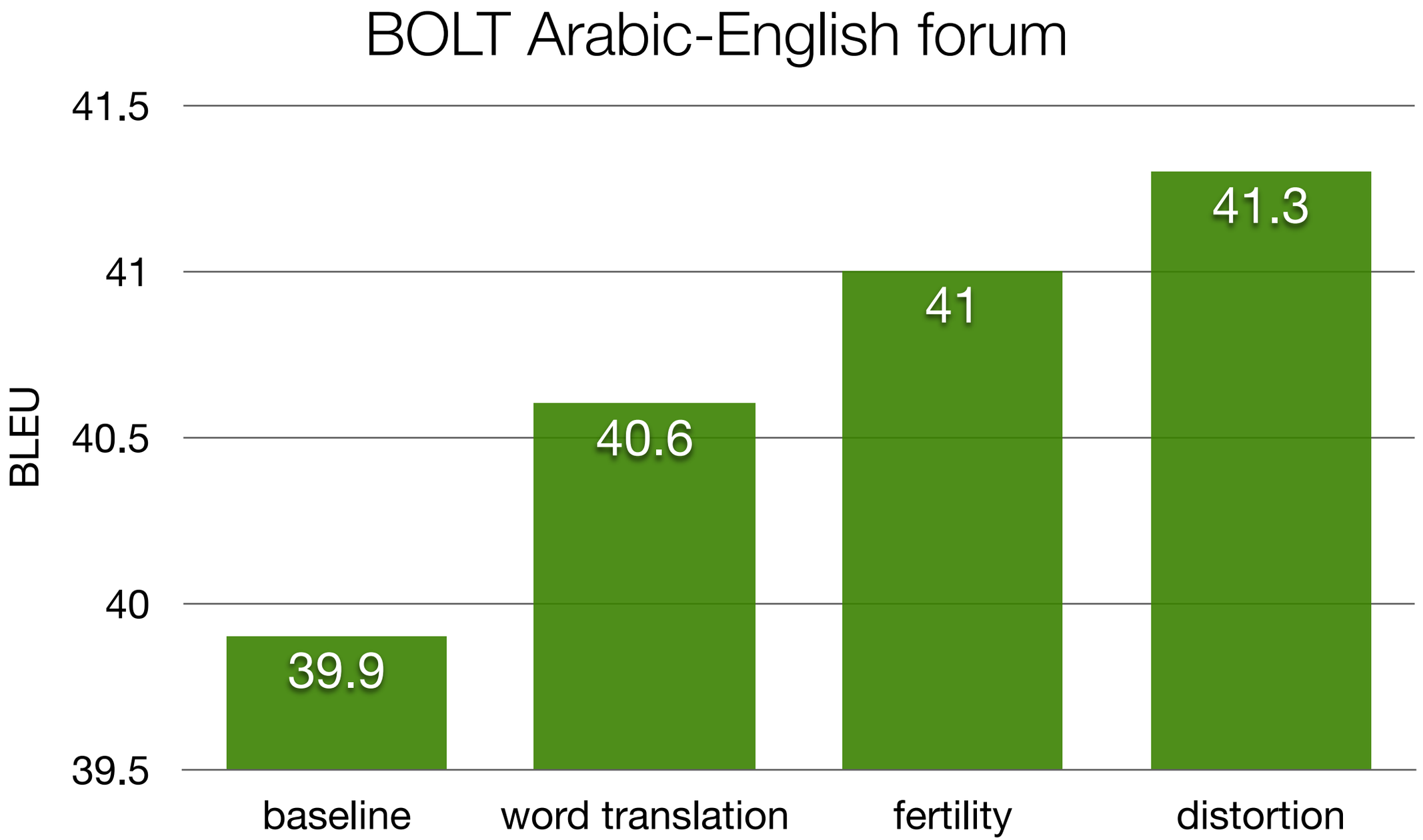
澳洲 是 与 北韩 有 帮教 的 少数 国家 之一

Neural network translation models

Distortion: Predict reordering given Chinese words and previous English words



Translation results



Conclusions

- Neural networks can be thought of as automatically inventing features, even for very large contexts

- A major obstacle (large output vocabulary) is removed

- Deep networks, pretraining, etc. are not needed?

- Large improvements are possible