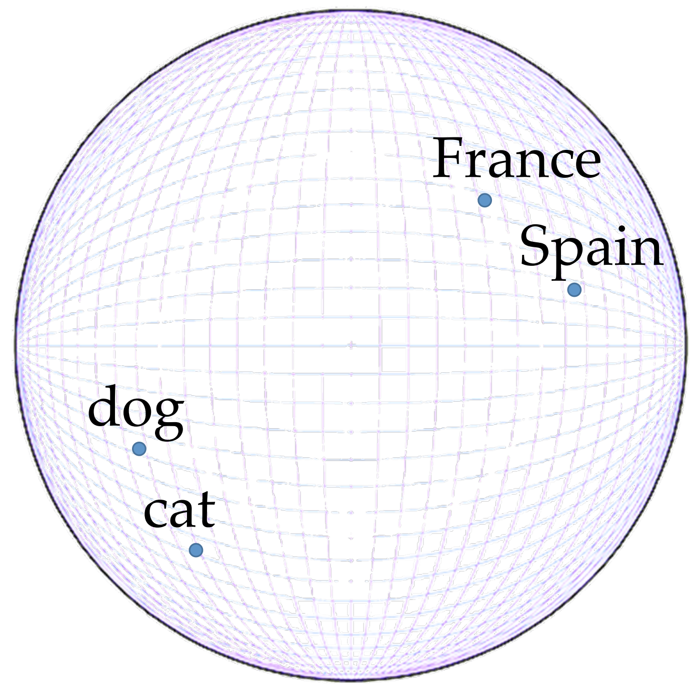
Learning to translate with neural networks

Michael Auli

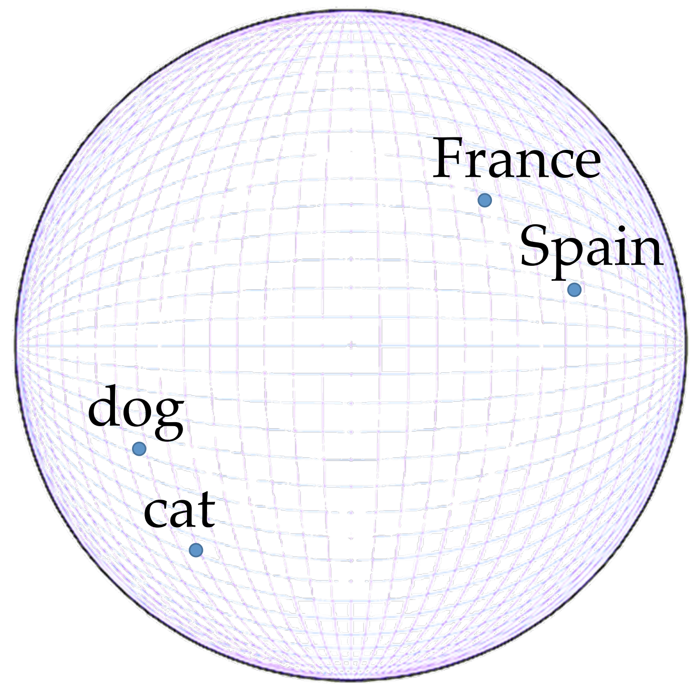
Neural networks for text processing

- Similar words near each other



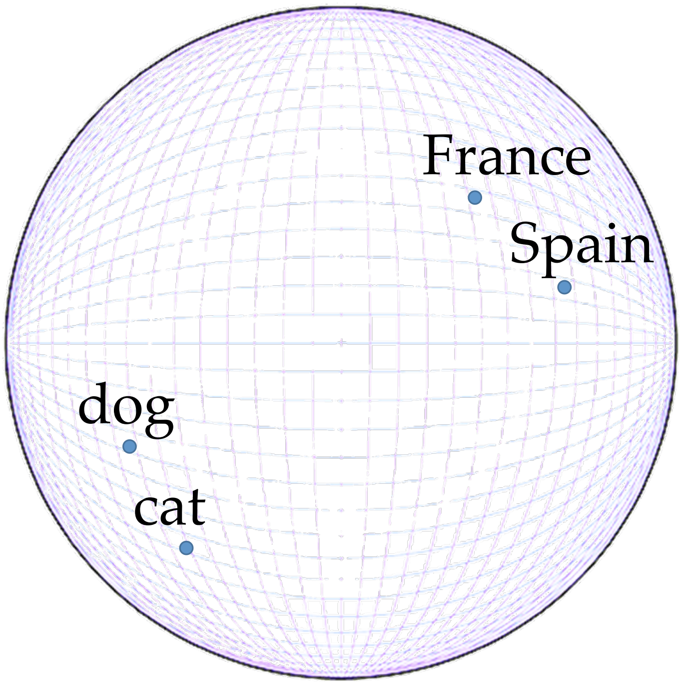
Neural networks for text processing

- Similar words near each other
- Changing model parameters for one example effects similar words in similar contexts



Neural networks for text processing

- Similar words near each other
- Changing model parameters for one example effects similar words in similar contexts
- Traditional discrete models treat each word separately



Neural networks

IMAGENET	Error 26% → 15% (Krizhevsky 2012)
	Error 27% → 18 % (Hinton 2012)
Language Modeling	PPLX 141 → 101 (Mikolov 2011)

Neural networks

IMAGENET Error 26% → 15% (Krizhevsky 2012)

Error 27% → 18 % (Hinton 2012)

Language Modeling PPLX 141 → 101 (Mikolov 2011)

Machine Translation This talk

Le (2012), Kalchbrenner (2013),
Devlin (2014), Sutskever (2014),
Cho (2014), …

What happened in MT over the past 10 years?

What happened in MT over the past 10 years?

“Learning **simple models** from large bi-texts is a solved problem”

(Lopez & Post, 2013)

What happened in MT over the past 10 years?

"Learning **simple models** from large bi-texts is a solved problem"

(Lopez & Post, 2013)

Google Translate
Break through language barriers.

bing Translator

WMT 2013  9/10 times

Machine translation

本 地区 的 发展 和 进步 。

development and progress of the region .

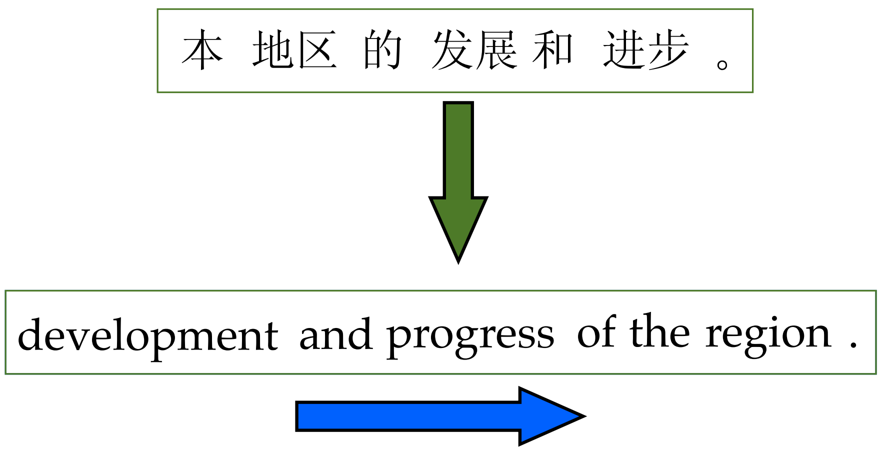
Machine translation

本 地区 的 发展 和 进步 。

development and progress of the region .

Translation modeling

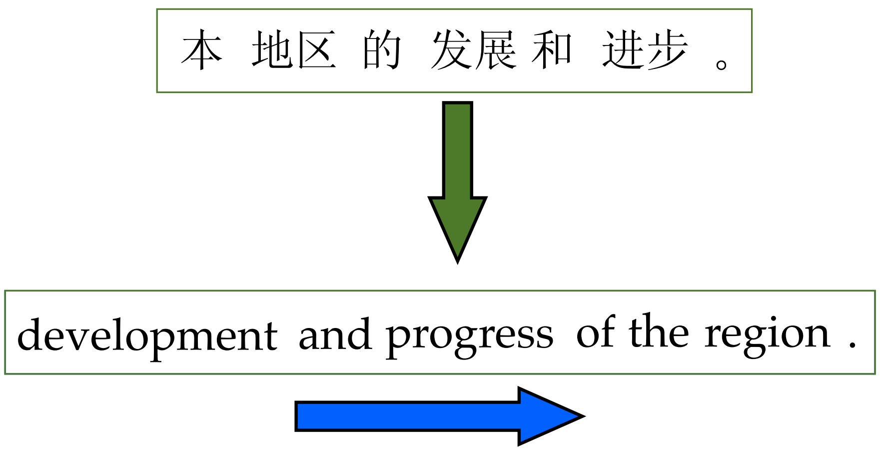
Machine translation



Translation modeling

Language modeling

Machine translation

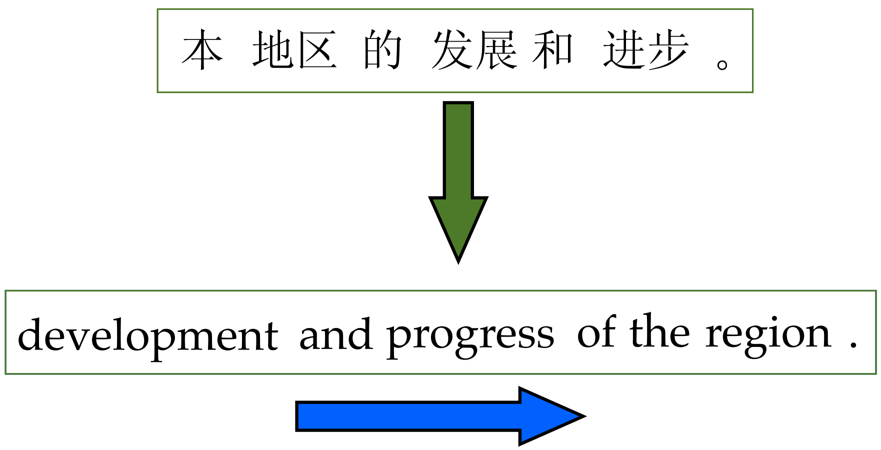


Translation modeling

Language modeling

Optimization

Machine translation



Translation modeling

Language modeling

Optimization

Reordering

Machine translation

本 地区 的 发展 和 进步 。

development and progress of the region .

Translation modeling 

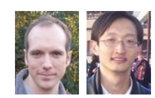
Auli et al., EMNLP 2013; Hu et al., EACL 2014

Language modeling 

Optimization

Auli & Gao, ACL 2014 

Reordering

Auli et al., EMNLP 2014 

Koehn et al. (2003)

Discrete phrase-based translation

本 地区 的 发展 和 进步 。

development and progress of the region .

Koehn et al. (2003)

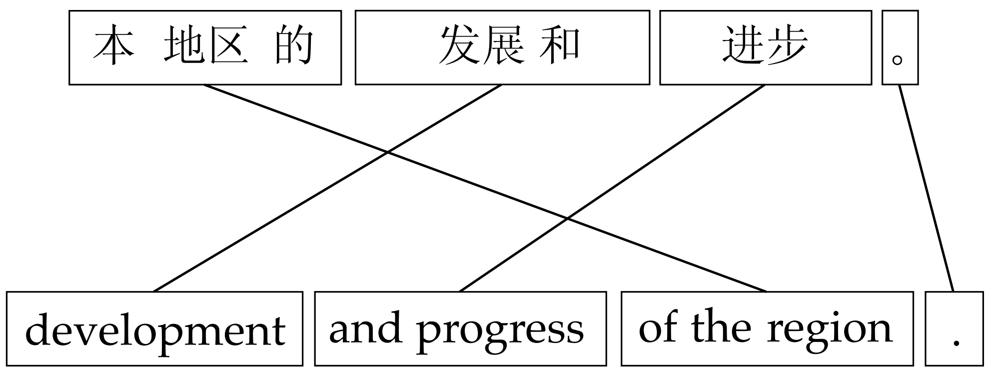
Discrete phrase-based translation

本 地区 的	发展 和	进步	。

development	and progress	of the region	.

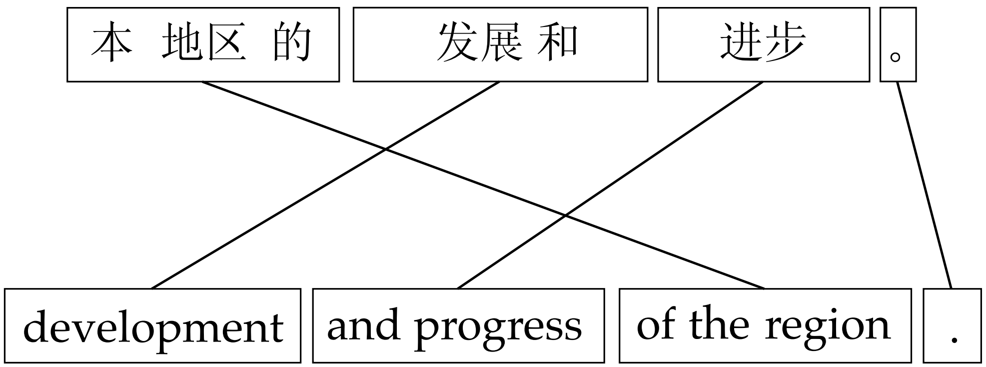
Koehn et al. (2003)

Discrete phrase-based translation



Koehn et al. (2003)

Discrete phrase-based translation



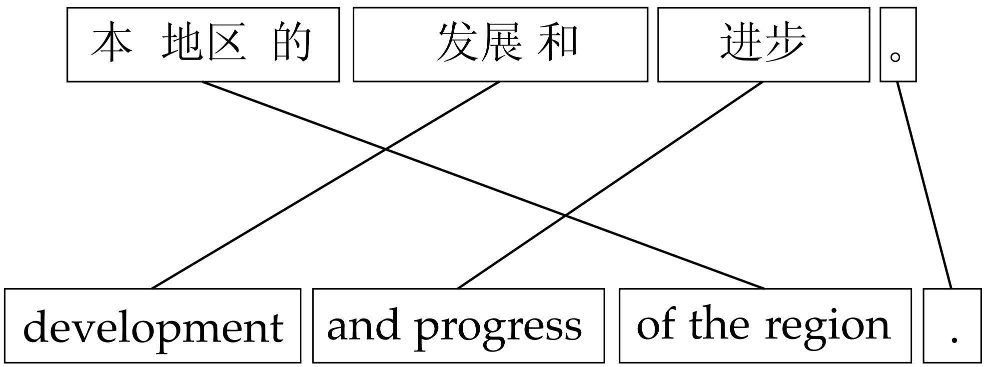
本 地区 的 → of the region

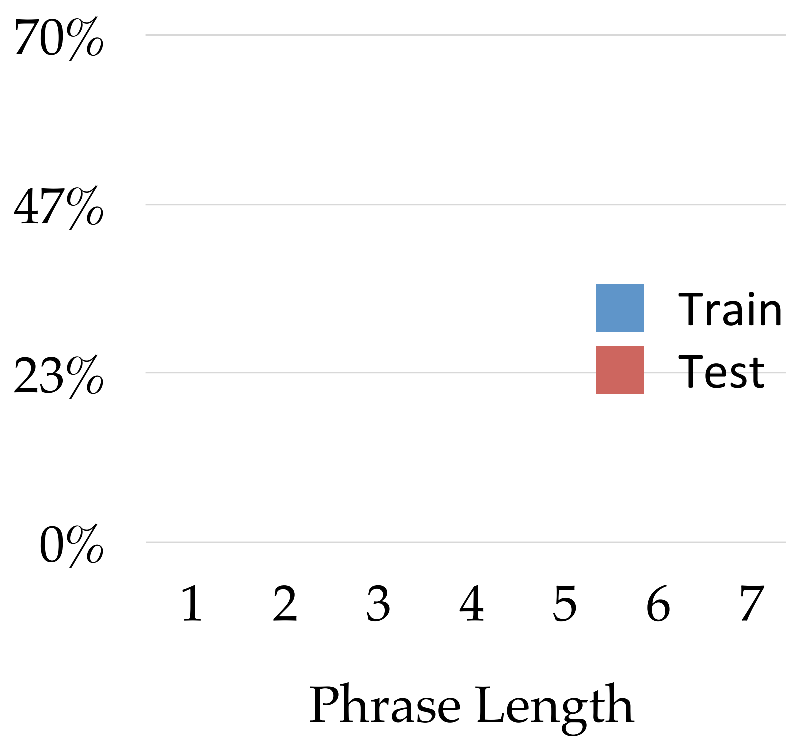
发展 → development

和 进步 → and progress

Koehn et al. (2003)

Discrete phrase-based translation





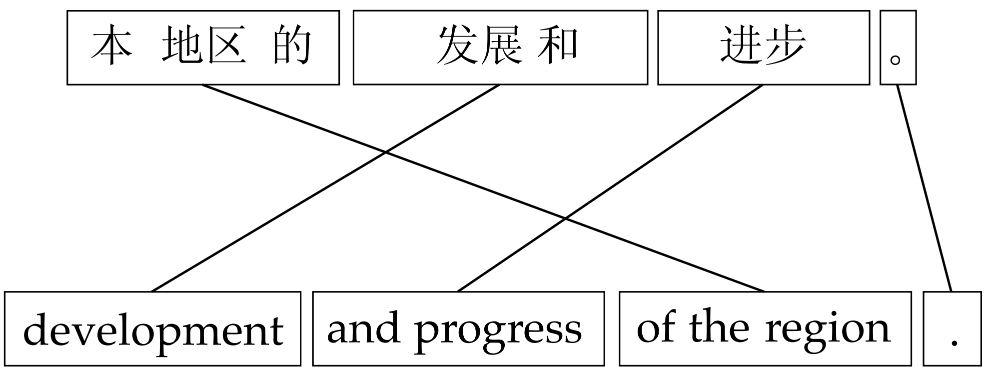
本 地区 的 → of the region

发展 → development

和 进步 → and progress

Koehn et al. (2003)

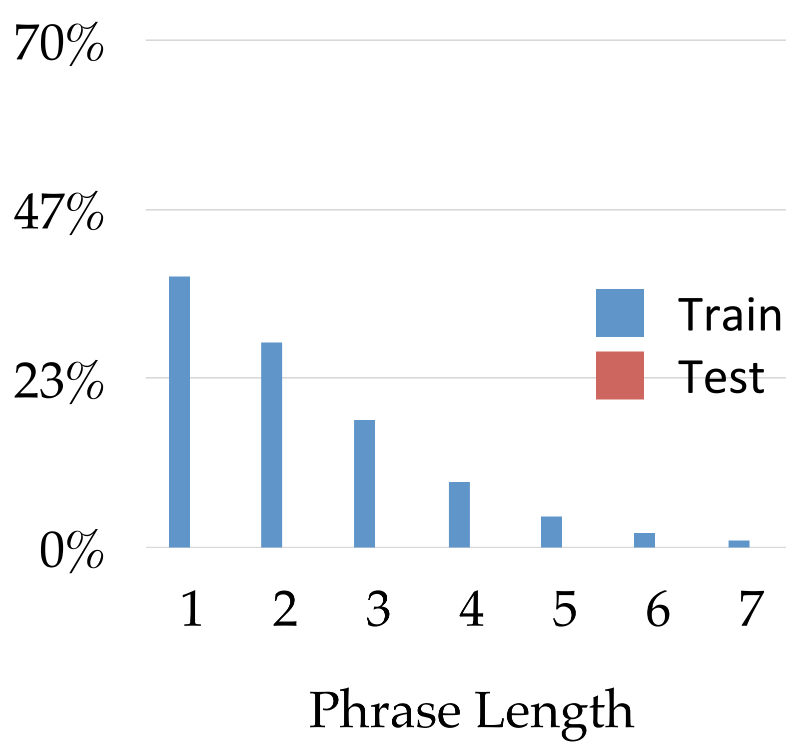
Discrete phrase-based translation



本 地区 的 → of the region

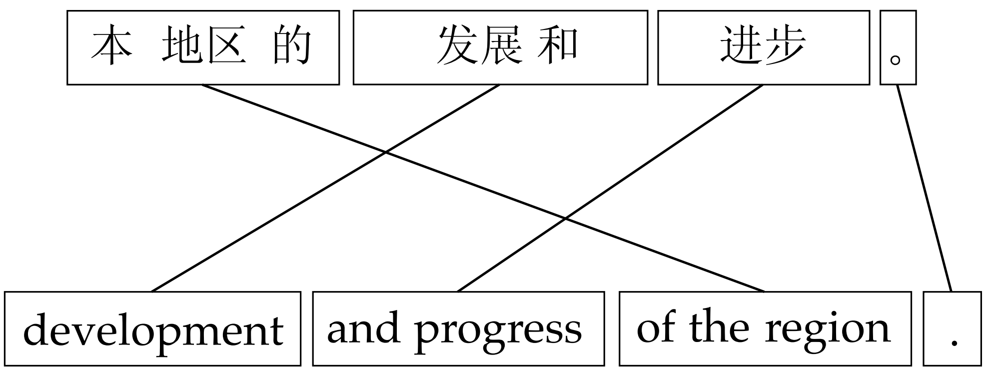
发展 → development

和 进步 → and progress



Koehn et al. (2003)

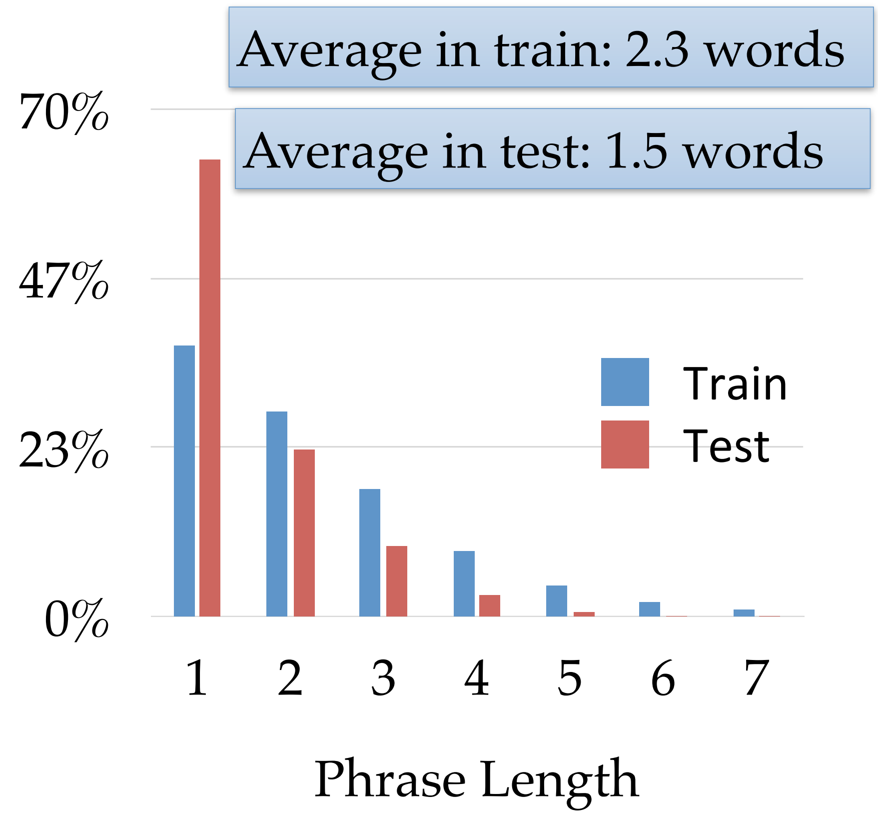
Discrete phrase-based translation



本 地区 的 → of the region

发展 → development

和 进步 → and progress



Kneser & Ney (1996)

Discrete n-gram language modeling

p(progress in the region) =

Kneser & Ney (1996)

Discrete n-gram language modeling

p(progress in the region) =

Train data:

…
development and progress of
the region. in …
…

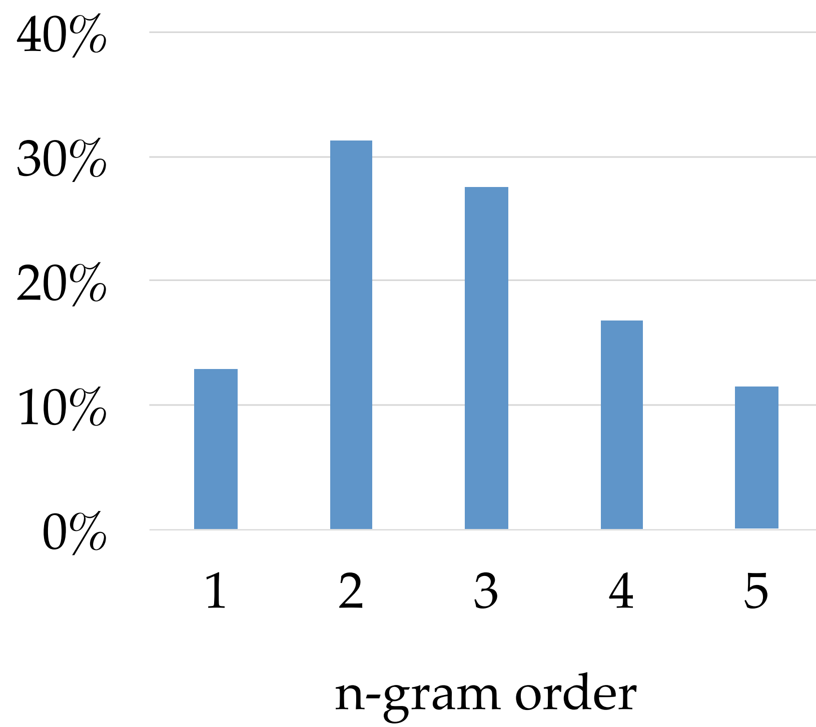
Kneser & Ney (1996)

Discrete n-gram language modeling

p(progress in the region) =
  p(progress) p(in)
  p(the) p(region | the)

Train data:

…
development and progress of
the region. in …
…



Does not include out-of-vocabulary tokens

Kneser & Ney (1996)

Discrete n-gram language modeling

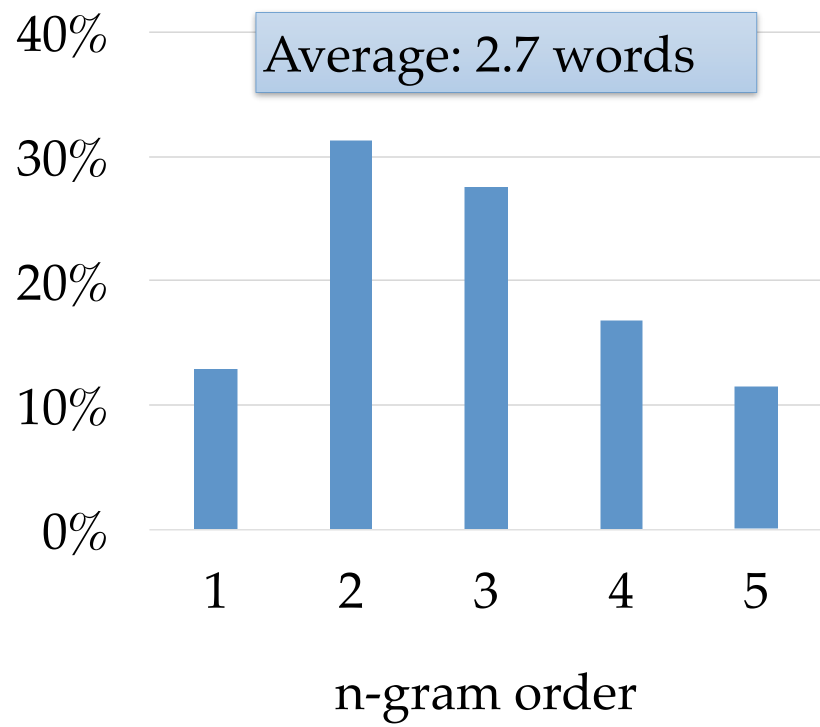
p(progress in the region) =

p(progress) p(in)

p(the) p(region | the)

Train data:

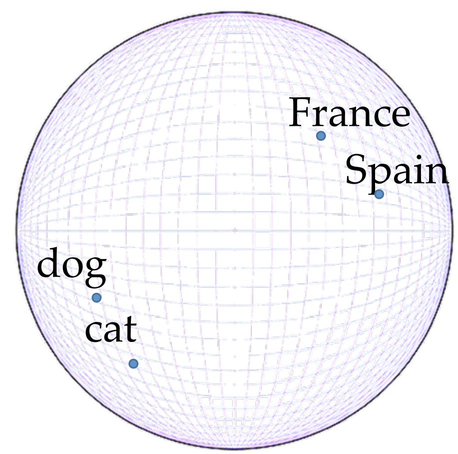
…
development and progress of
the region. in …
…



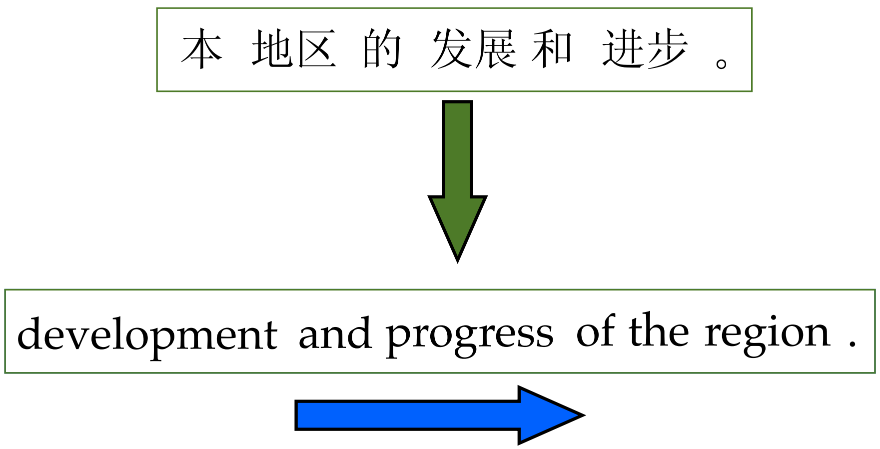
Does not include out-of-vocabulary tokens

How can we improve this?

- Or: how to capture relationships beyond 1.5 to 2.7 words?
- Neural nets: distributional representations make it easier to capture relationships
- Recurrent nets: easy to model variable-length sequences



This talk



Translation modeling

Auli et al., EMNLP 2013; Hu et al., EACL 2014

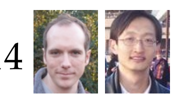
Language modeling

Optimization

Auli & Gao, ACL 2014

Reordering

Auli et al., EMNLP 2014



This talk

本 地区 的 发展 和 进步 。

development and progress of the region .

Translation modeling 

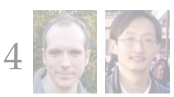
Auli et al., EMNLP 2013; Hu et al., EACL 2014

Language modeling 

Optimization

Auli & Gao, ACL 2014 

Reordering

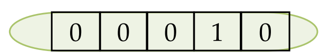
Auli et al., EMNLP 2014 

Feed-forward network

e.g. bigram LM

Feed-forward network

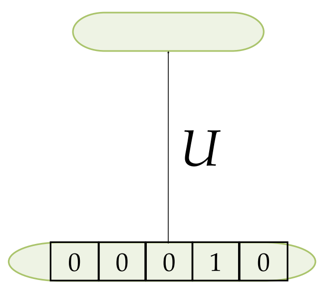
e.g. bigram LM



and

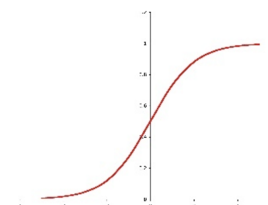
Feed-forward network

e.g. bigram LM



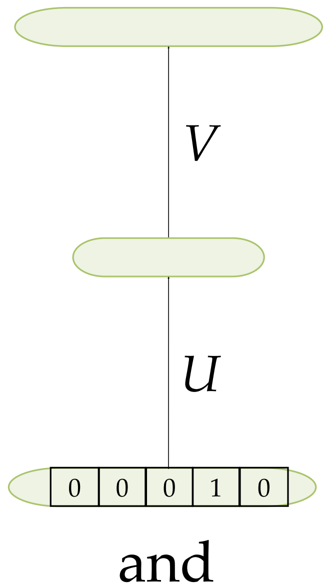
and

$$h_t = \sigma(Ux_t)$$
$$\sigma(z) = \frac{1}{1 + exp\{-z\}}$$



Feed-forward network

e.g. bigram LM

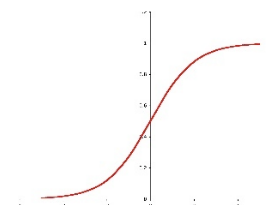


$$y_t = s(Vh_t)$$

$$s(z) = \frac{exp\{z\}}{\sum_{z'} exp\{z'\}}$$

$$h_t = \sigma(Ux_t)$$

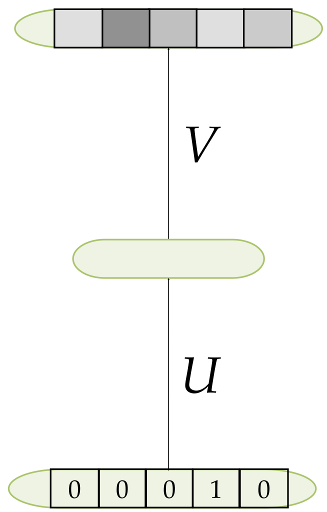
$$\sigma(z) = \frac{1}{1 + exp\{-z\}}$$



Feed-forward network

e.g. bigram LM

p(progress | and)



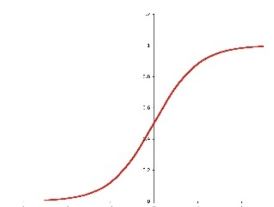
and

$$y_t = s(Vh_t)$$

$$s(z) = \frac{exp\{z\}}{\sum_{z'} exp\{z'\}}$$

$$h_t = \sigma(Ux_t)$$

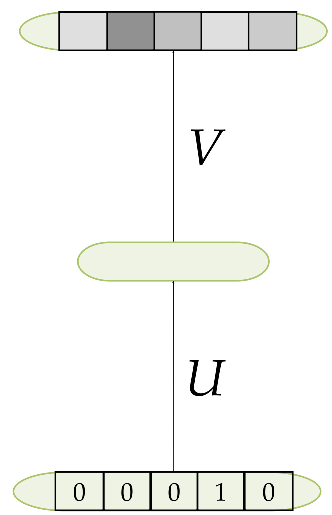
$$\sigma(z) = \frac{1}{1 + exp\{-z\}}$$



Feed-forward network

e.g. bigram LM

p(progress | and)



and

Classification

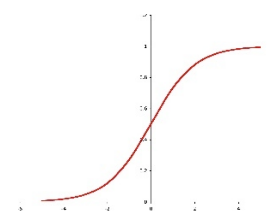
Feature learning

$$y_t = s(Vh_t)$$

$$s(z) = \frac{exp\{z\}}{\sum_{z'} exp\{z'\}}$$

$$h_t = \sigma(Ux_t)$$

$$\sigma(z) = \frac{1}{1 + exp\{-z\}}$$



Feed-forward network

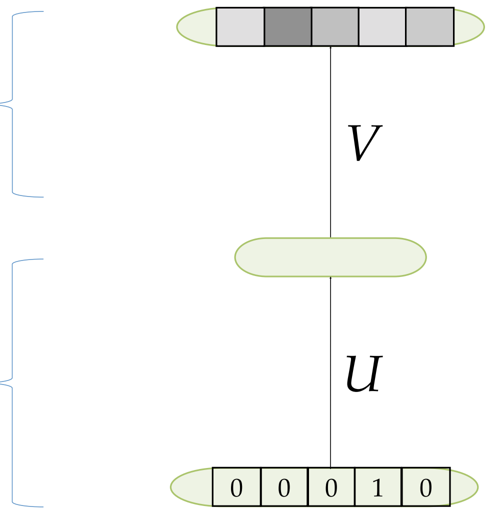
Still based on limited context!

e.g. bigram LM

p(progress | and)

Classification

Feature learning



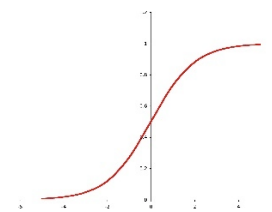
and

$$y_t = s(Vh_t)$$

$$s(z) = \frac{exp\{z\}}{\sum_{z'} exp\{z'\}}$$

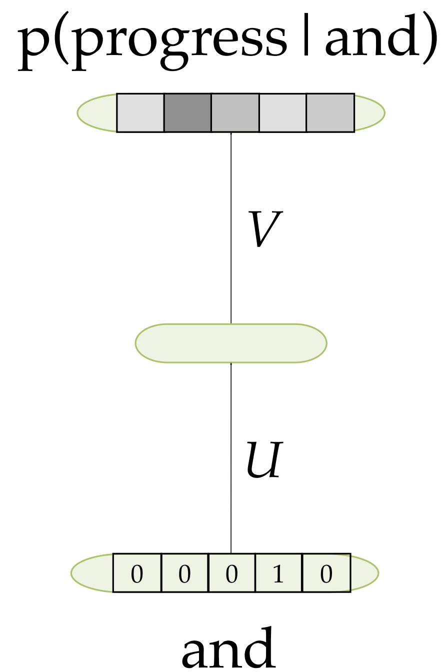
$$h_t = \sigma(Ux_t)$$

$$\sigma(z) = \frac{1}{1 + exp\{-z\}}$$



Recurrent network

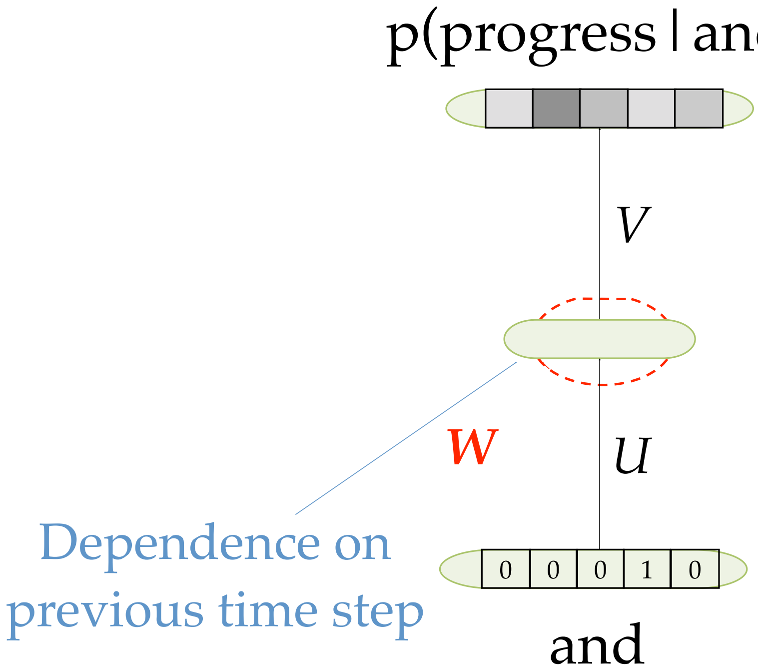
e.g. bigram LM



$$h_t = \sigma(Ux_t)$$

Recurrent network

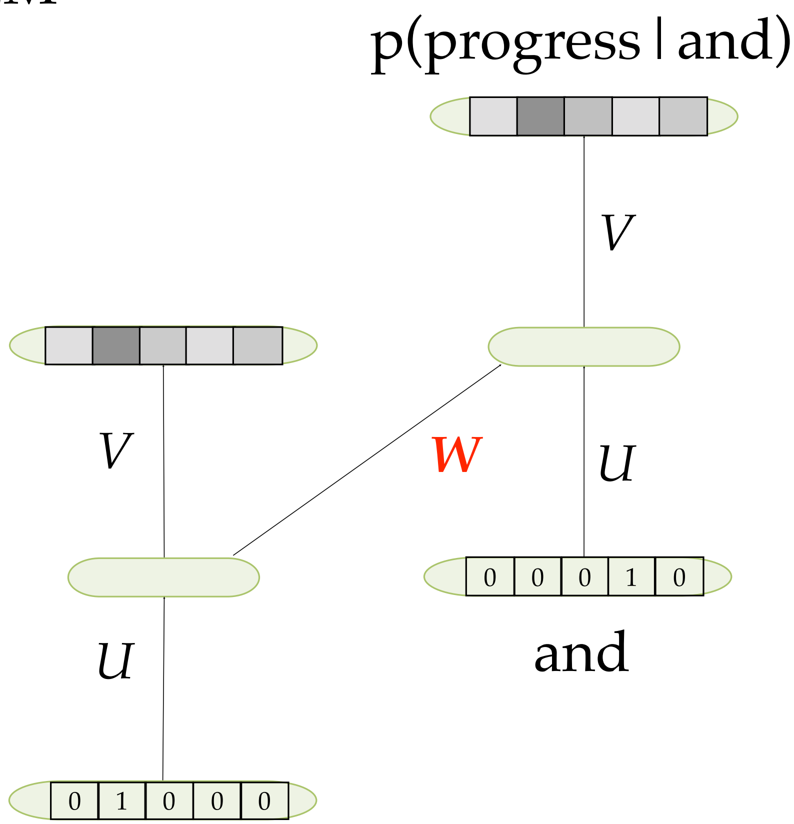
e.g. bigram LM



$$h_t = \sigma(Ux_t)$$

Recurrent network

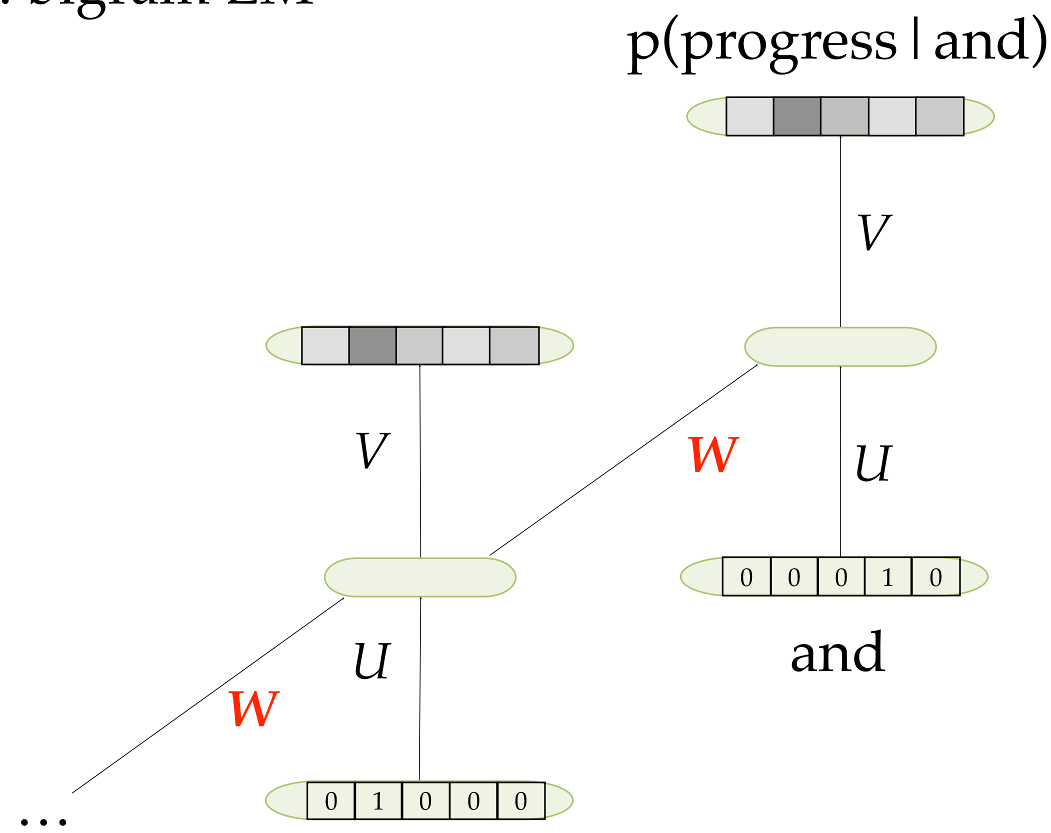
e.g. bigram LM



$$h_t = \sigma(Ux_t + \color{red}{W} h_{t-1})$$

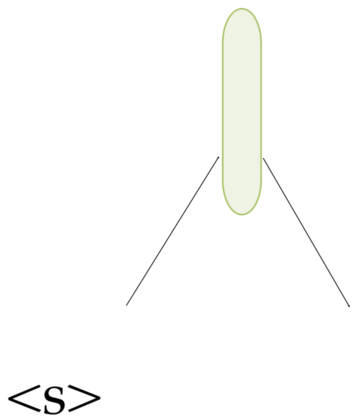
Recurrent network

e.g. bigram LM

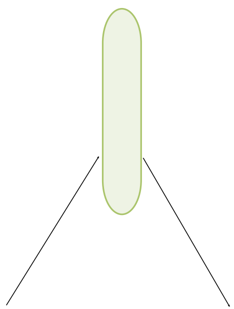


$$h_t = \sigma(Ux_t + {\color{red}W} h_{t-1})$$

Recurrent network

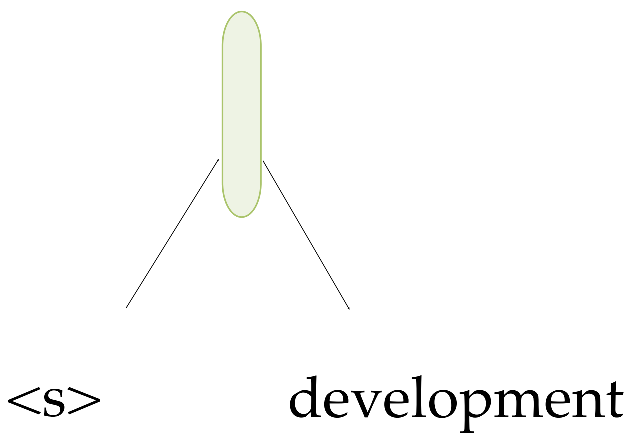


Recurrent network

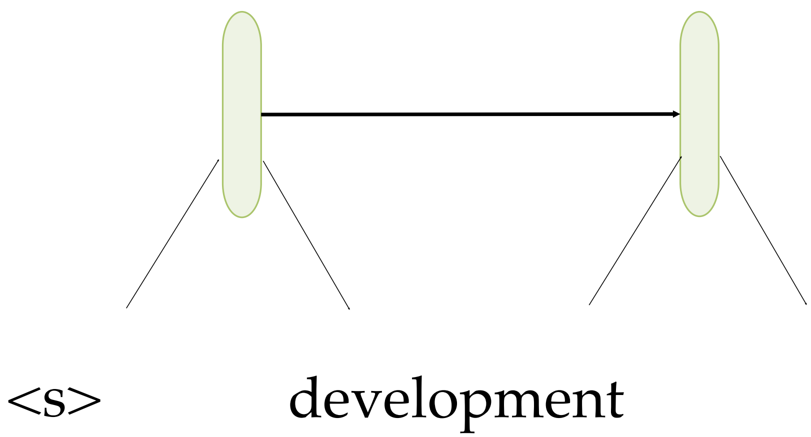


<s> p(development | <s>)

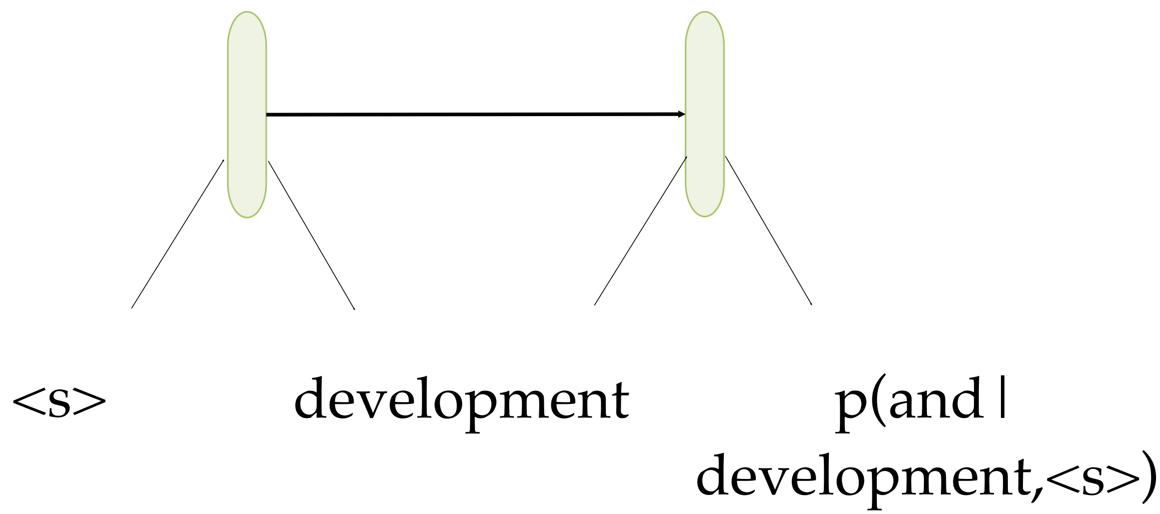
Recurrent network



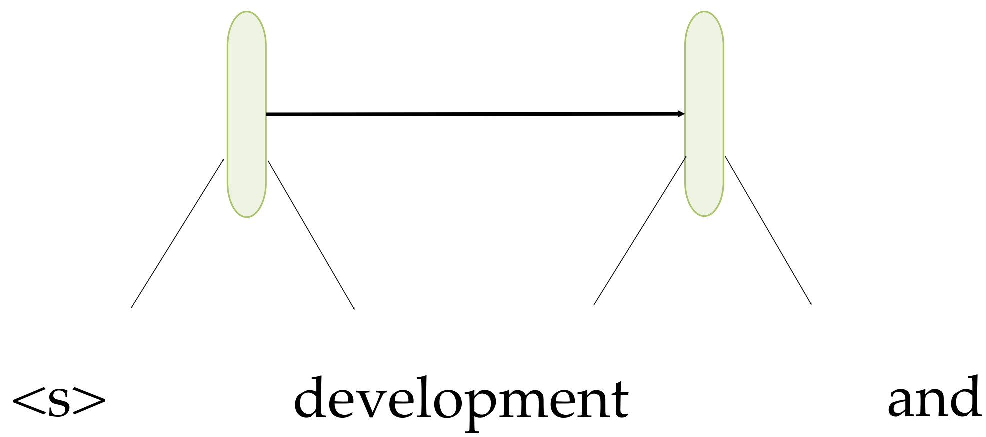
Recurrent network



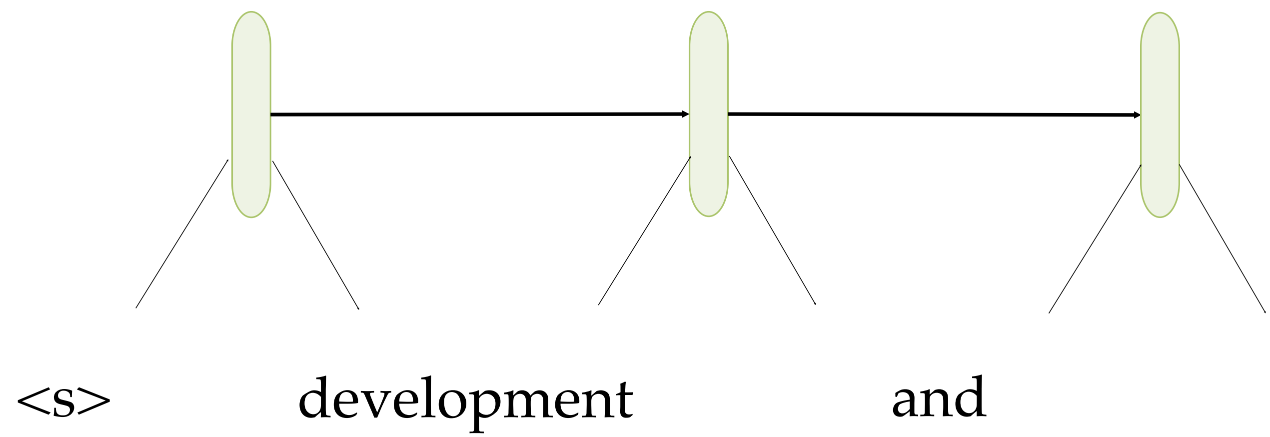
Recurrent network



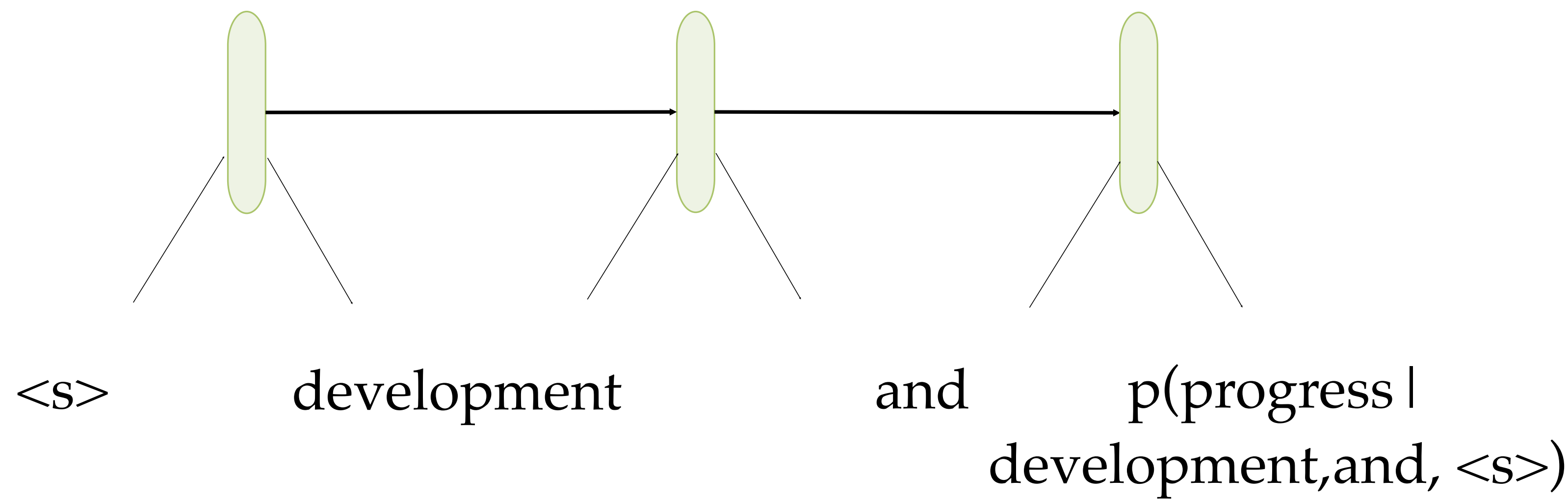
Recurrent network



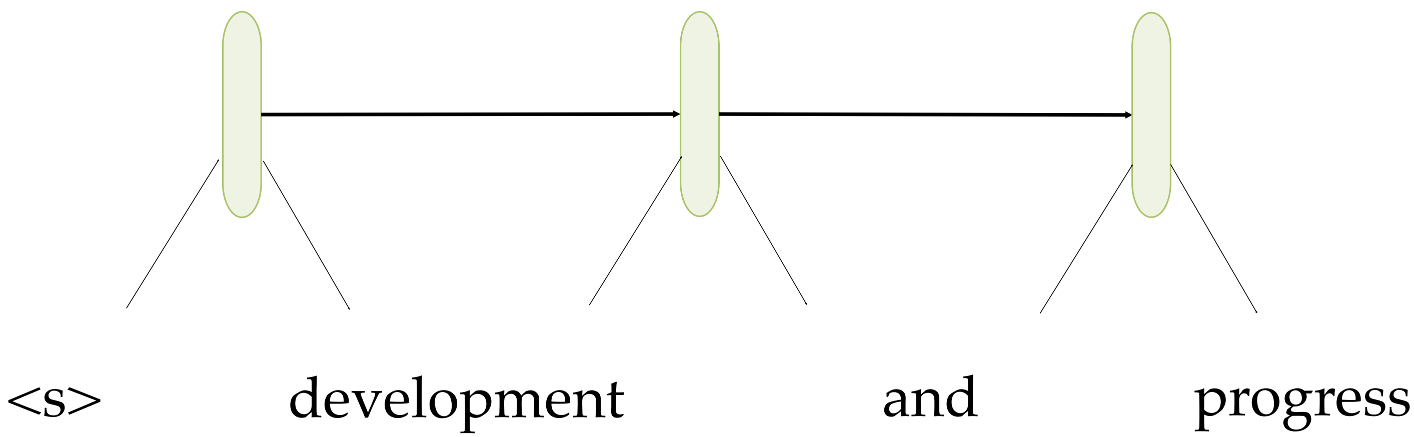
Recurrent network



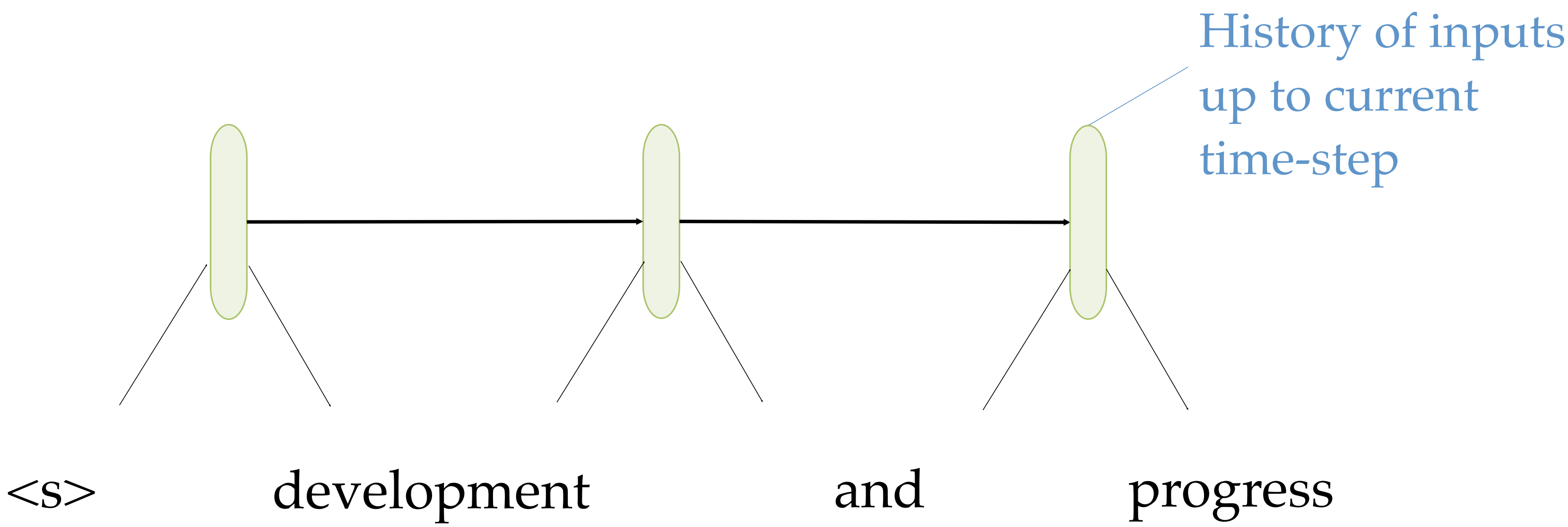
Recurrent network



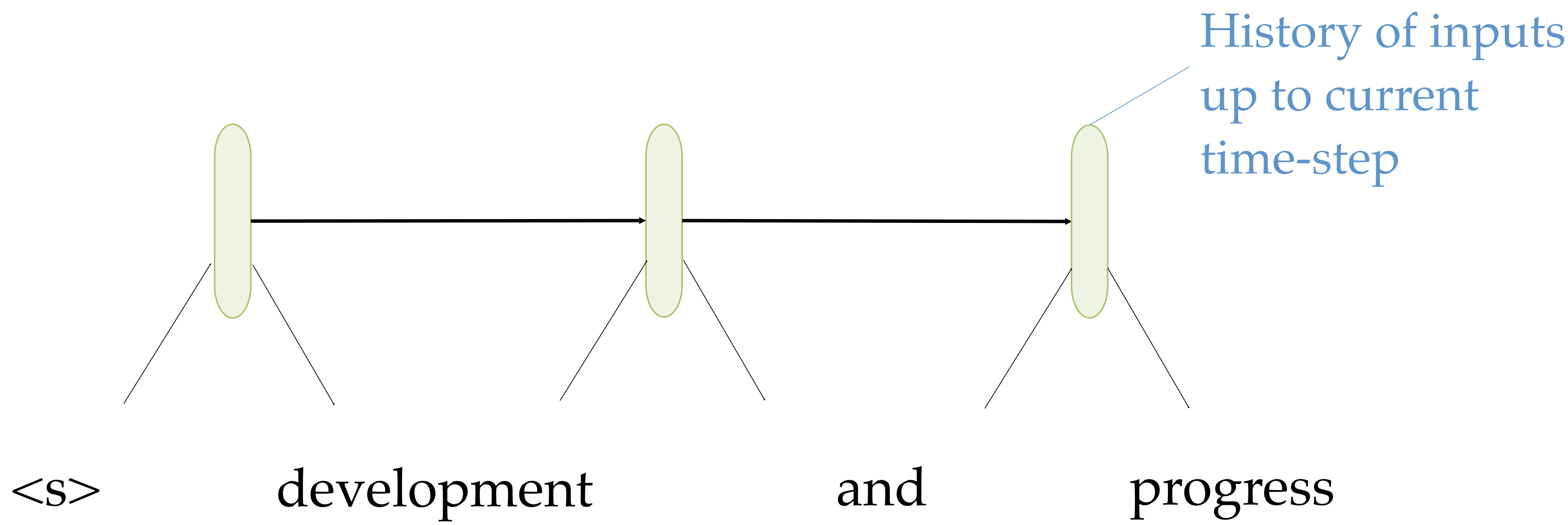
Recurrent network



Recurrent network



Recurrent network



State of the art in language modeling (Mikolov 2011)

More accurate than feed-forward nets (Sundermeyer 2013)

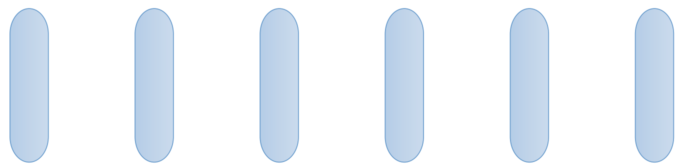
Combined language and translation model

本 地区 的 发展 和 进步

Auli et al., EMNLP 2013

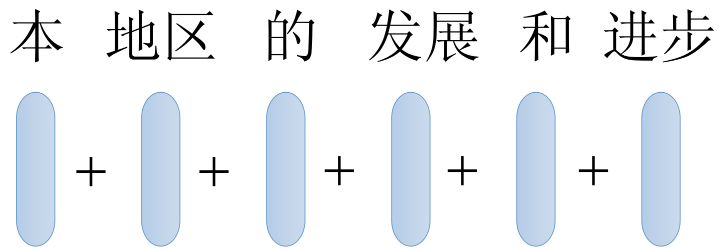
Combined language and translation model

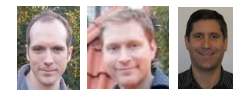
本 地区 的 发展 和 进步



Auli et al., EMNLP 2013

Combined language and translation model

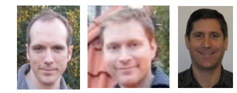




Auli et al., EMNLP 2013

Combined language and translation model

本 地区 的 发展 和 进步

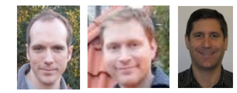


Auli et al., EMNLP 2013

Entire source sentence representation

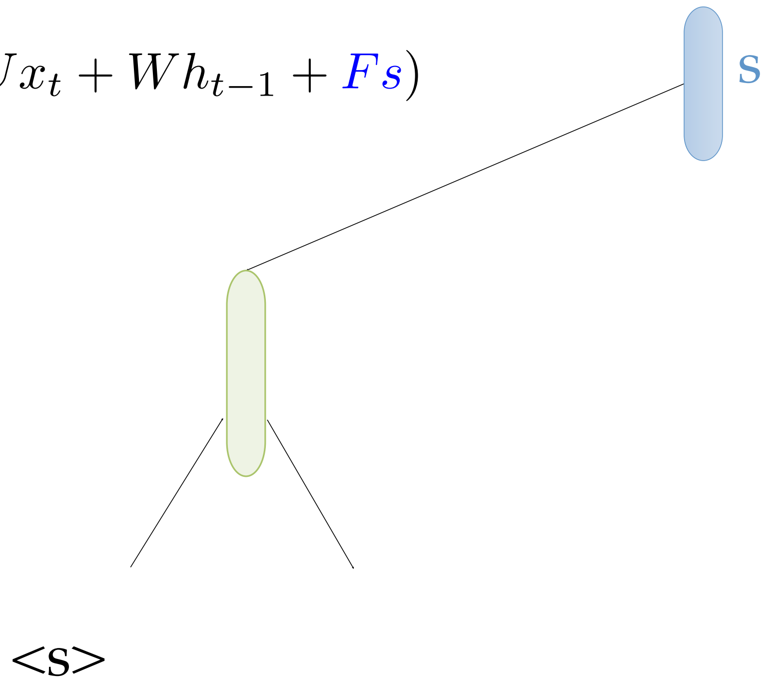
Combined language and translation model

本 地区 的 发展 和 进步



Auli et al., EMNLP 2013

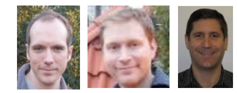
$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Entire source sentence representation

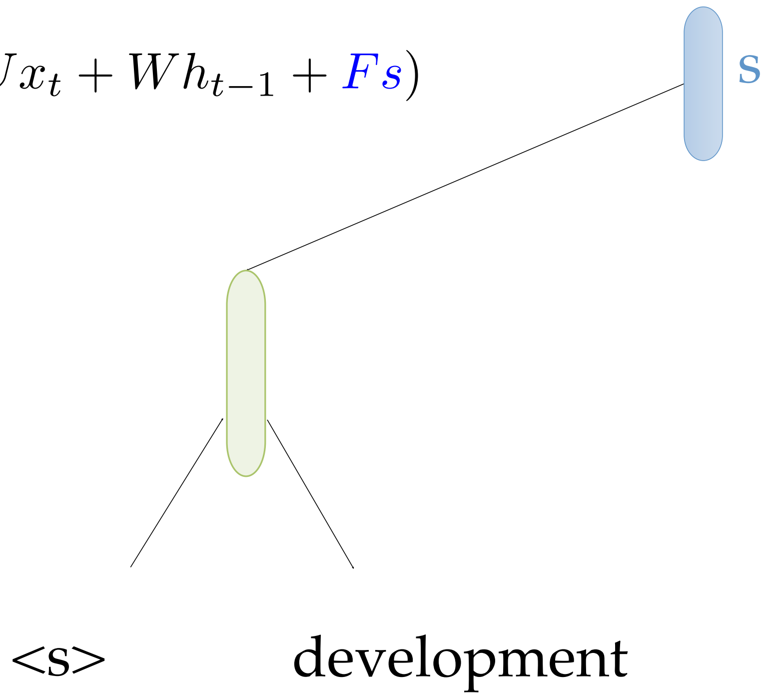
Combined language and translation model

本 地区 的 发展 和 进步



Auli et al., EMNLP 2013

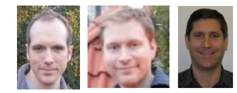
$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Entire source sentence representation

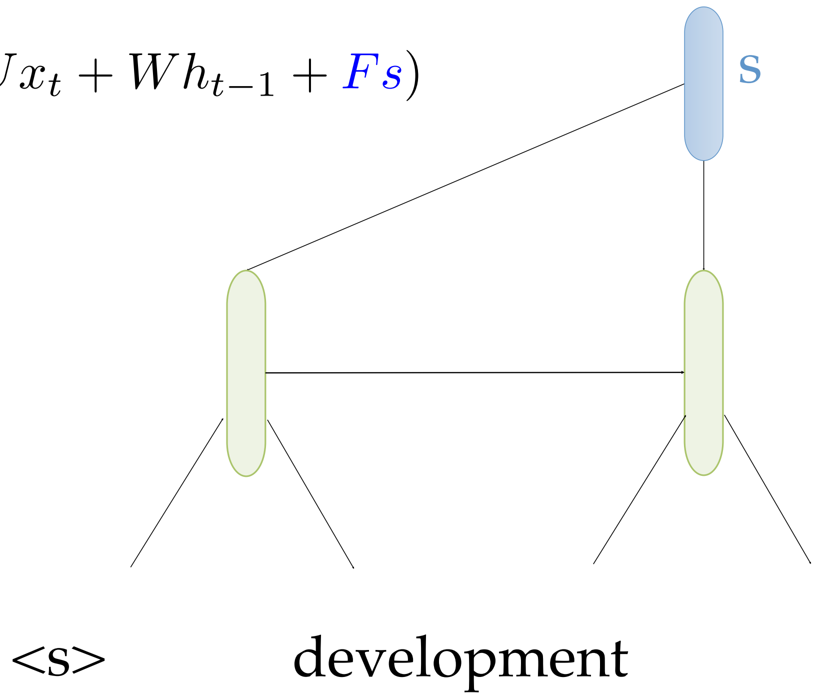
Combined language and translation model

本 地区 的 发展 和 进步



Auli et al., EMNLP 2013

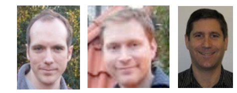
$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Entire source sentence representation

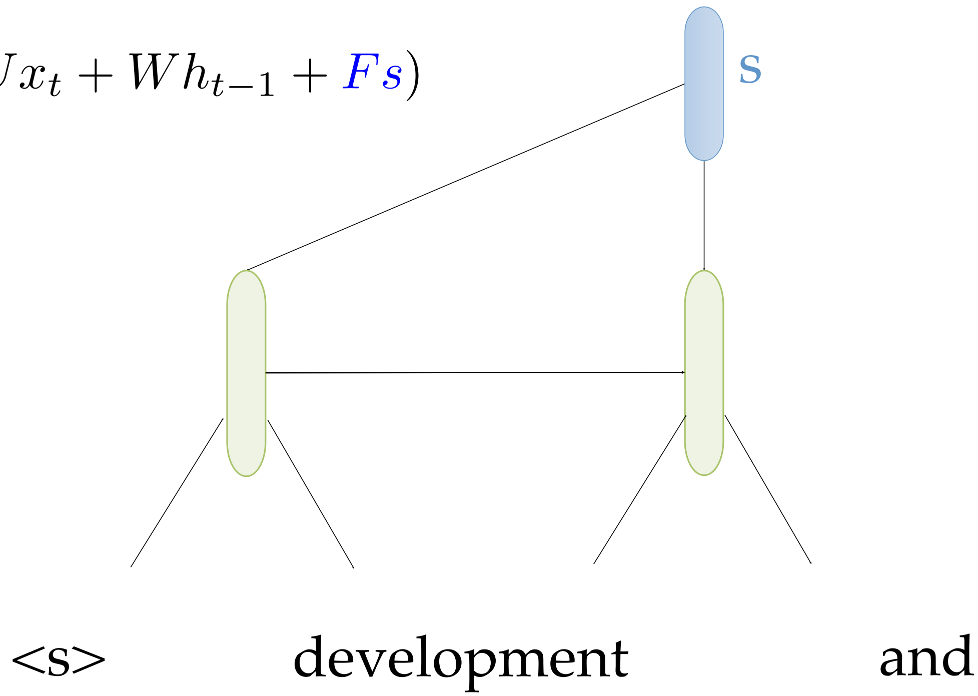
Combined language and translation model

本 地区 的 发展 和 进步



Auli et al., EMNLP 2013

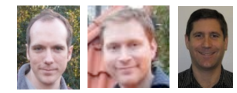
$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Entire source sentence representation

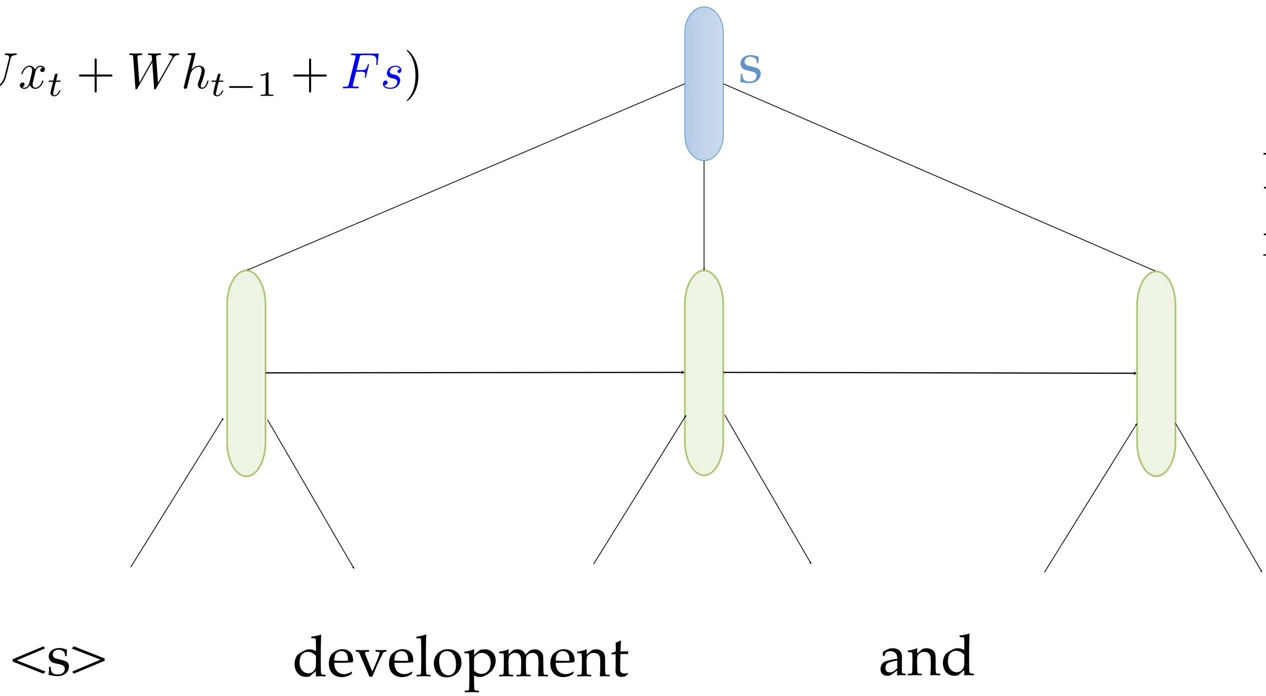
Combined language and translation model

本 地区 的 发展 和 进步



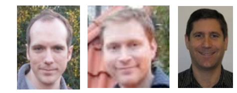
Auli et al., EMNLP 2013

$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$



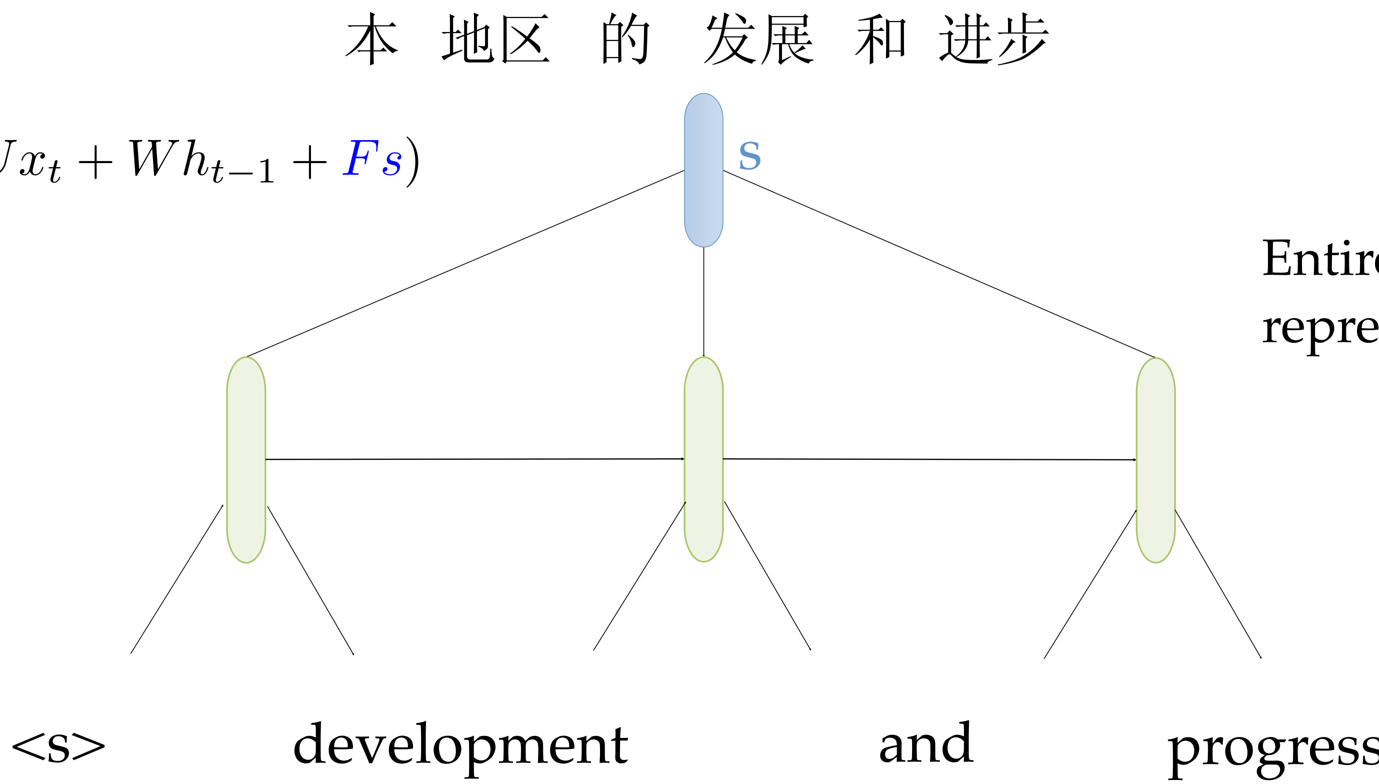
Entire source sentence representation

Combined language and translation model



Auli et al., EMNLP 2013

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Entire source sentence representation

Combined language and translation model

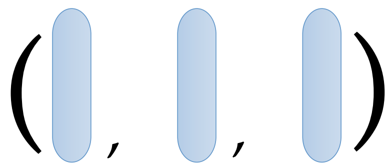
本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$

Combined language and translation model

本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$

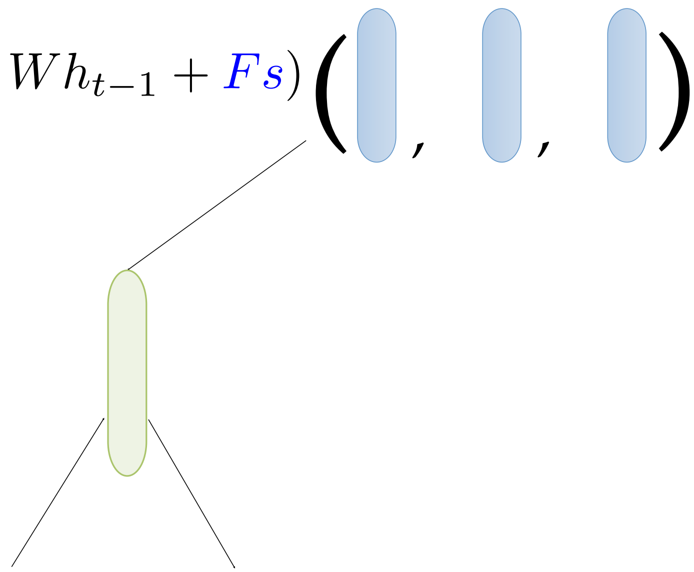


Source word-window

Combined language and translation model

本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$

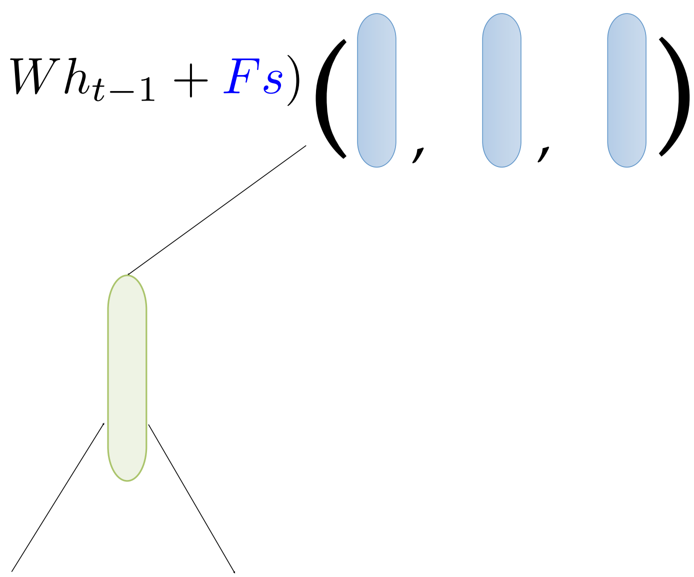


Source word-window

<s>

Combined language and translation model

$$h_t = \sigma(Ux_t + Wh_{t-1} + {\color{blue}Fs})$$

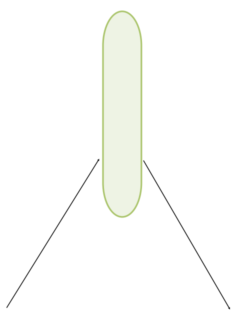


Combined language and translation model

本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$

Source word-window

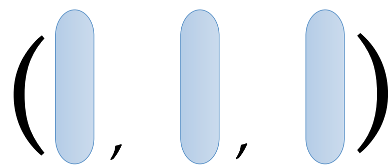


<s> development

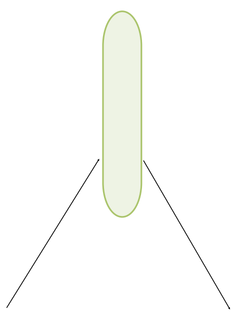
Combined language and translation model

本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



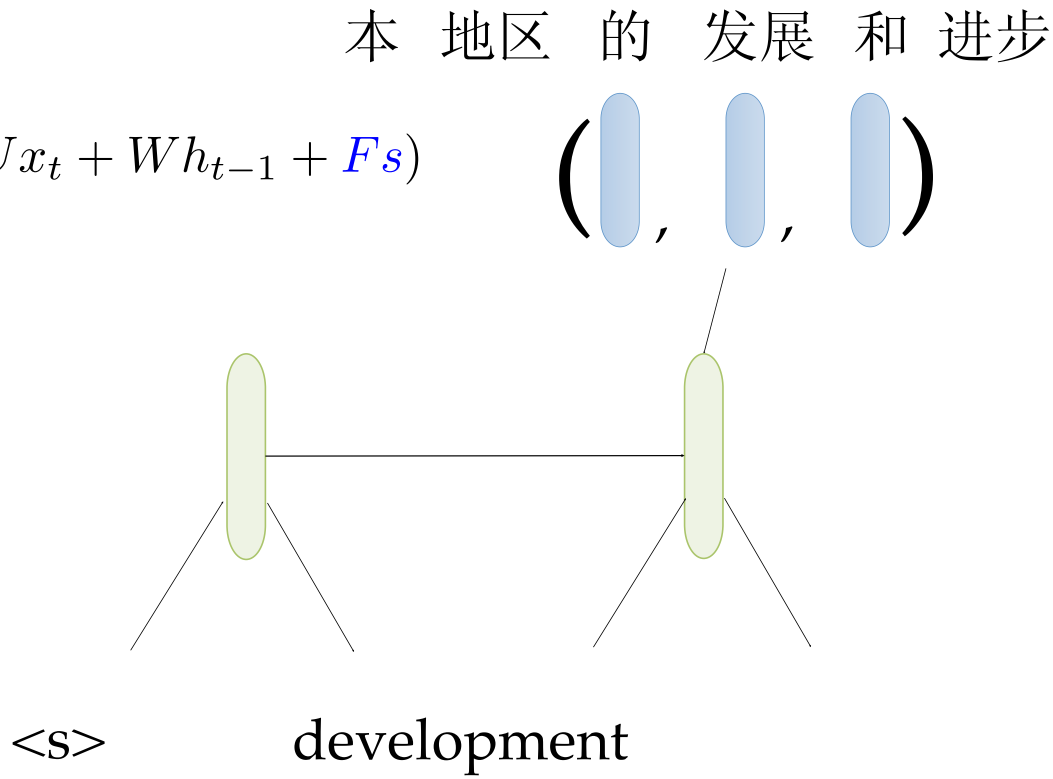
Source word-window



<s> development

Combined language and translation model

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$

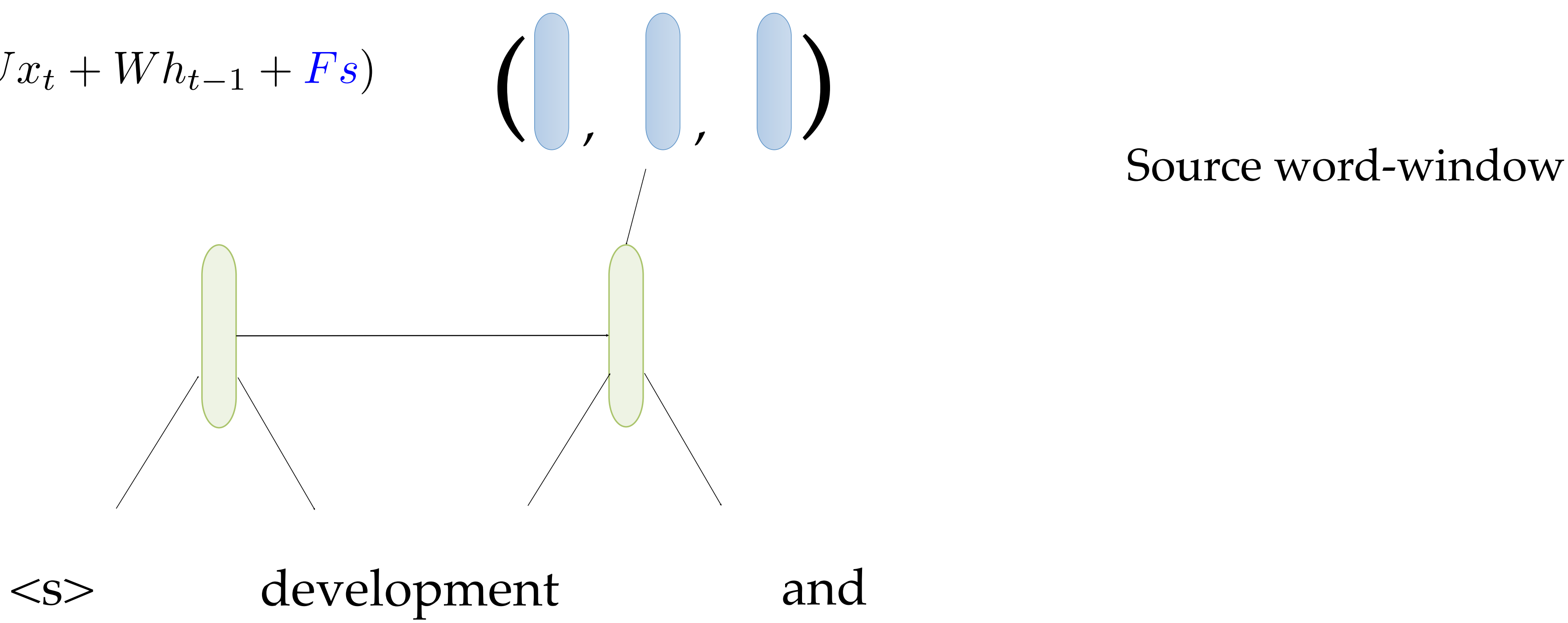


Source word-window

Combined language and translation model

本 地区 的 发展 和 进步

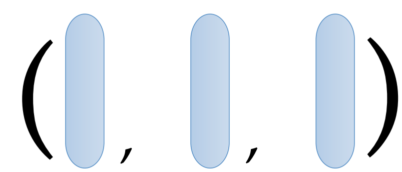
$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



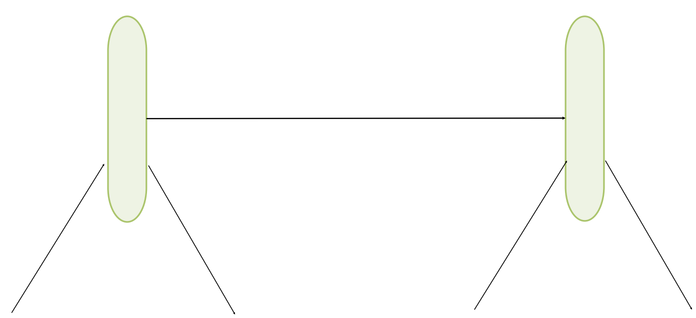
Combined language and translation model

本 地区 的 发展 和 进步

$$h_t = \sigma(Ux_t + Wh_{t-1} + {\color{blue}Fs})$$



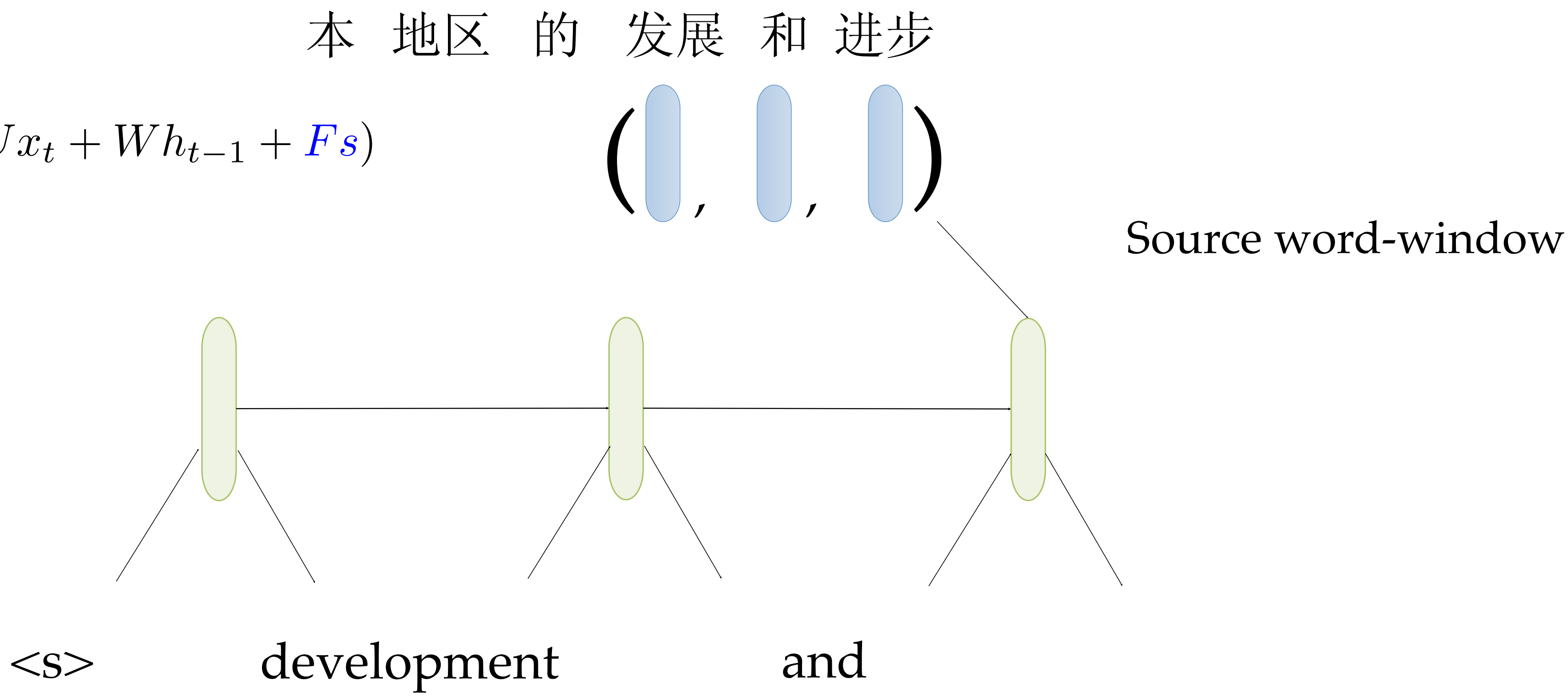
Source word-window



<s> development and

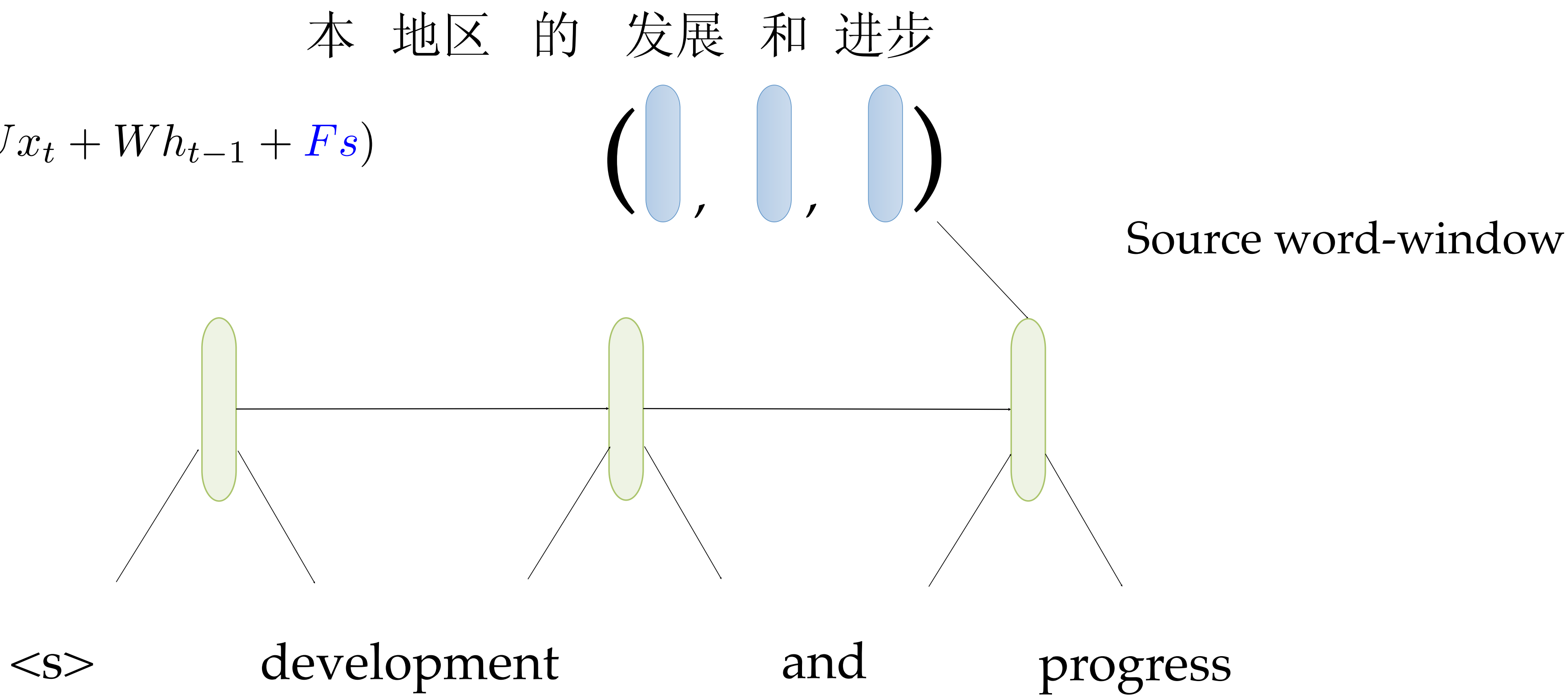
Combined language and translation model

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



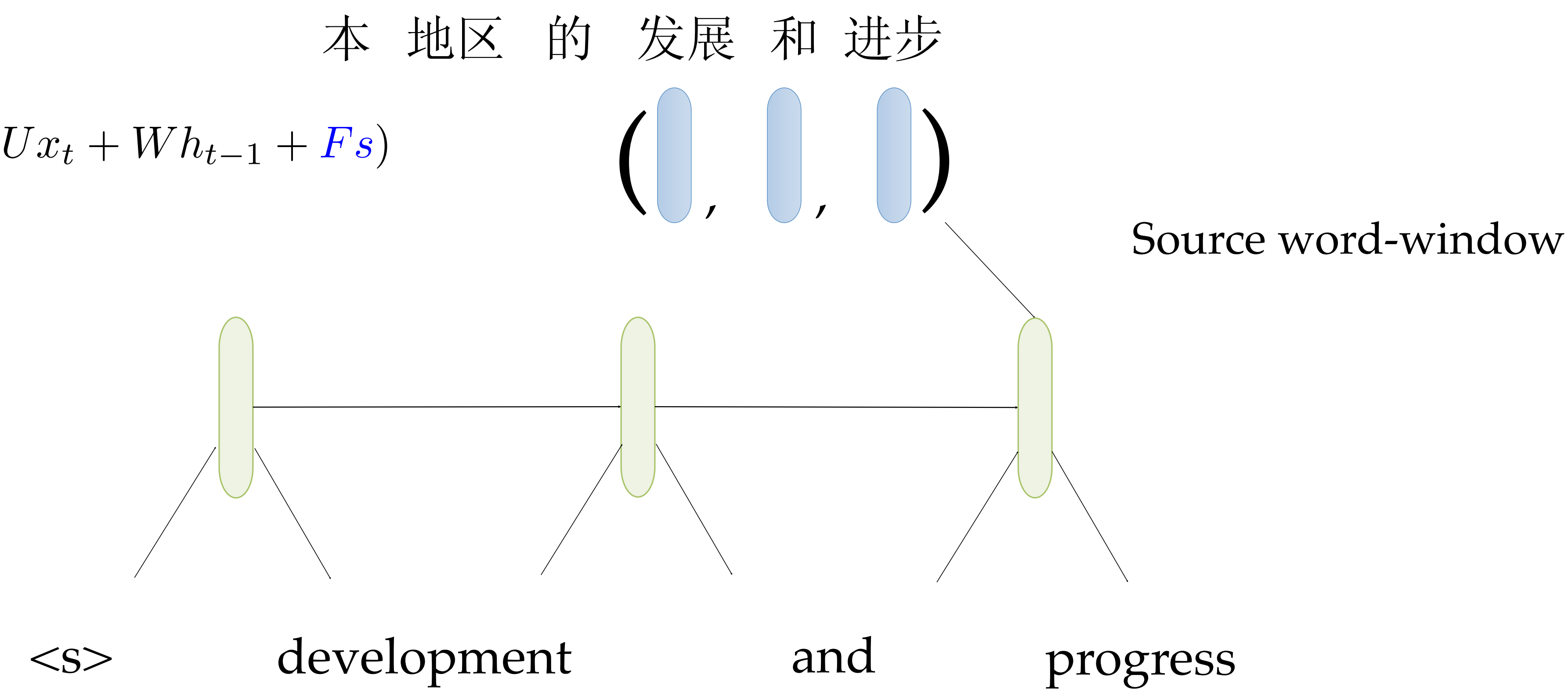
Combined language and translation model

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Combined language and translation model

$$h_t = \sigma(Ux_t + Wh_{t-1} + Fs)$$



Le (2012), Kalchbrenner (2013), Devlin (2014), Cho (2014), Sutskever (2014)

Experimental setup

- WMT 2012 French-English translation task
- Data: 100M words
- Baseline: Phrase-based model similar to Moses
- Rescoring
- Mini-batch gradient descent
- Class-structured output layer (Goodman, 1996)

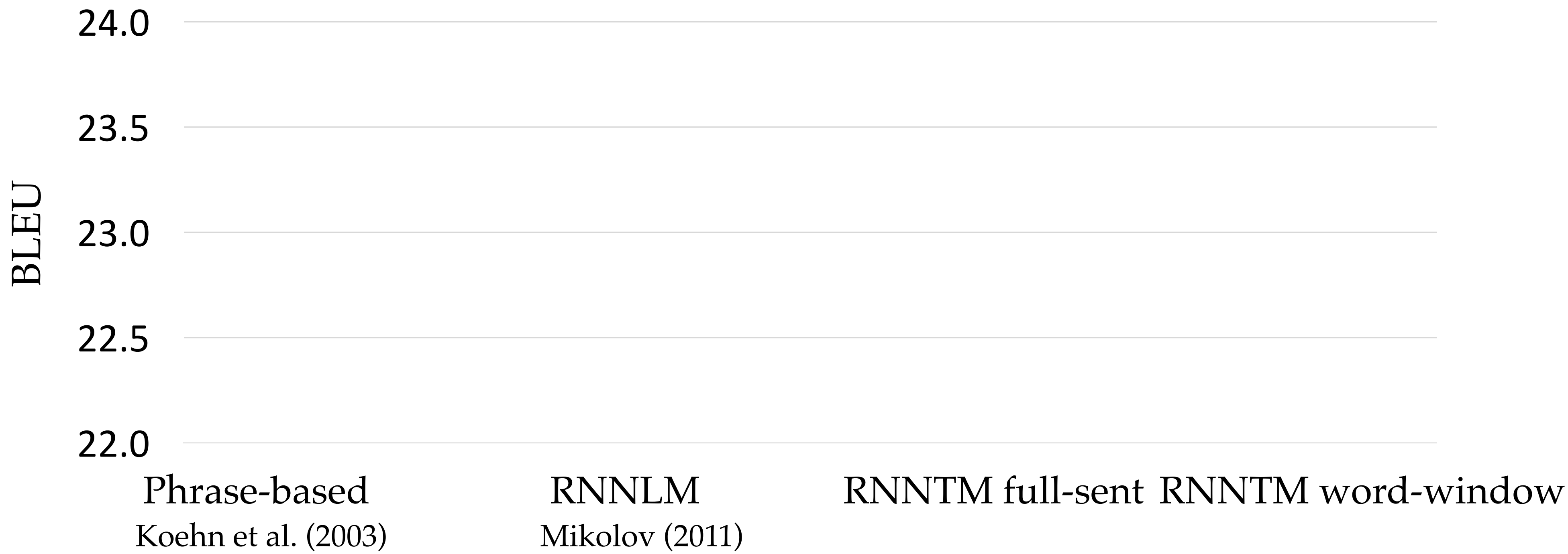
Does the neural model learn to translate?

-Discrete translation model

WMT 2012 French-English, 100M words, phrase-based baseline, n-best rescoring

Does the neural model learn to translate?

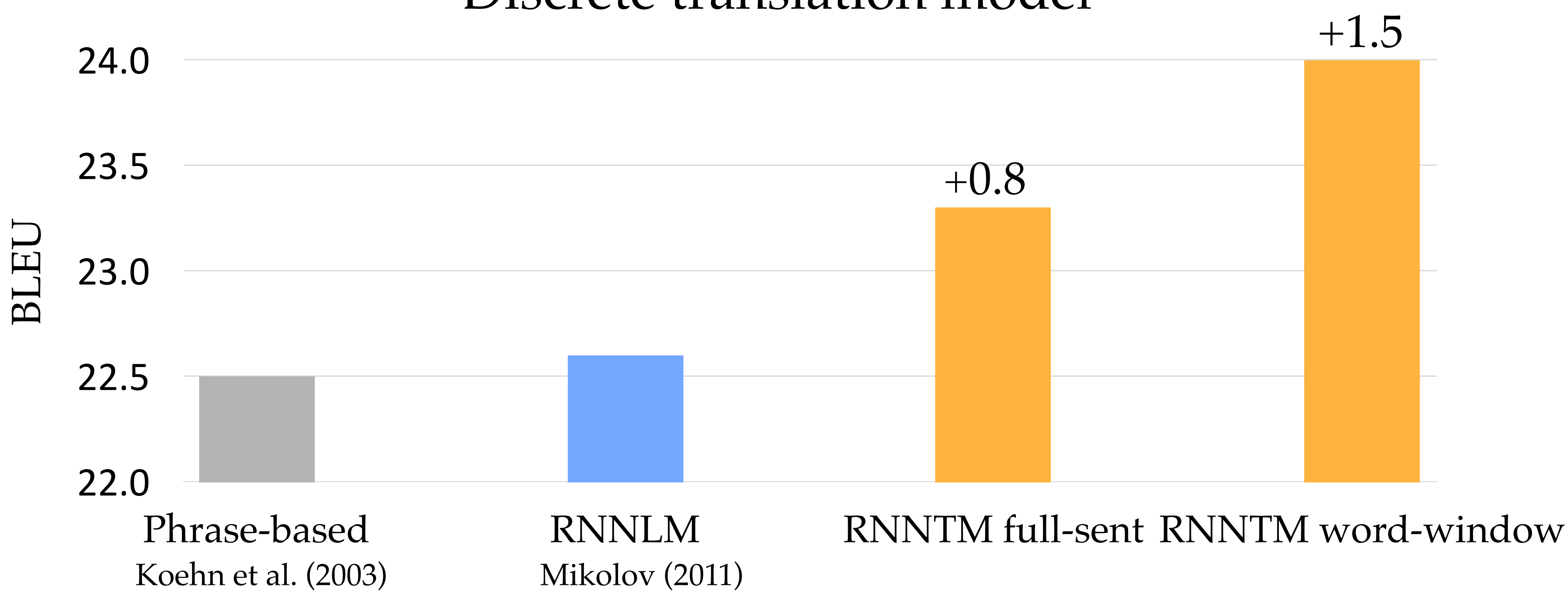
-Discrete translation model



WMT 2012 French-English, 100M words, phrase-based baseline, n-best rescoring

Does the neural model learn to translate?

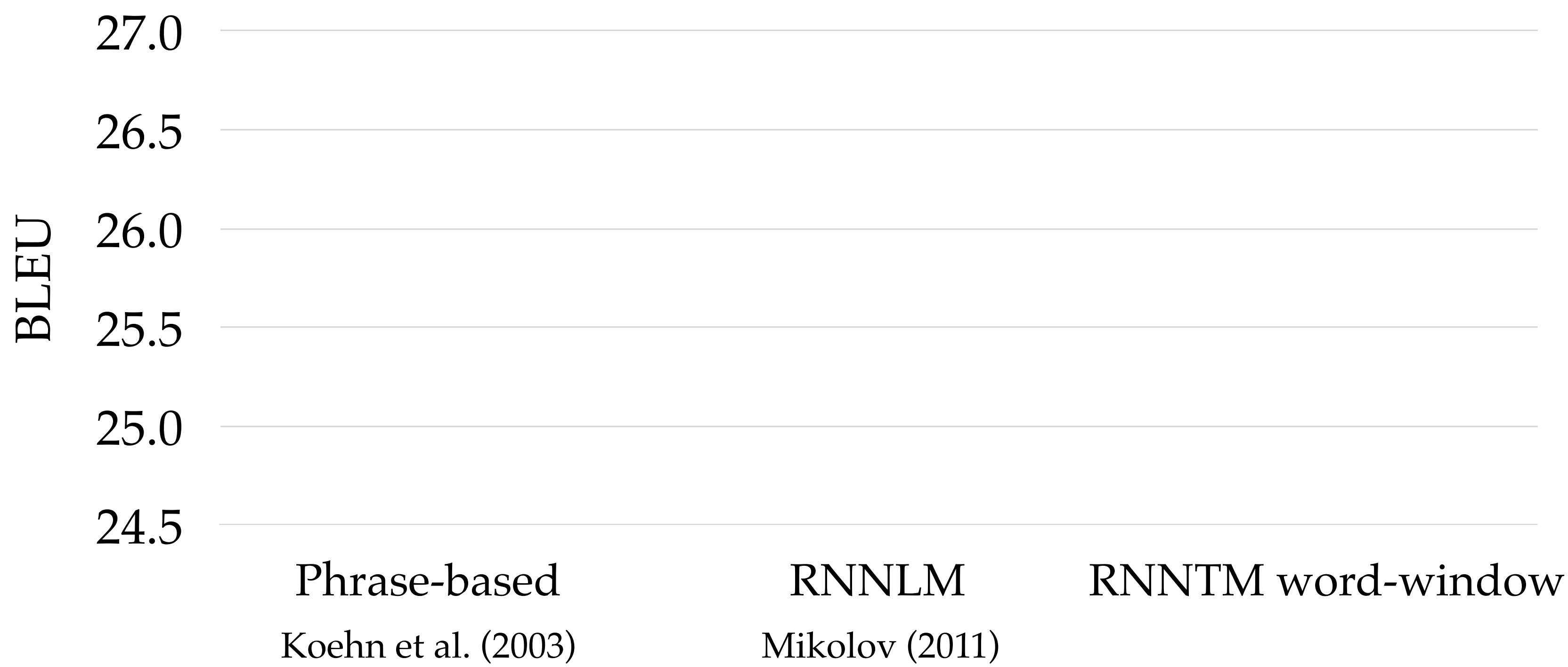
-Discrete translation model



WMT 2012 French-English, 100M words, phrase-based baseline, n-best rescoring

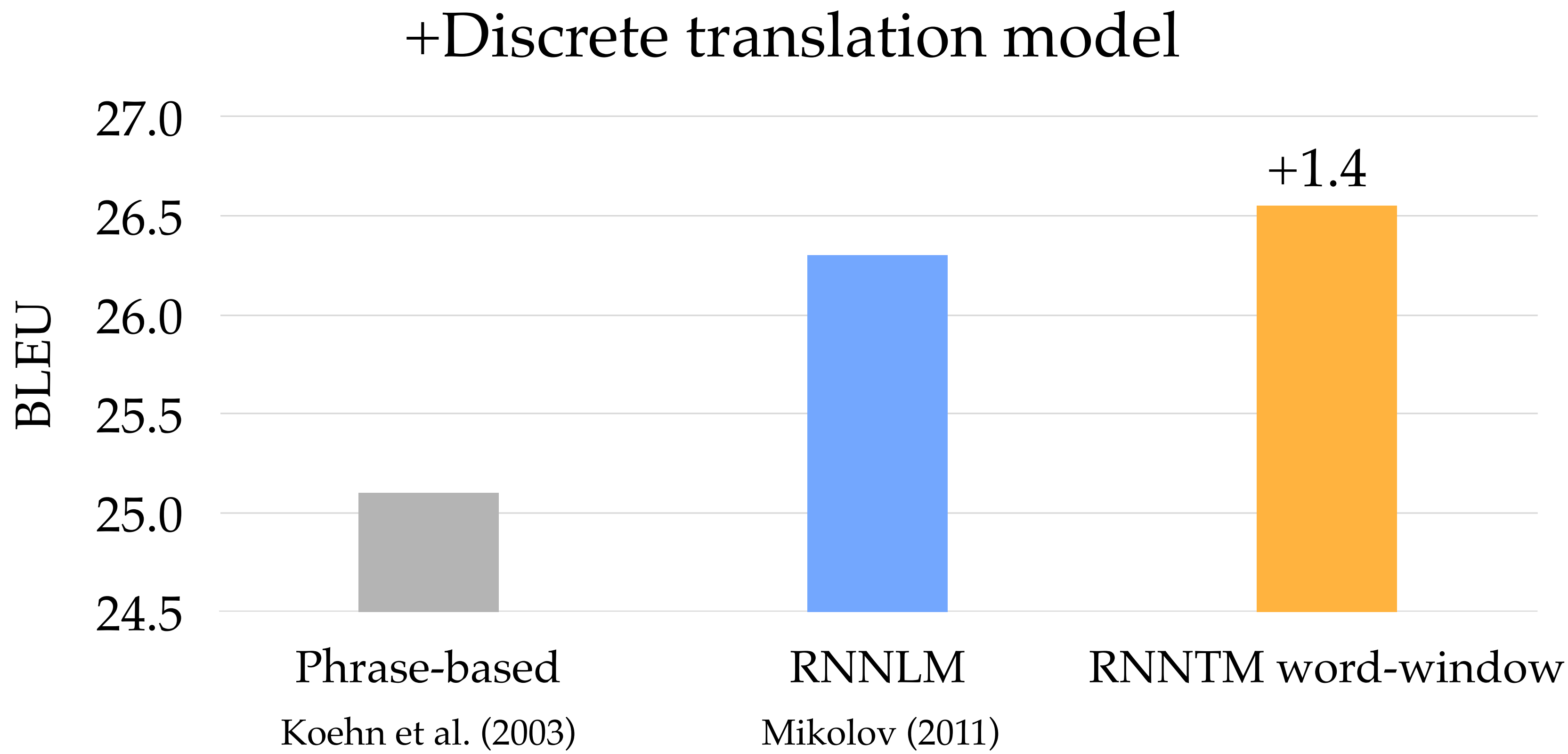
Improving a phrase-based baseline

+Discrete translation model



WMT 2012 French-English, 100M words, phrase-based baseline, lattice rescoring

Improving a phrase-based baseline



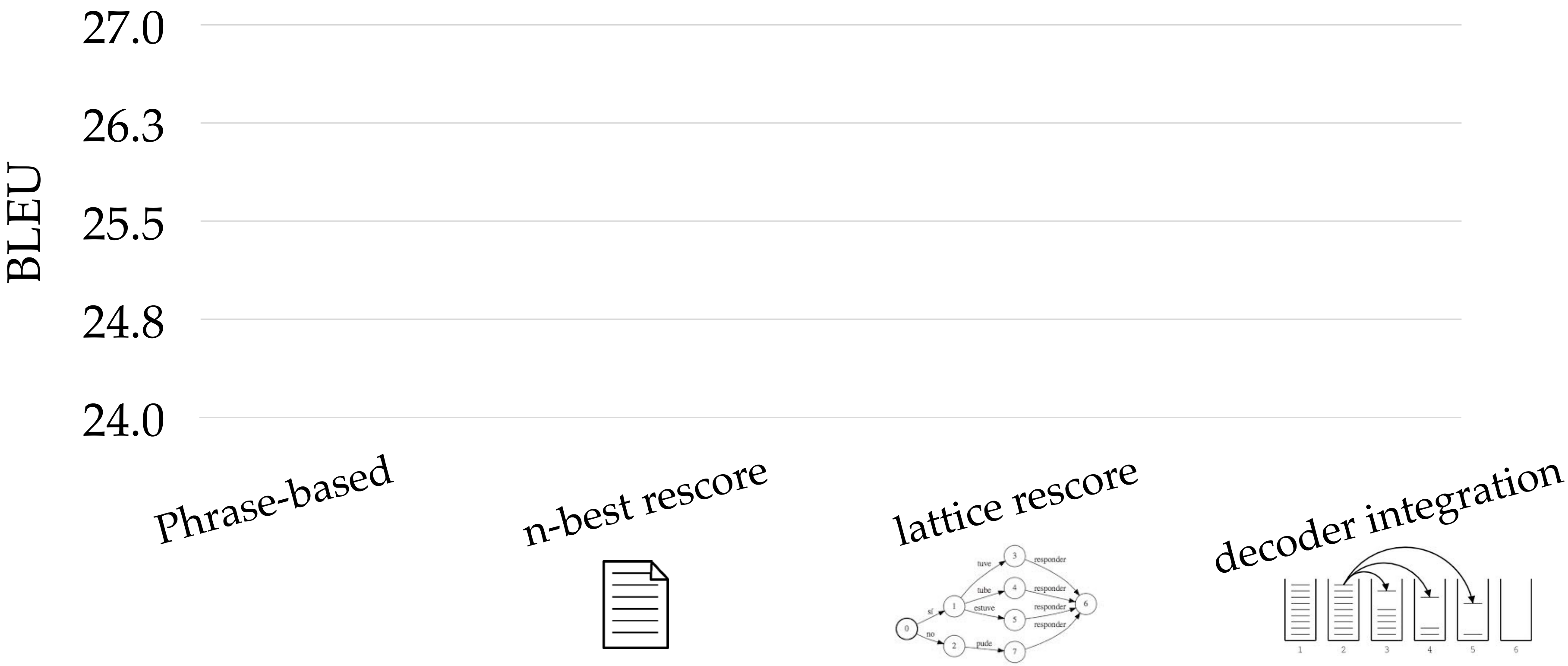
WMT 2012 French-English, 100M words, phrase-based baseline, lattice rescoring

Combining recurrent nets with discrete models

Auli & Gao, ACL 2014 

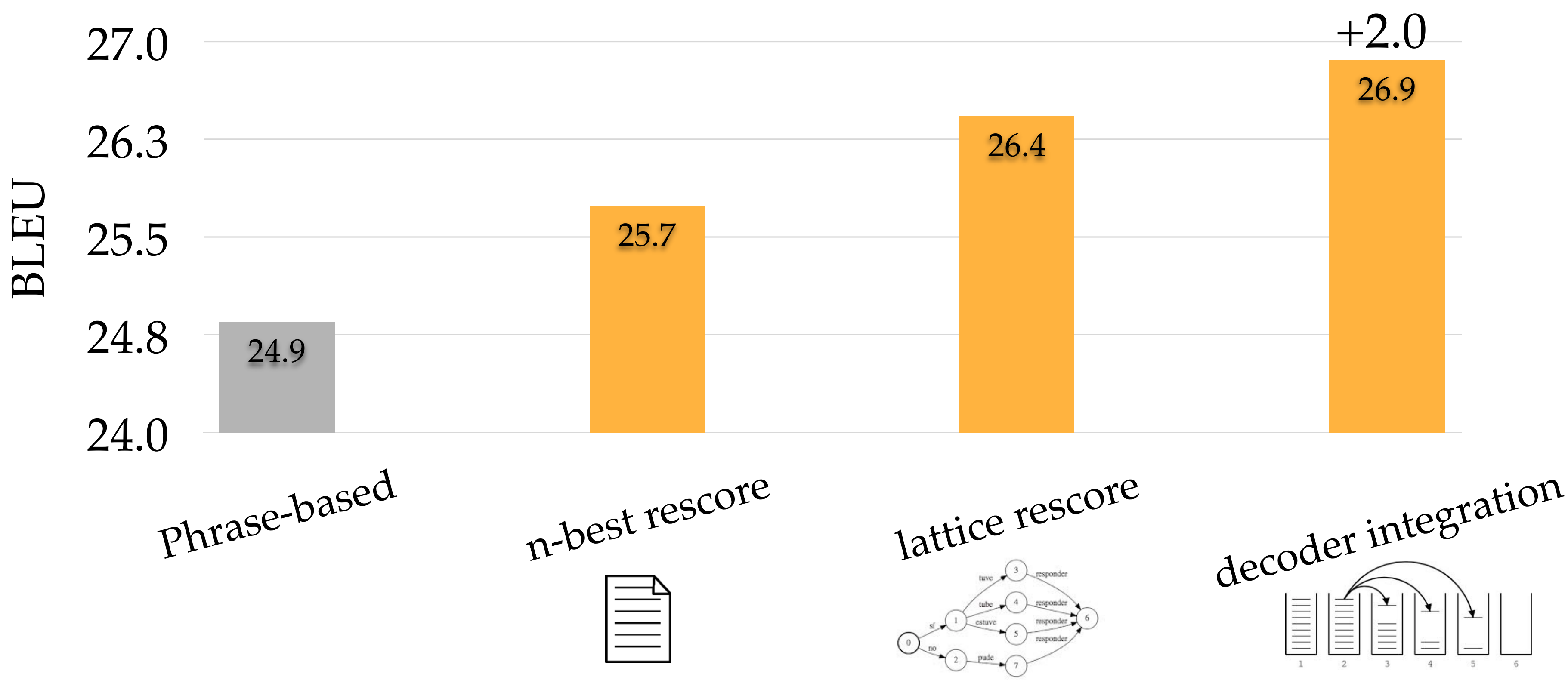
Combining recurrent nets with discrete models

Auli & Gao, ACL 2014 

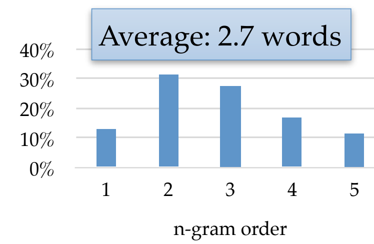


Combining recurrent nets with discrete models

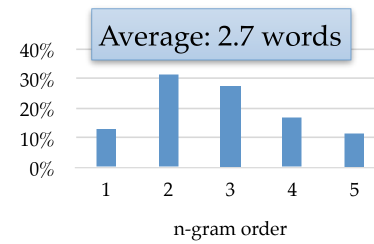
Auli & Gao, ACL 2014



Neural nets vs discrete models

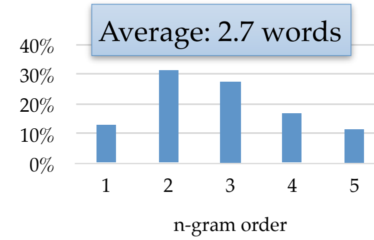


Neural nets vs discrete models



- Neural models more robust when discrete models rely on sparse estimates?

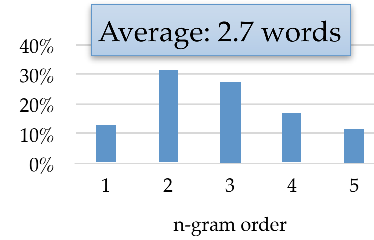
Neural nets vs discrete models



- Neural models more robust when discrete models rely on sparse estimates?
- Language modeling: n-gram LM and neural net LM two components of log-linear model of translation

$$\hat{e} = argmax_e \sum_i \lambda_i h_i(f, e) \qquad h_1(f, e) = log\ p_{lm}(e) \qquad h_2(f, e) = log\ p_{rnn}(e)$$

Neural nets vs discrete models



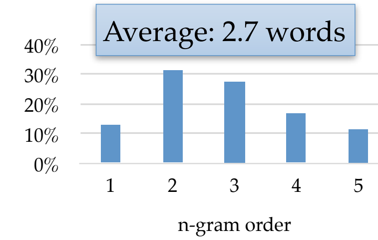
- Neural models more robust when discrete models rely on sparse estimates?
- Language modeling: n-gram LM and neural net LM two components of log-linear model of translation

$$\hat{e} = argmax_e \sum_i \lambda_i h_i(f,e) \qquad h_1(f,e) = log\ p_{lm}(e) \qquad h_2(f,e) = log\ p_{rnn}(e)$$

- Split each model into five features, one for each n-gram order s.t.

$$log\ p_{lm}(e) = \sum_{i=1}^{5} h_i(f,e) \qquad log\ p_{rnn}(e) = \sum_{i=6}^{10} h_i(f,e)$$

Neural nets vs discrete models



- Neural models more robust when discrete models rely on sparse estimates?
- Language modeling: n-gram LM and neural net LM two components of log-linear model of translation

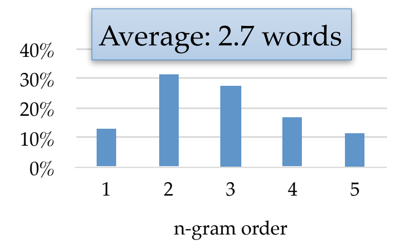
$$\hat{e} = argmax_e \sum_i \lambda_i h_i(f,e) \qquad h_1(f,e) = log\ p_{lm}(e) \qquad h_2(f,e) = log\ p_{rnn}(e)$$

- Split each model into five features, one for each n-gram order s.t.

$$log\ p_{lm}(e) = \sum_{i=1}^{5} h_i(f,e) \qquad log\ p_{rnn}(e) = \sum_{i=6}^{10} h_i(f,e)$$

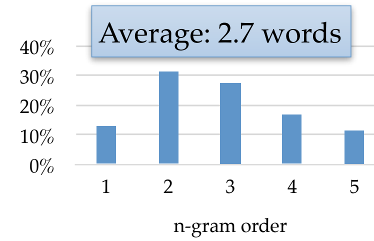
- Standard optimizer (MERT) to find weights for each n-gram order

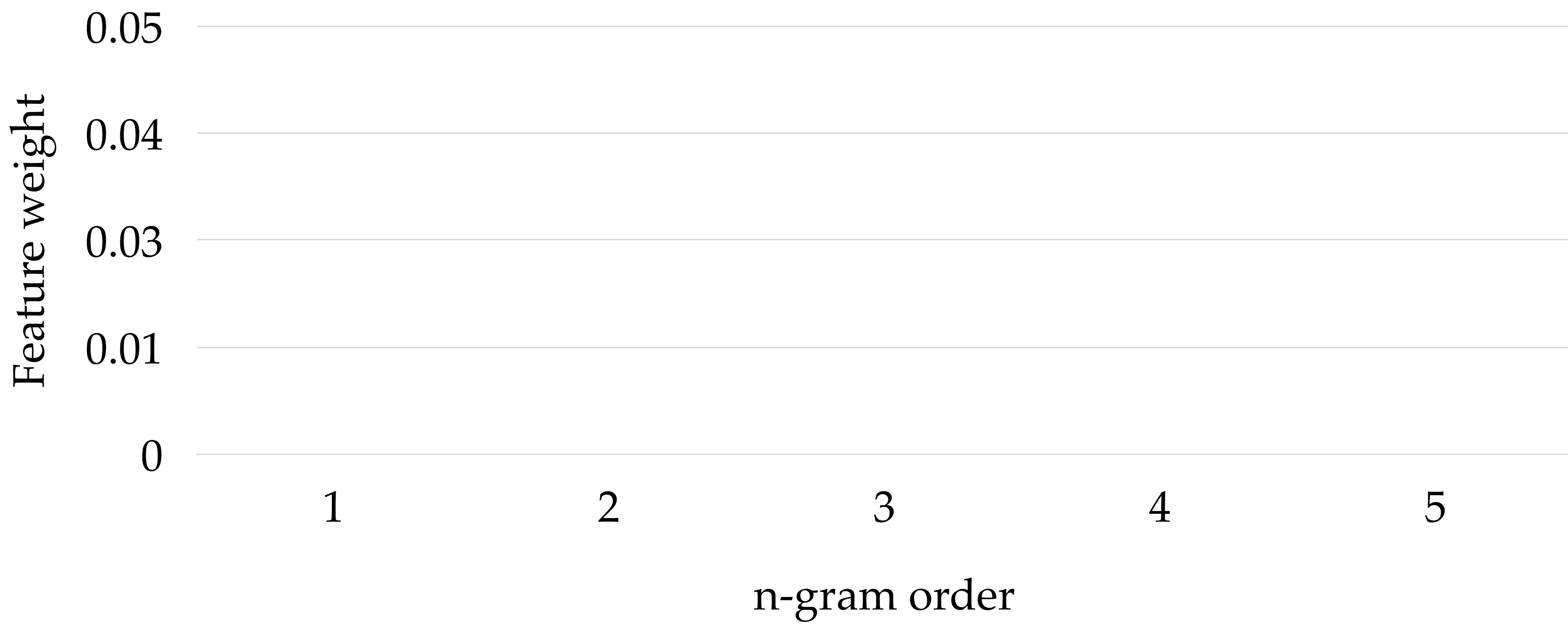
Neural nets vs discrete models



Optimzer: MERT

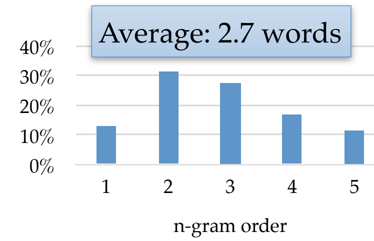
Neural nets vs discrete models

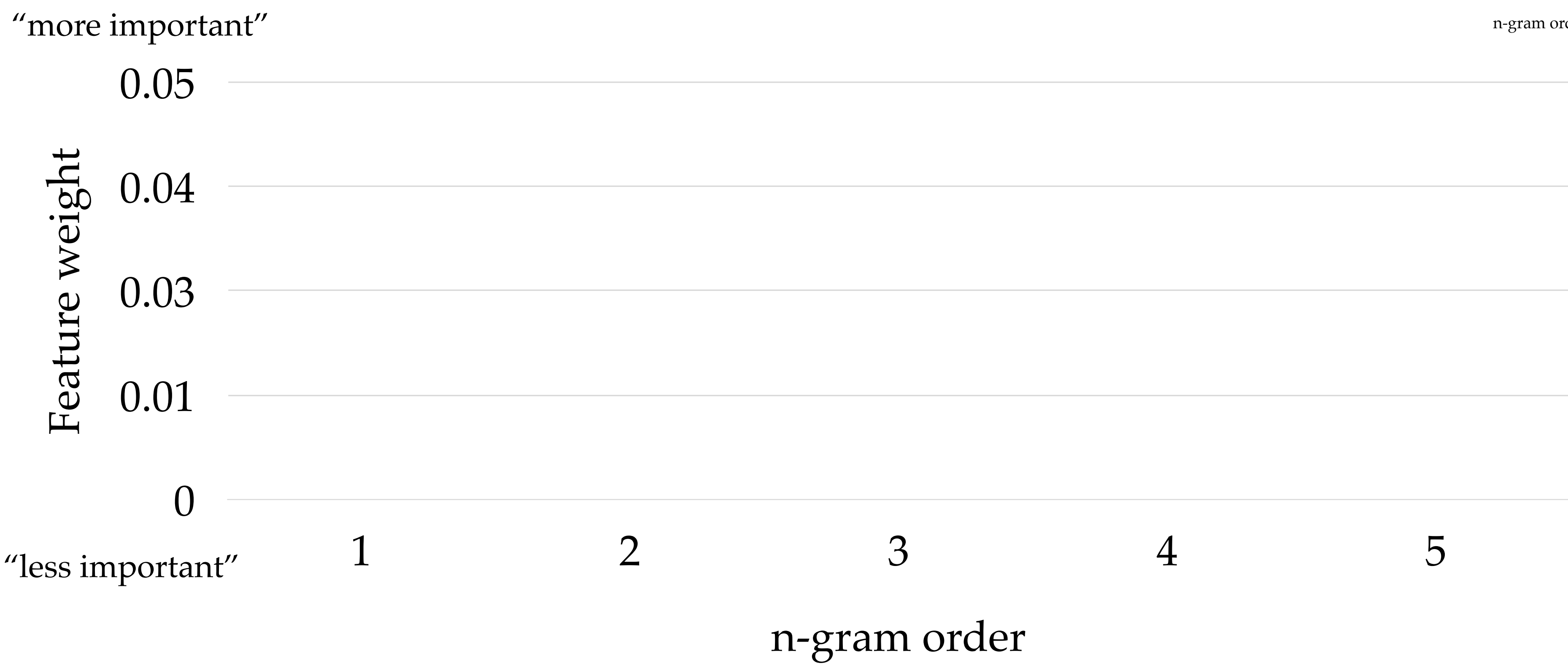




Optimzer: MERT

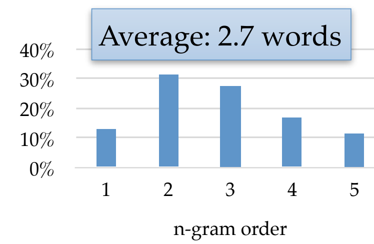
Neural nets vs discrete models

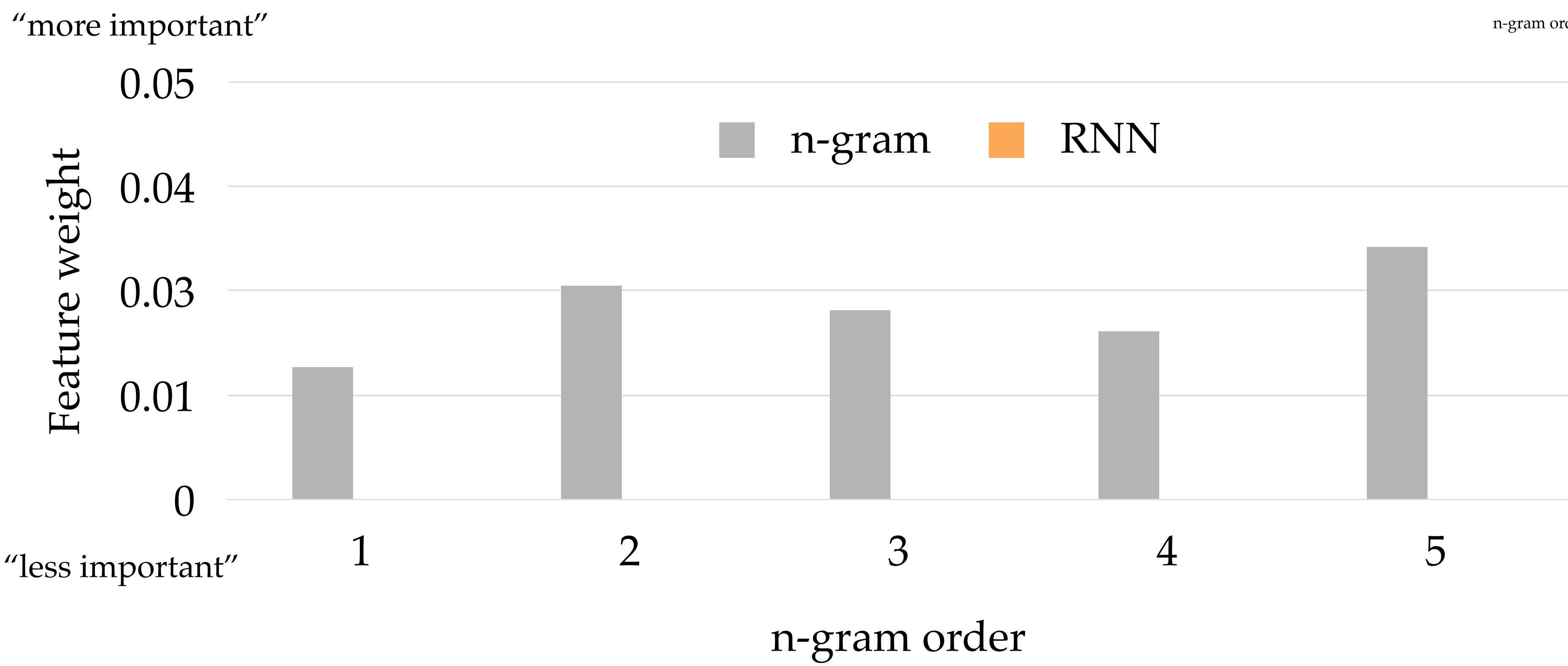




Optimzer: MERT

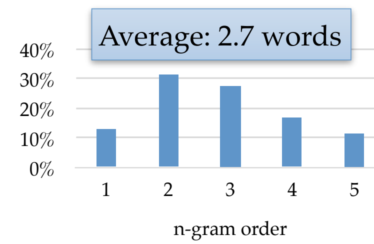
Neural nets vs discrete models

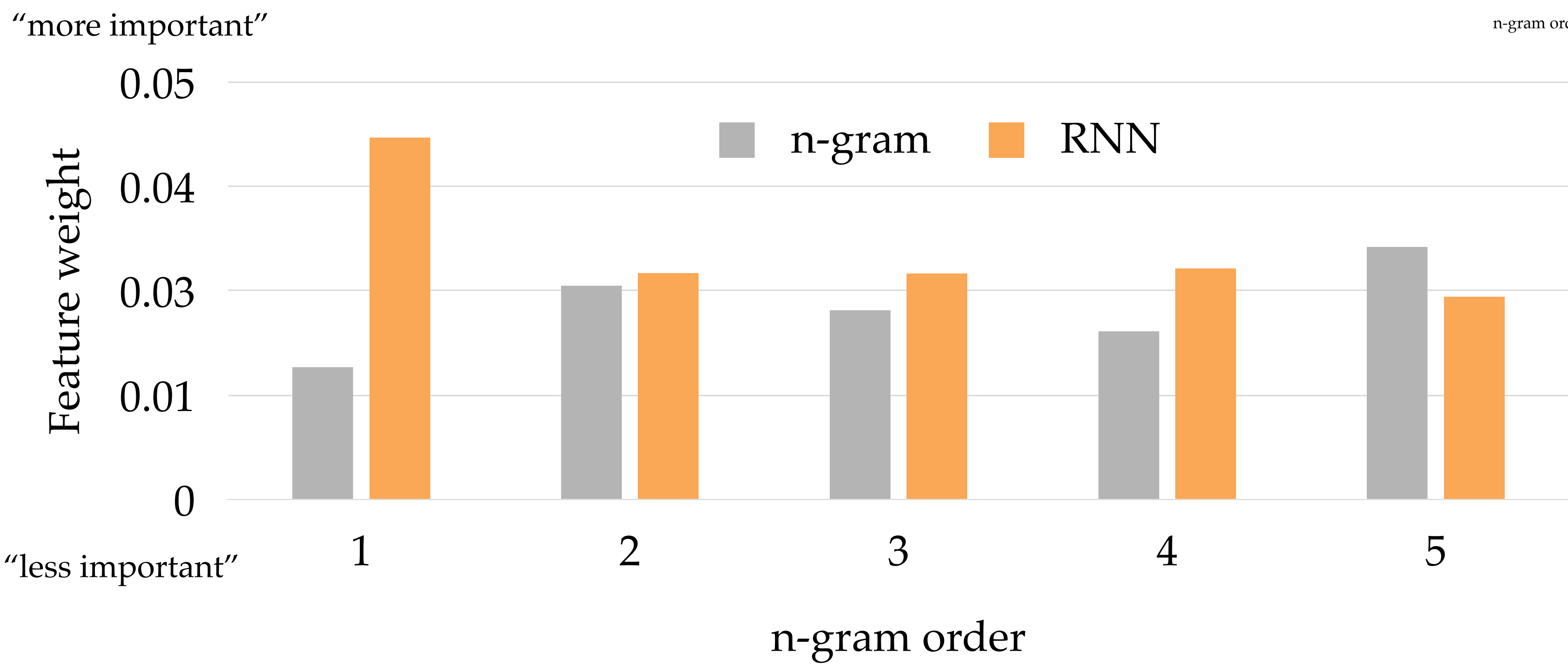




Optimzer: MERT

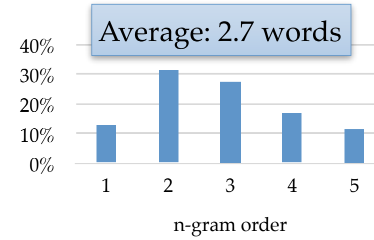
Neural nets vs discrete models

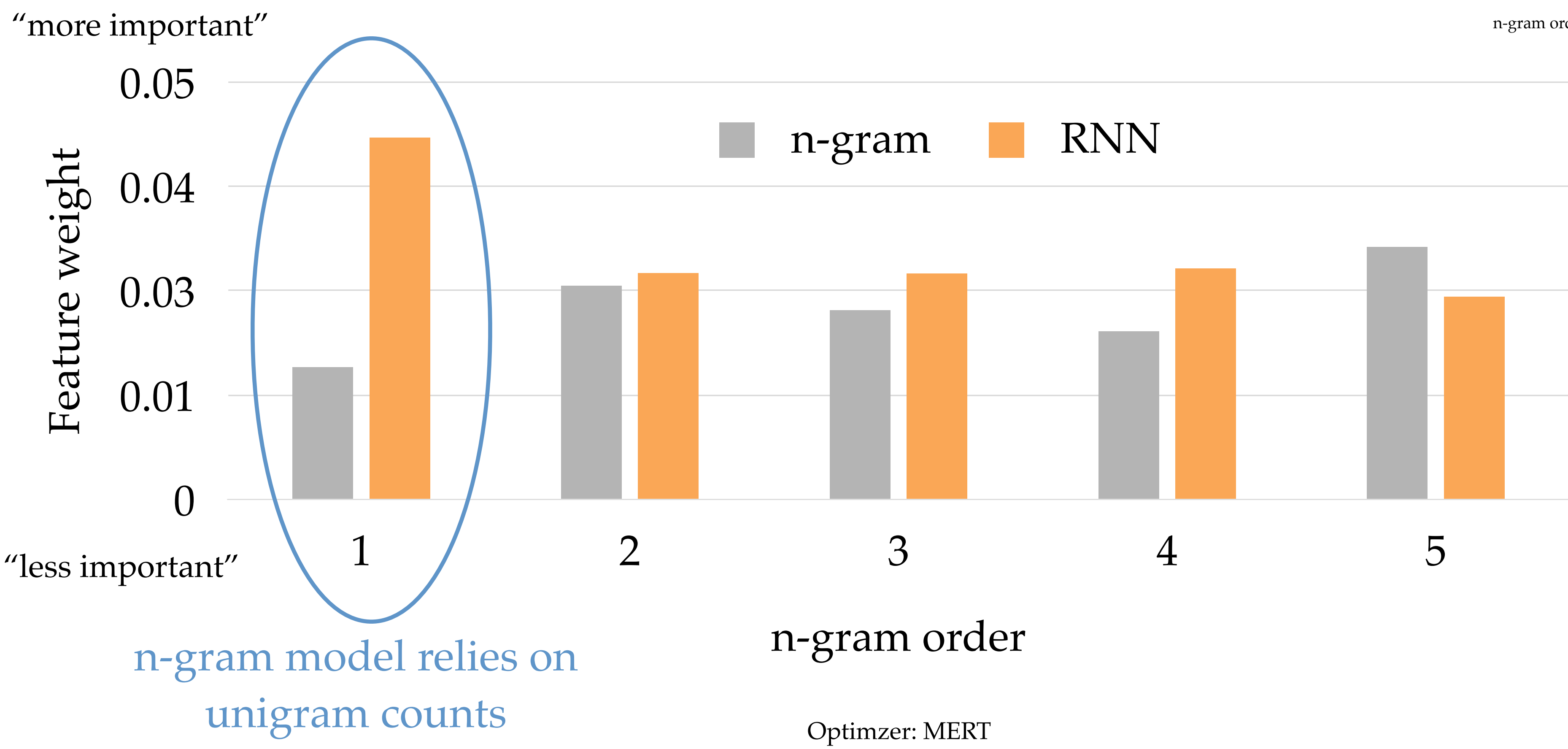




Optimzer: MERT

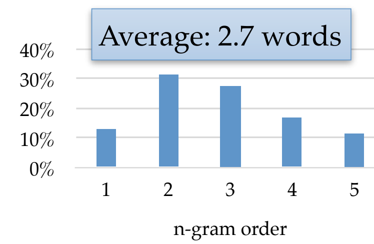
Neural nets vs discrete models

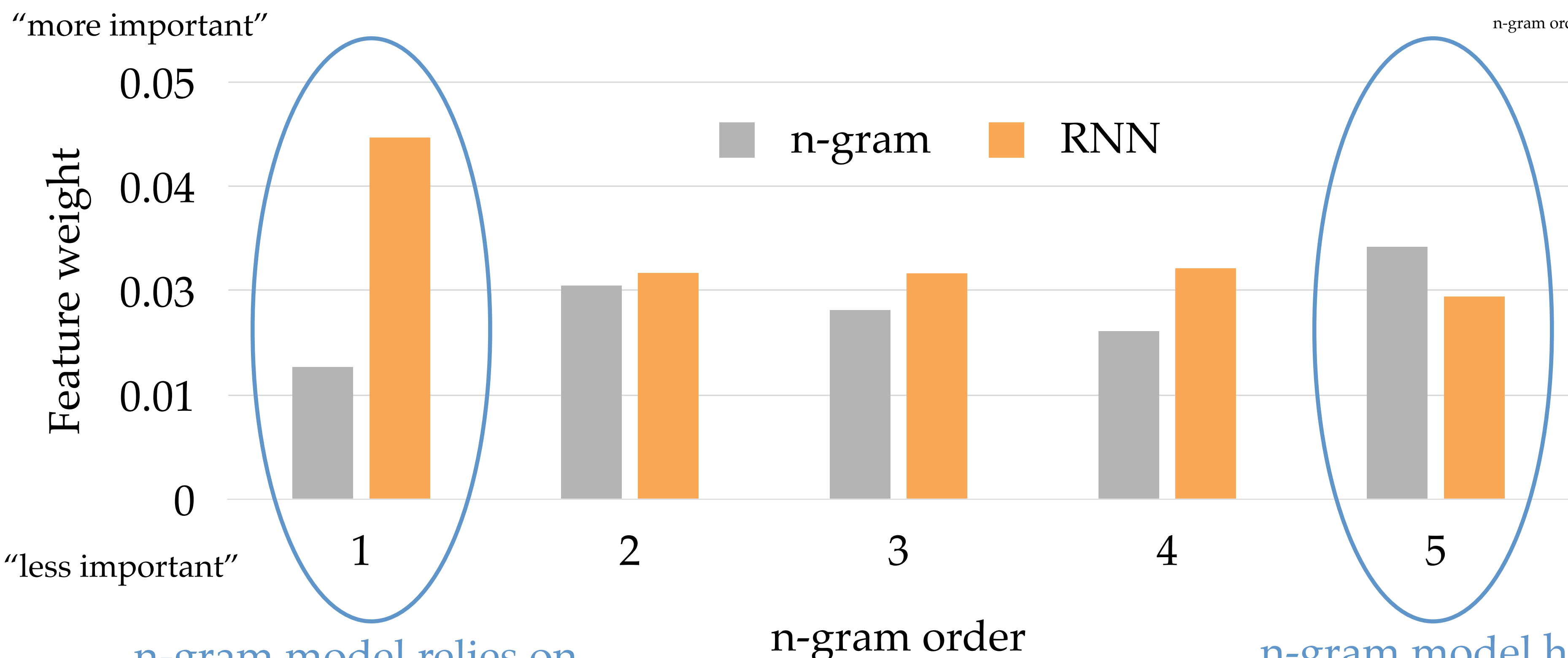




Optimzer: MERT

Neural nets vs discrete models





n-gram model relies on unigram counts

n-gram model has higher-order statistics

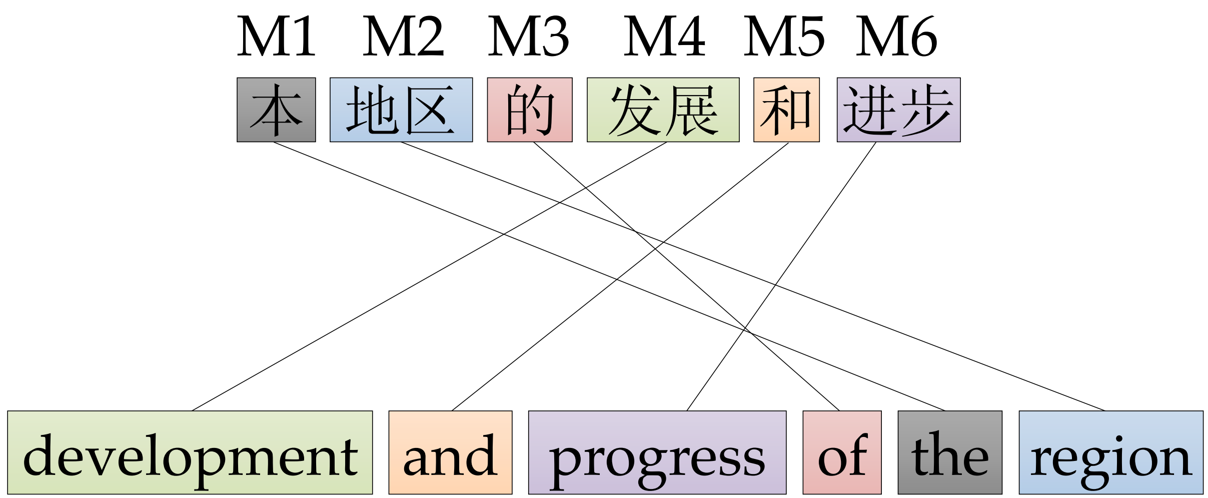
Optimzer: MERT

Recurrent Minimum Translation Unit Models

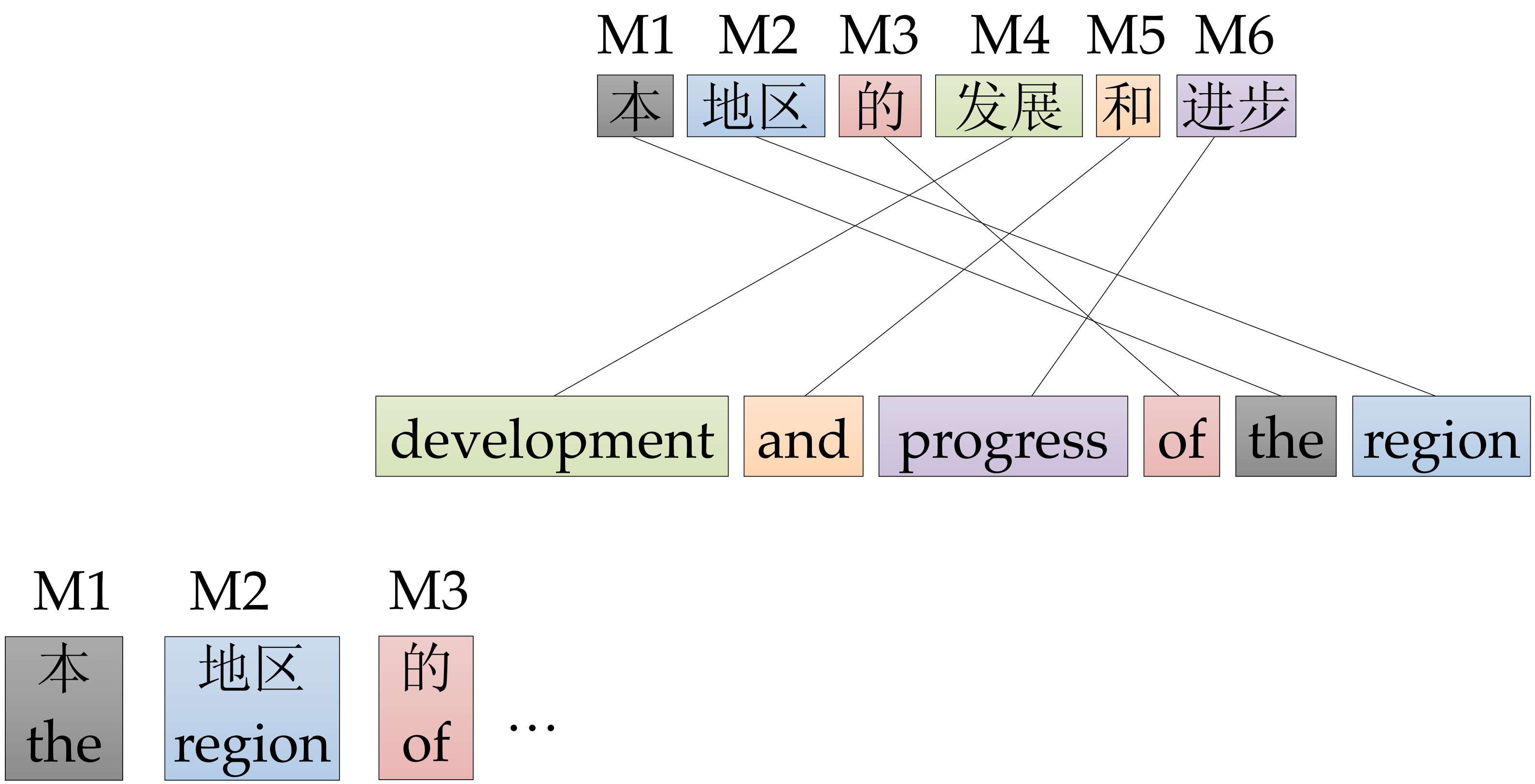
本 地区 的 发展 和 进步

development and progress of the region

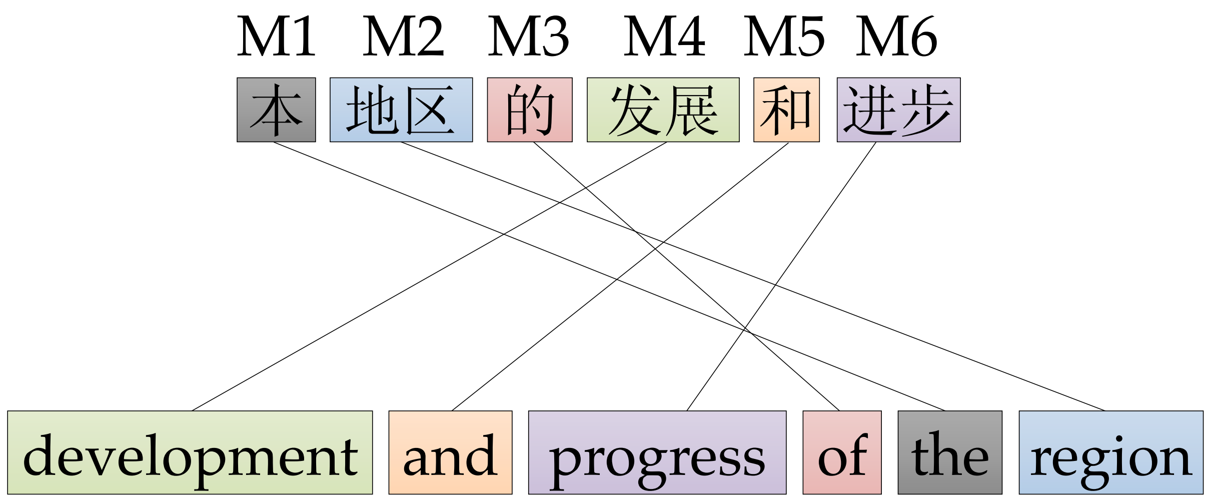
Recurrent Minimum Translation Unit Models

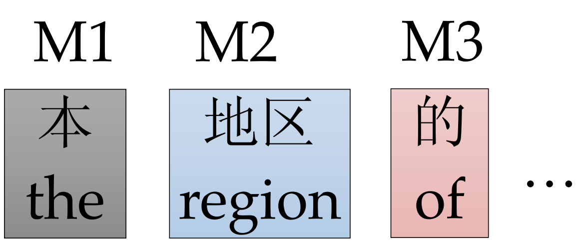


Recurrent Minimum Translation Unit Models



Recurrent Minimum Translation Unit Models





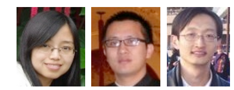
n-gram models over MTUs:

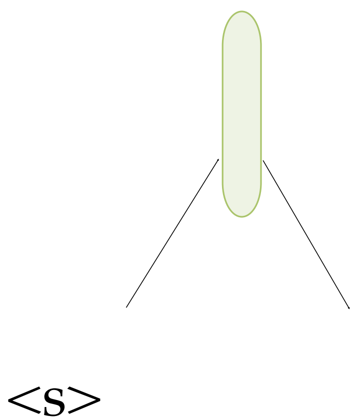
$$p(M1)\, p(M2 \mid M1)\, p(M3 \mid M1, M2) \ldots$$

Banchs et al. (2005), Quirk & Menezes (2006)

Recurrent Minimum Translation Unit Models

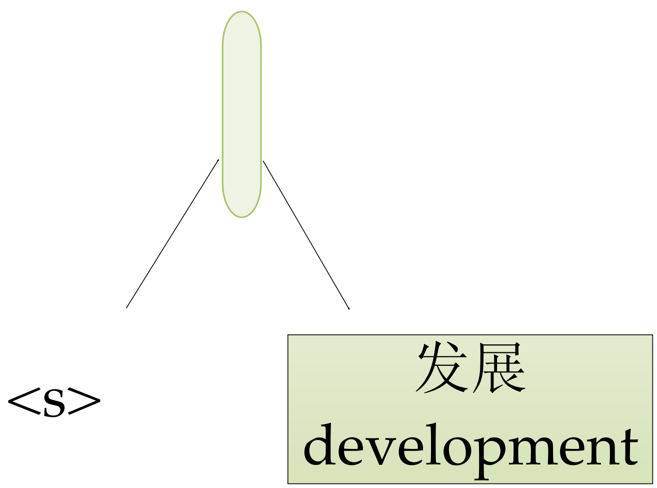
Hu et al., EACL 2014





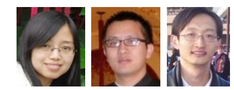
Recurrent Minimum Translation Unit Models

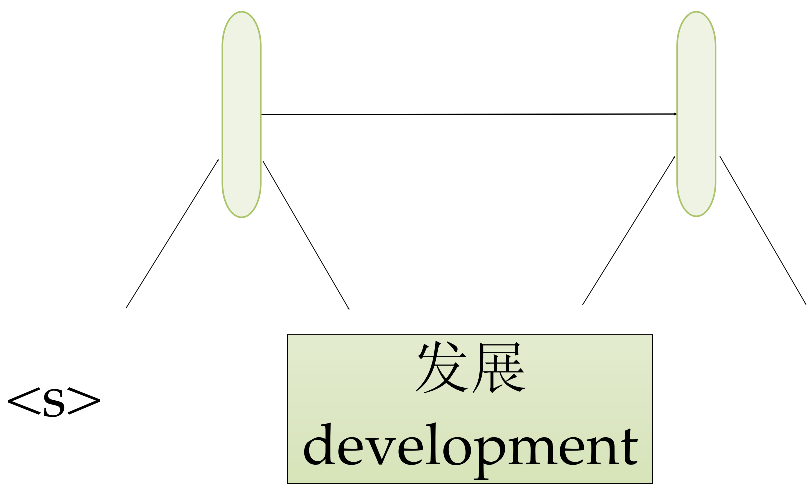
Hu et al., EACL 2014



Recurrent Minimum Translation Unit Models

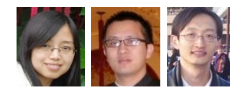
Hu et al., EACL 2014

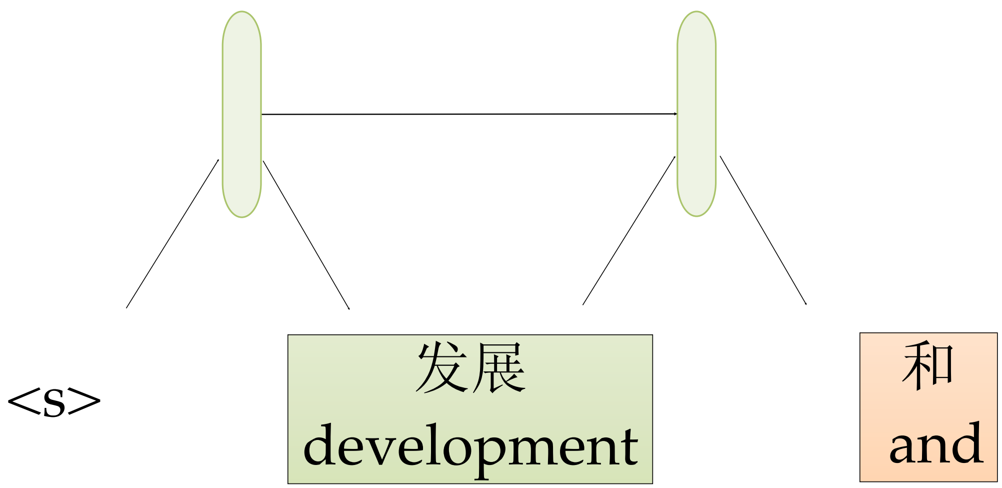




Recurrent Minimum Translation Unit Models

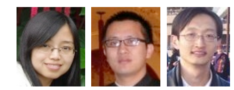
Hu et al., EACL 2014

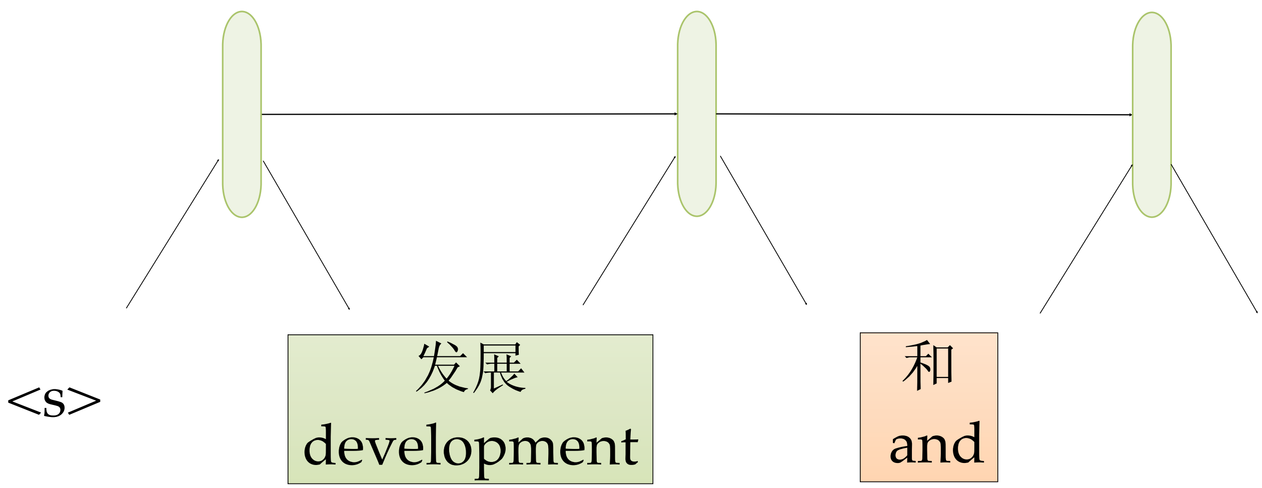




Recurrent Minimum Translation Unit Models

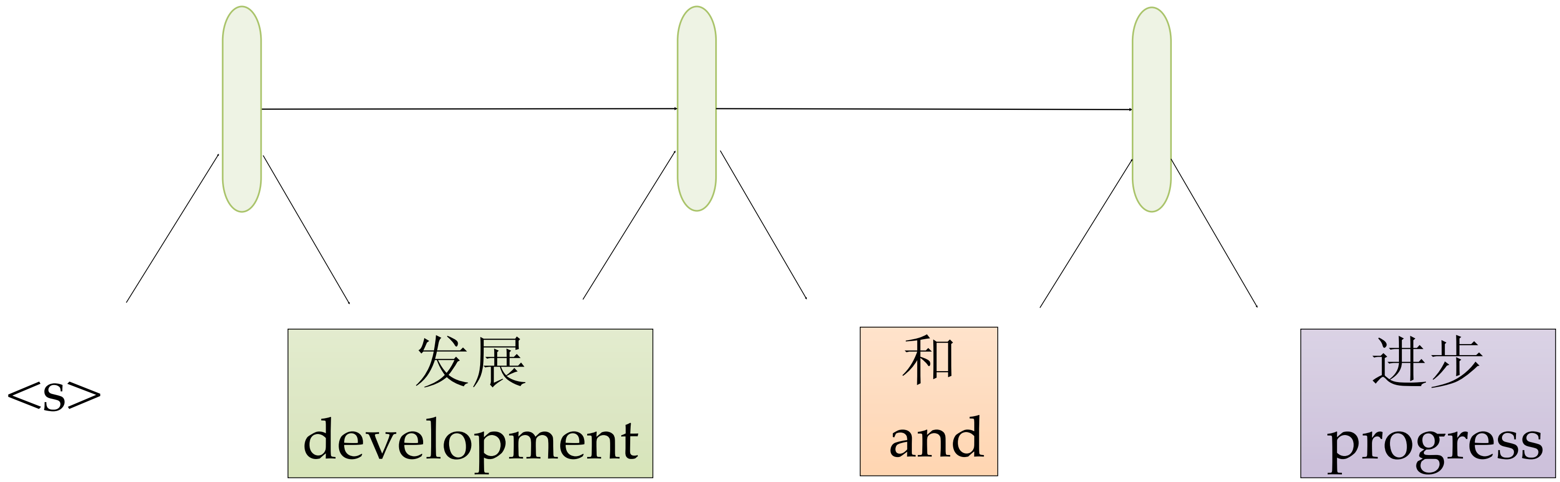
Hu et al., EACL 2014



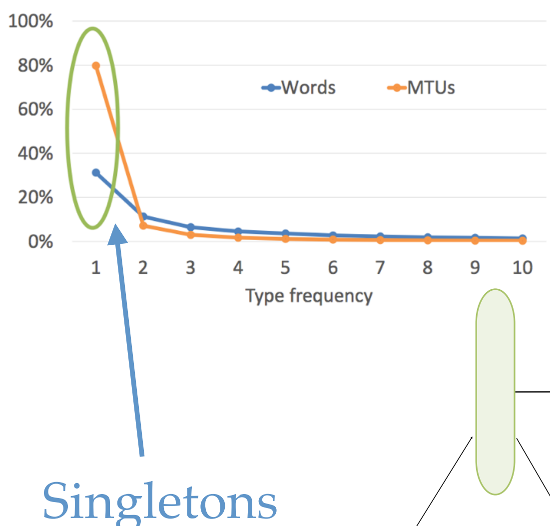


Recurrent Minimum Translation Unit Models

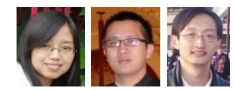
Hu et al., EACL 2014

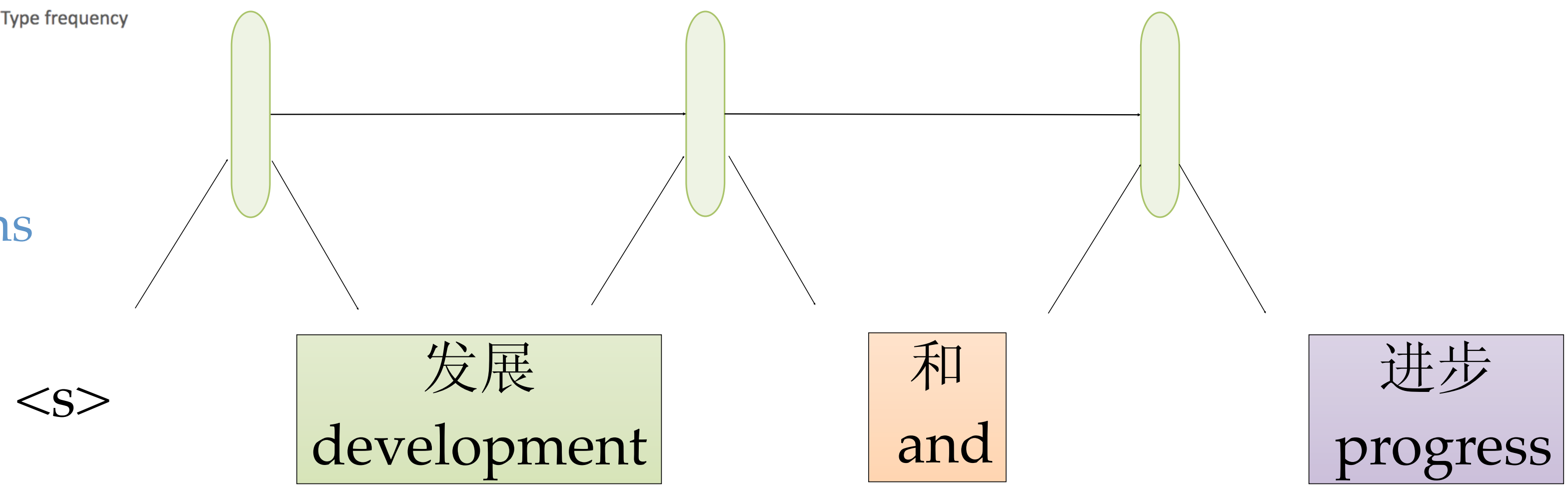


Recurrent Minimum Translation Unit Models



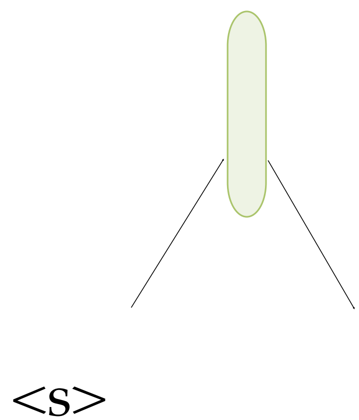
Hu et al., EACL 2014





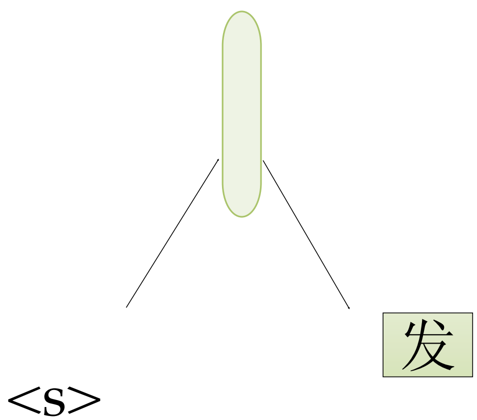
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



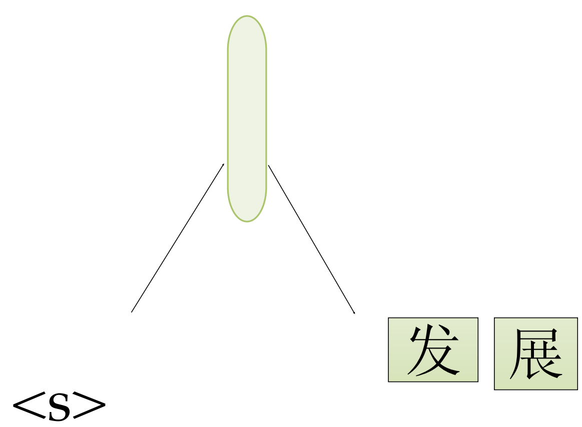
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



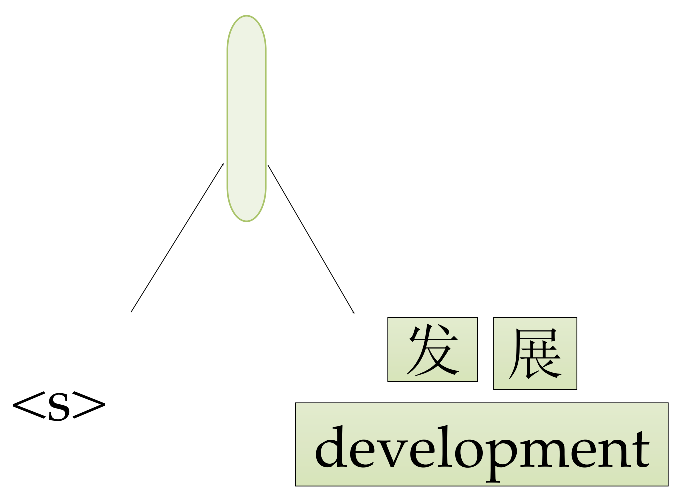
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



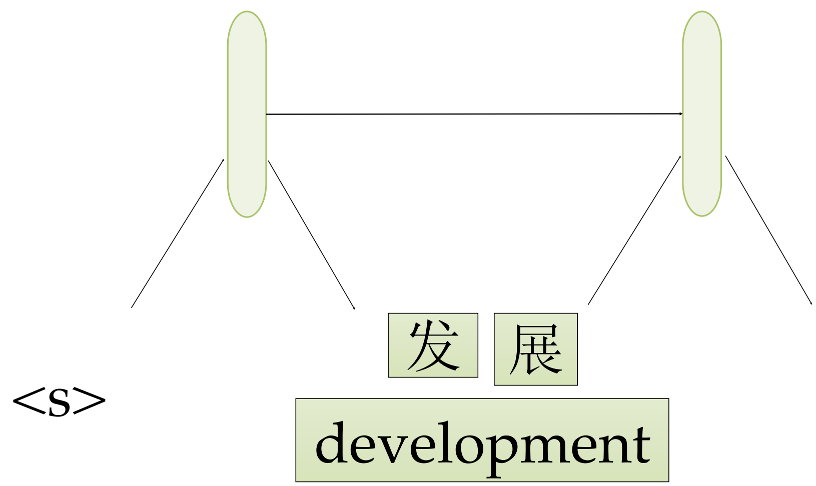
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



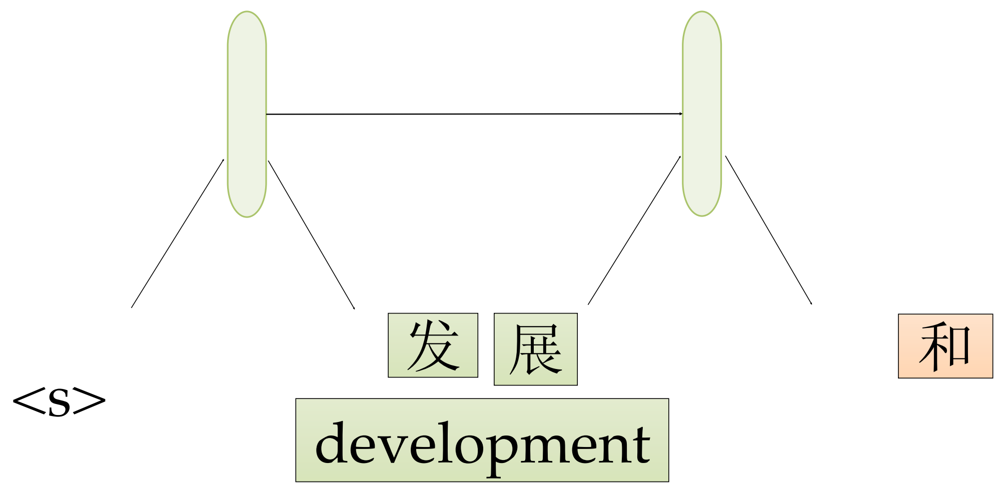
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



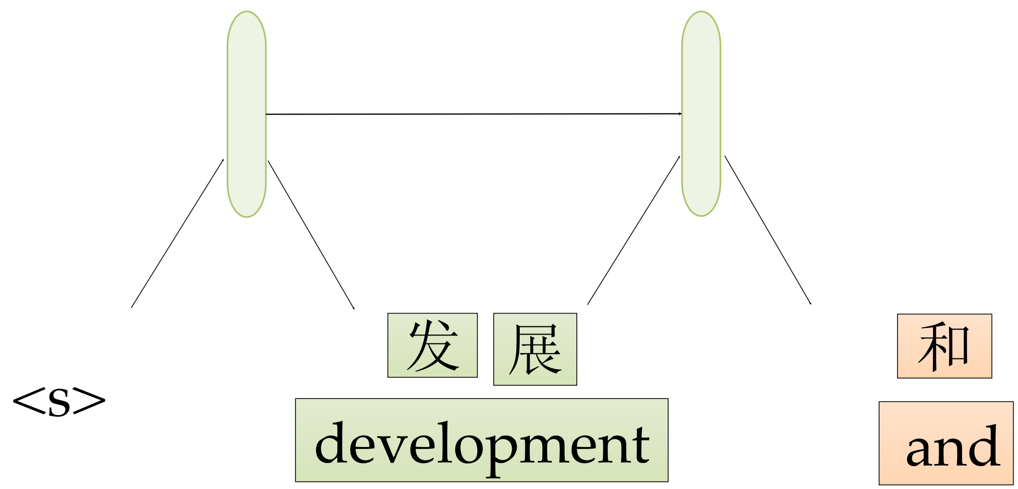
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



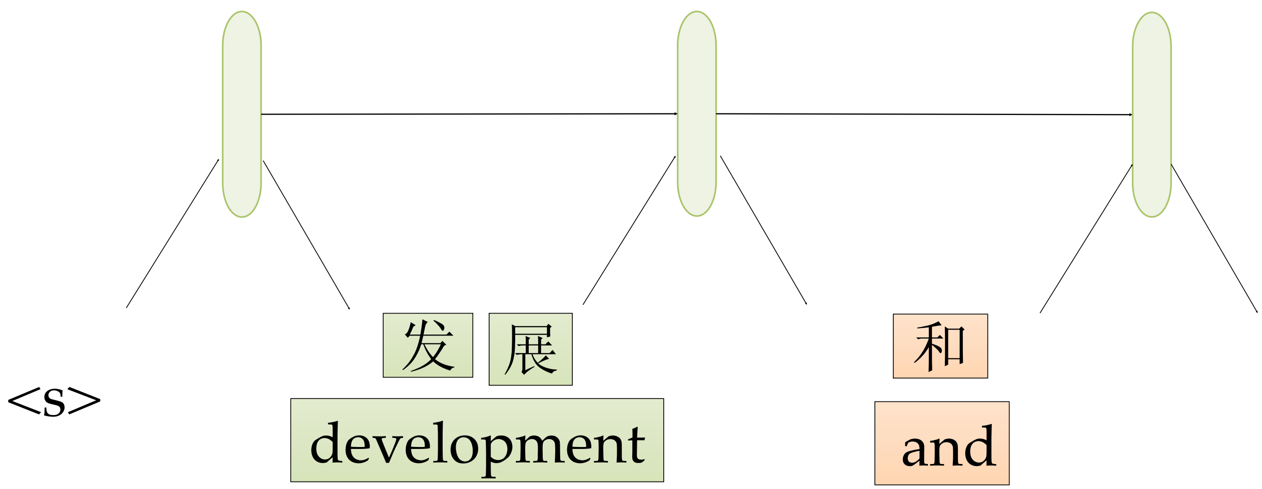
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



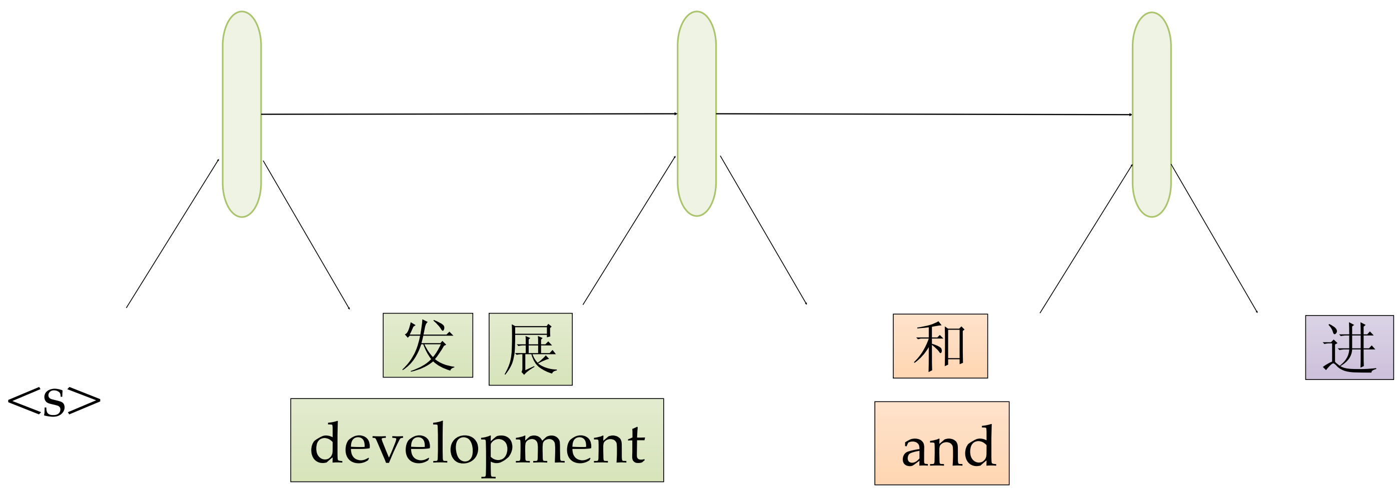
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



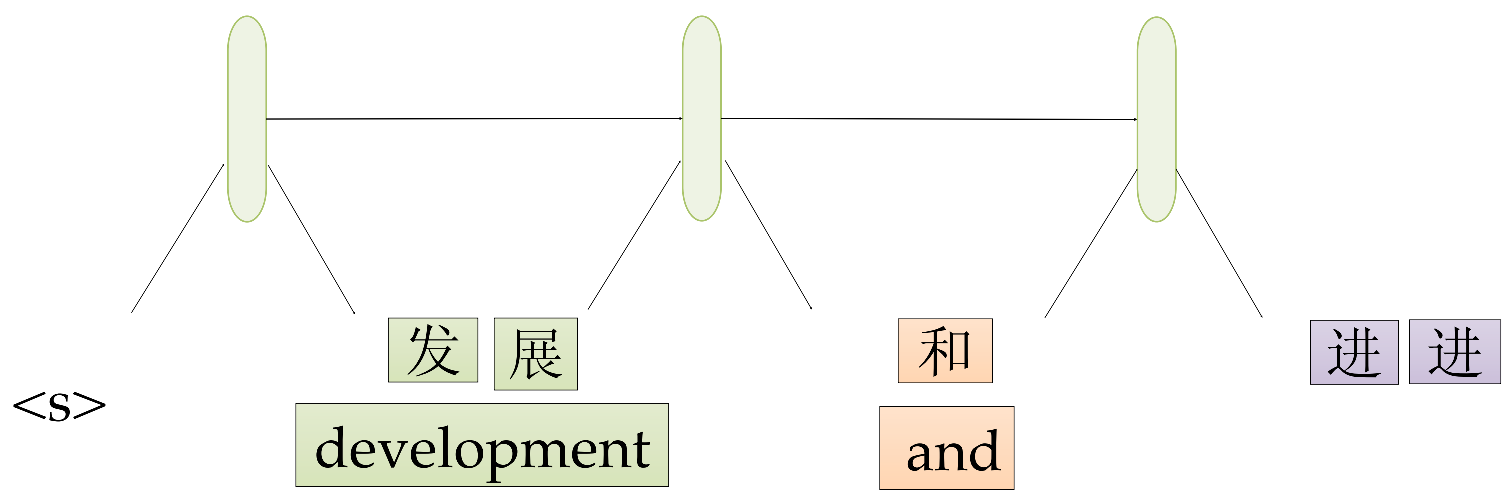
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



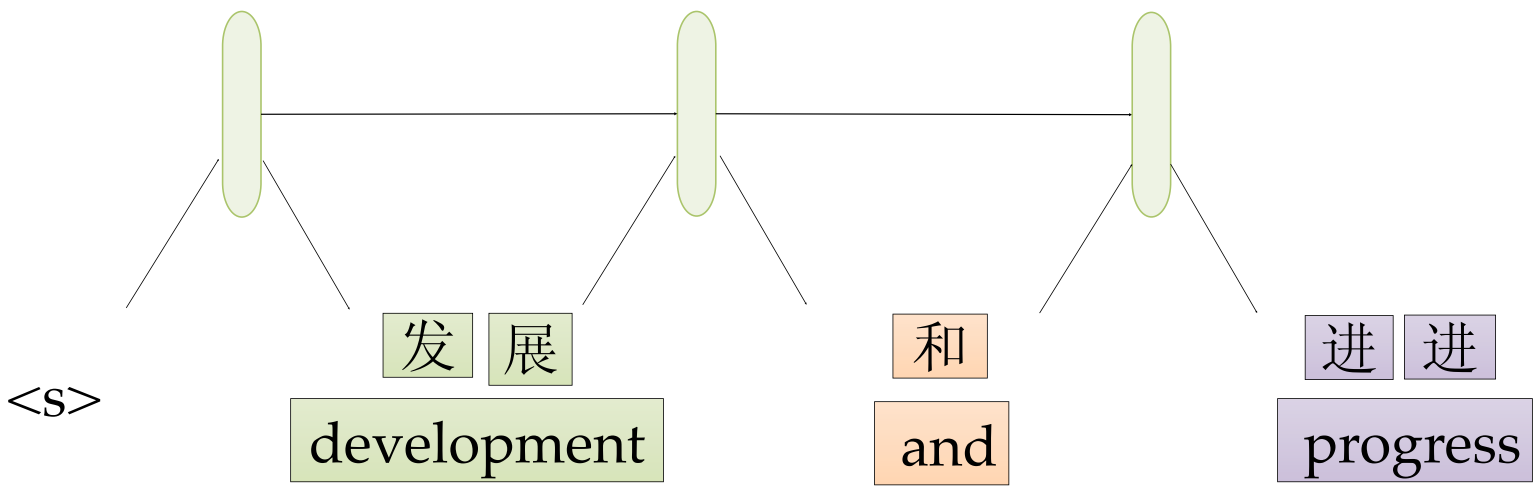
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation



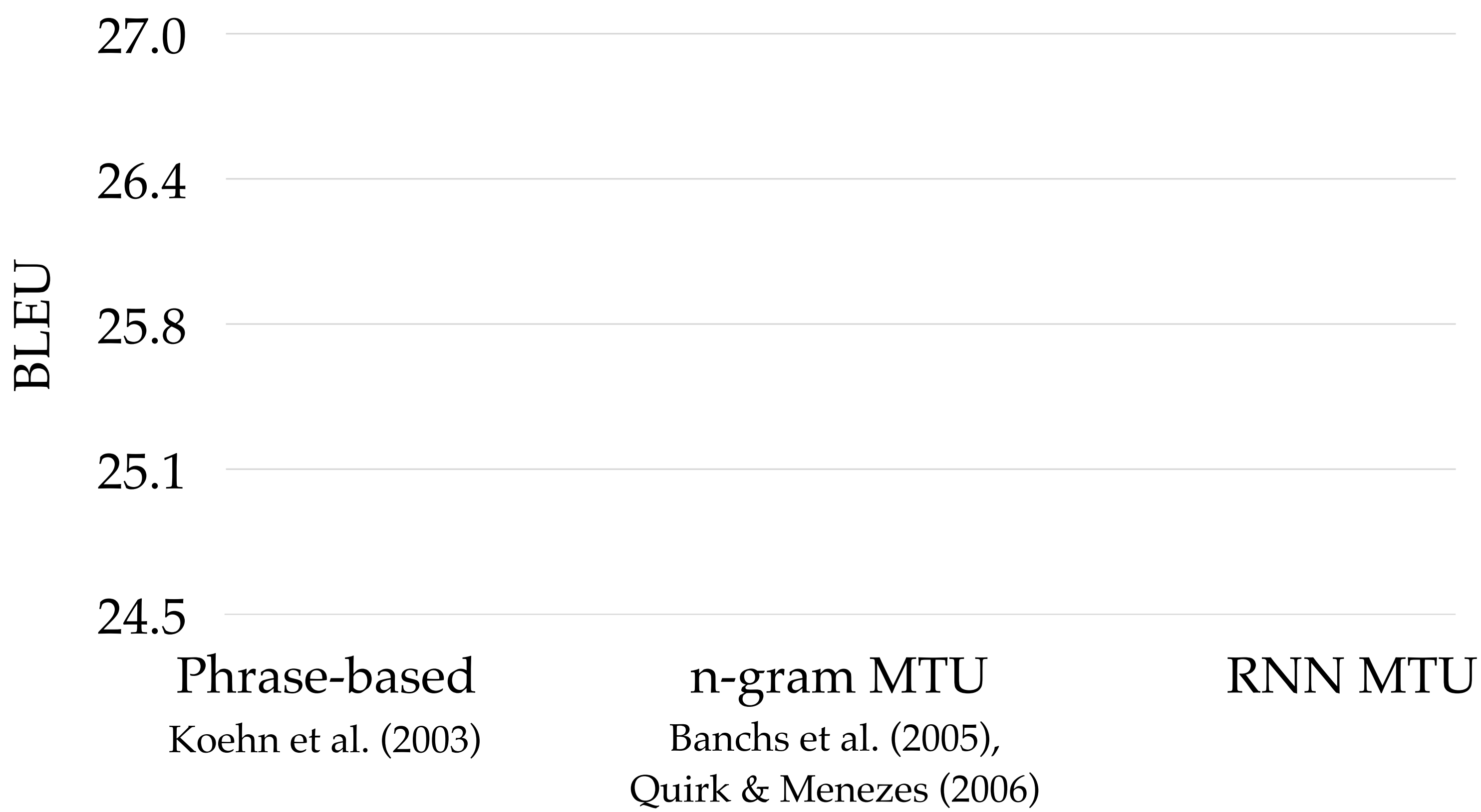
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation

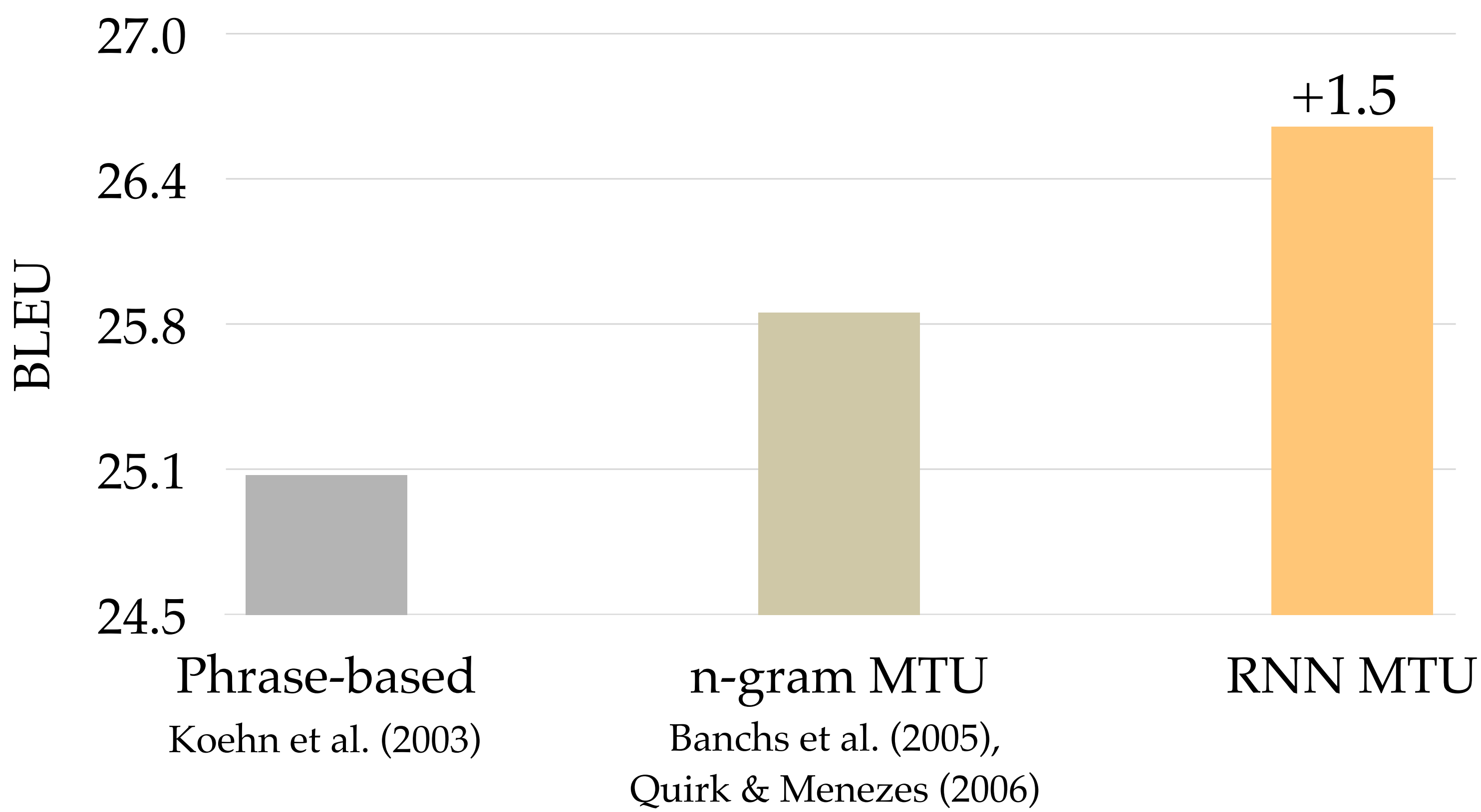


Recurrent Minimum Translation Unit Models

Recurrent Minimum Translation Unit Models



Recurrent Minimum Translation Unit Models



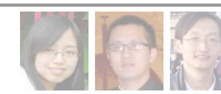
This talk

本 地区 的 发展 和 进步 。

development and progress of the region .

Translation modeling 

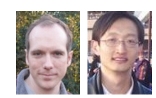
Auli et al., EMNLP 2013; Hu et al., EACL 2014

Language modeling 

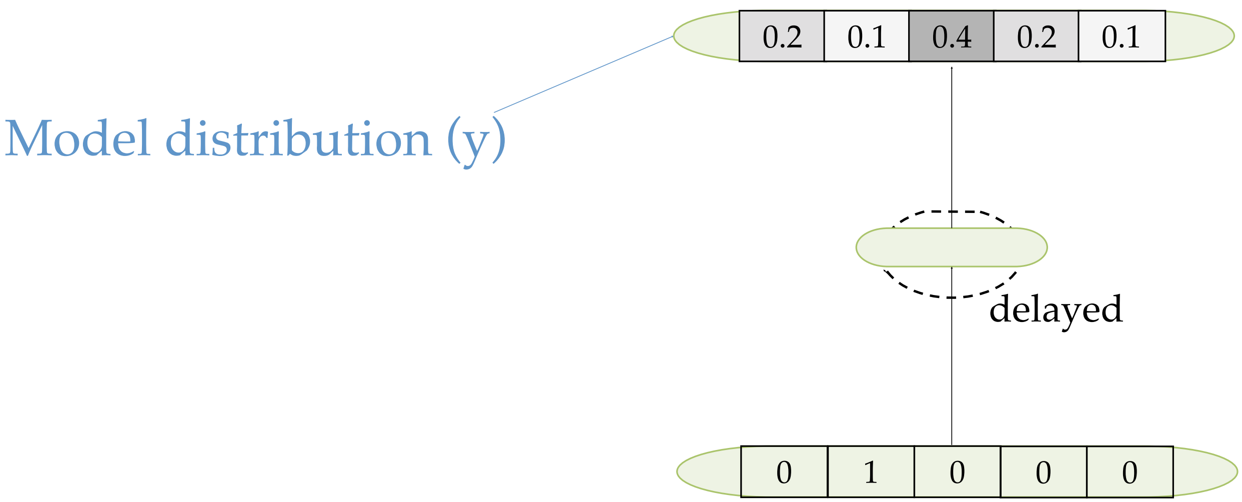
Optimization

Auli & Gao, ACL 2014 

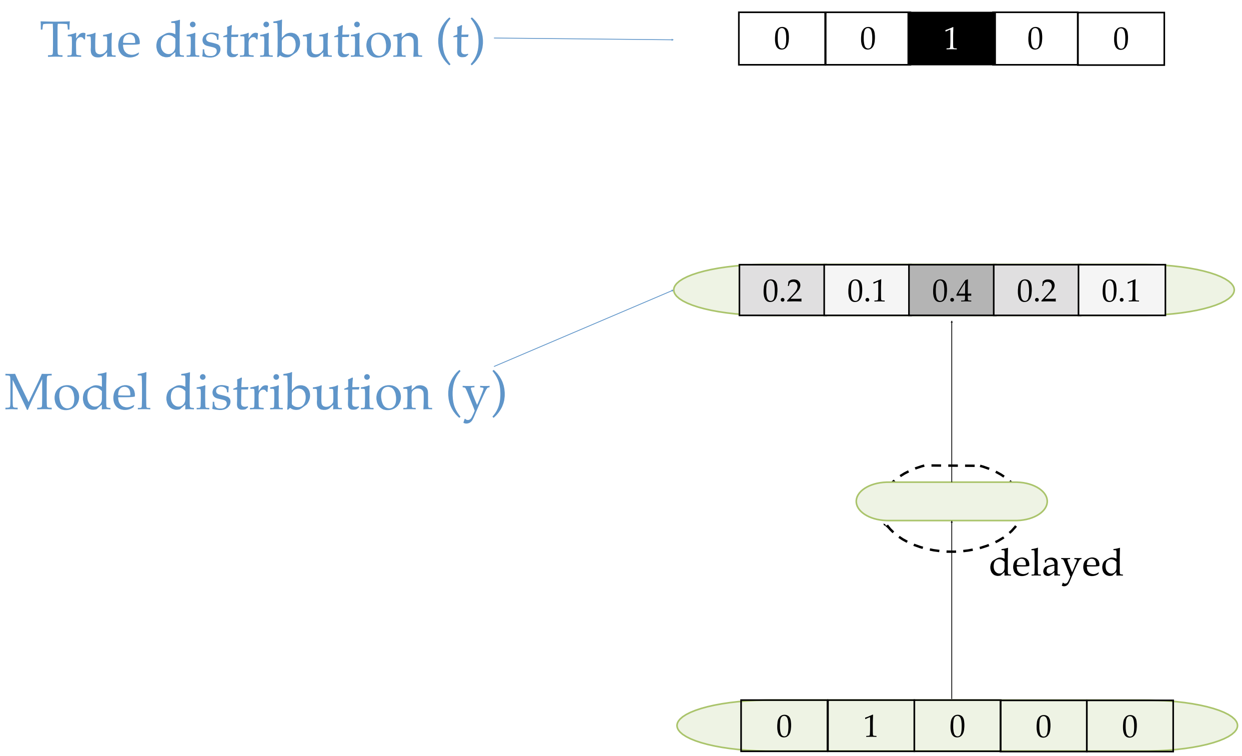
Reordering

Auli et al., EMNLP 2014 

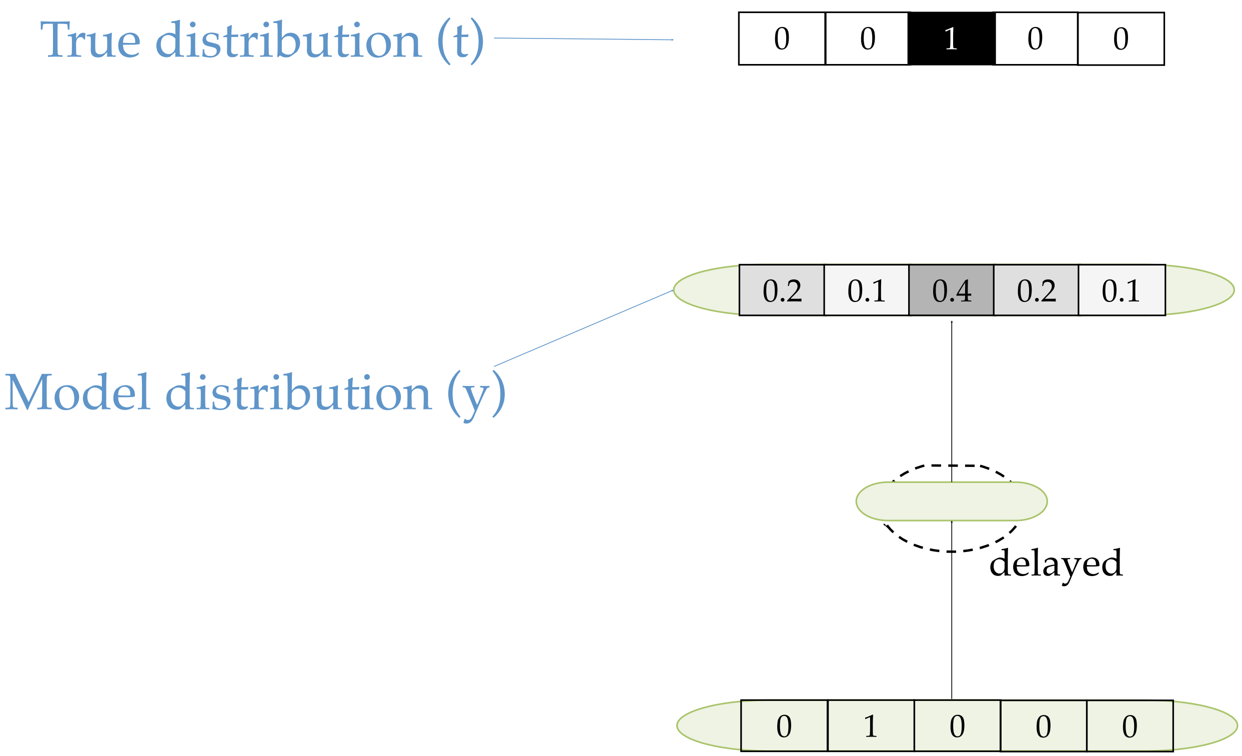
Back propagation with cross entropy error



Back propagation with cross entropy error



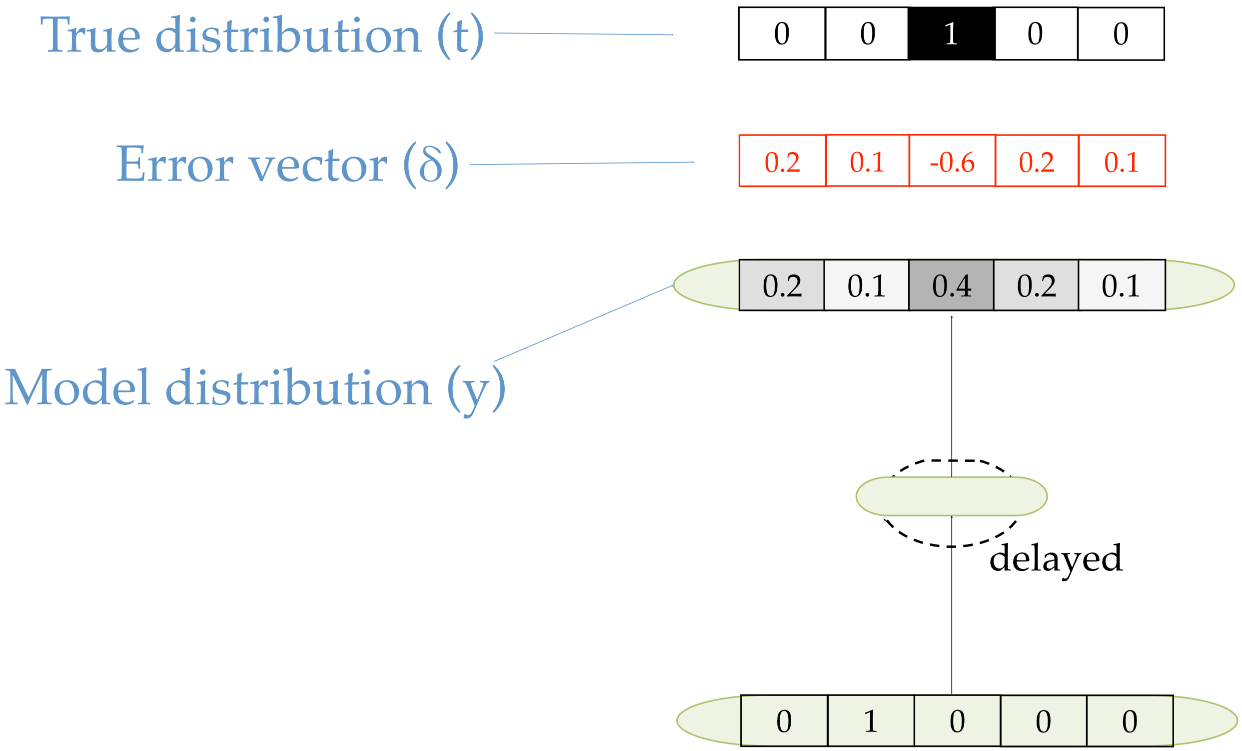
Back propagation with cross entropy error



$$J = -\sum_i t^i \ log \ y^i$$

Goal: Make correct outputs most likely

Back propagation with cross entropy error

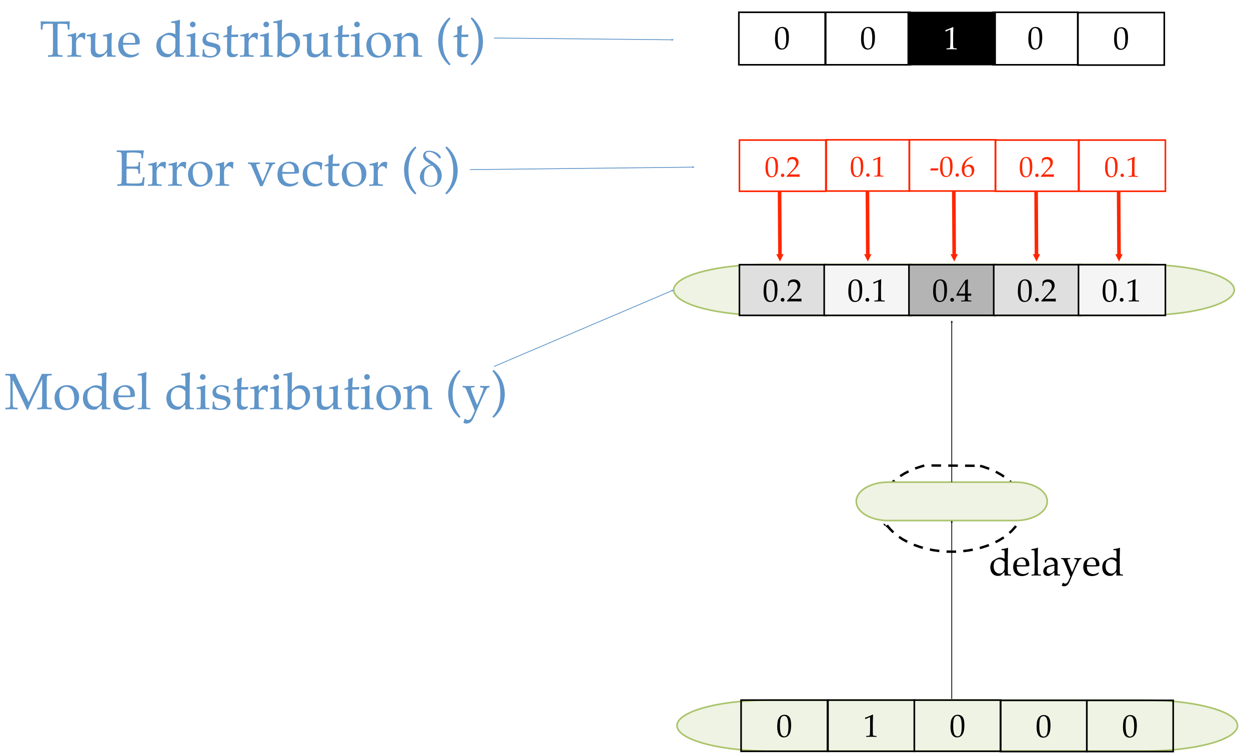


$$\delta^i = y^i - t^i$$

$$J = -\sum_i t^i \, log \, y^i$$

Goal: Make correct outputs most likely

Back propagation with cross entropy error

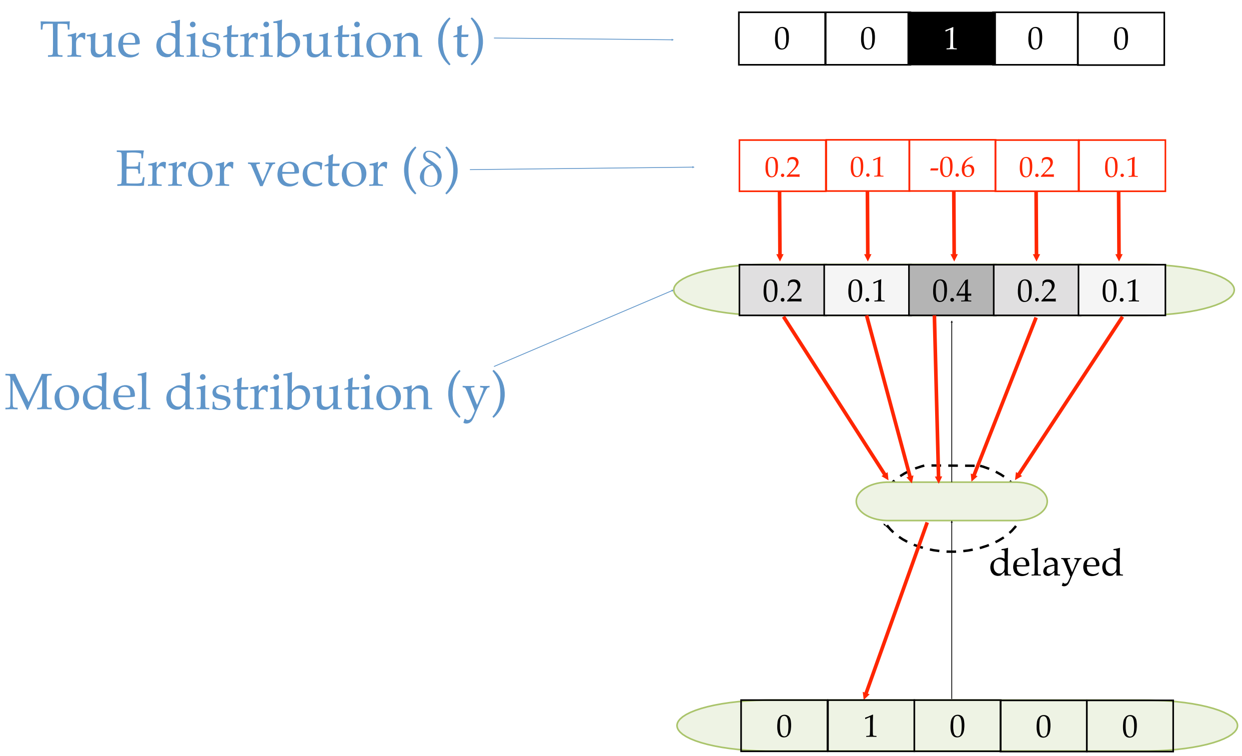


$$\delta^i = y^i - t^i$$

$$J = -\sum_i t^i \; log \; y^i$$

Goal: Make correct outputs most likely

Back propagation with cross entropy error

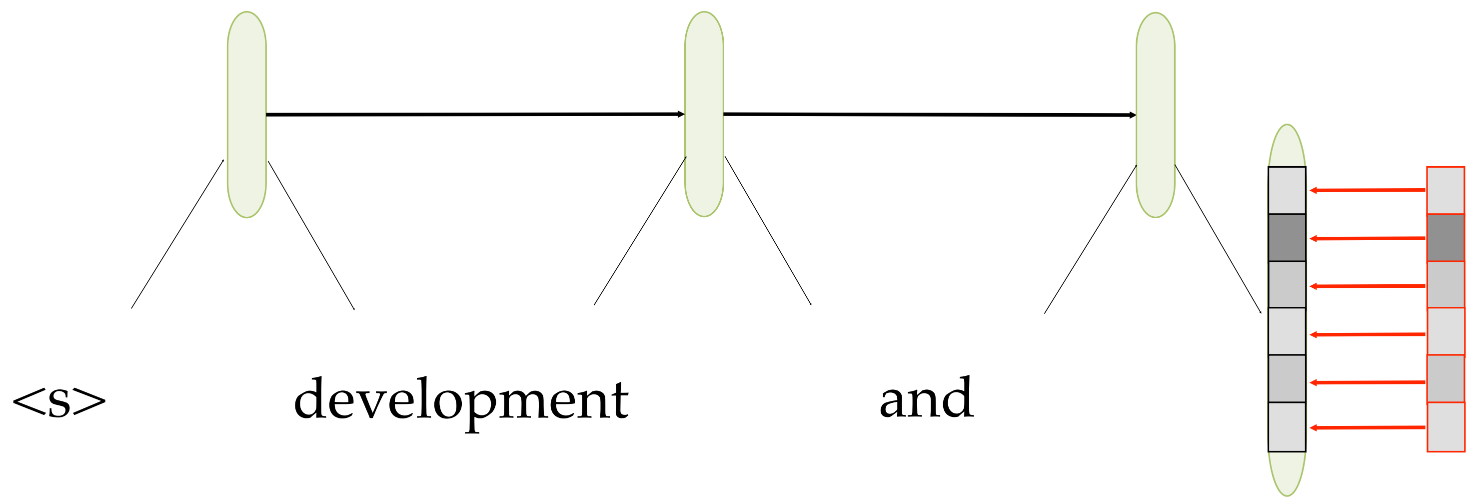


$$\delta^i = y^i - t^i$$

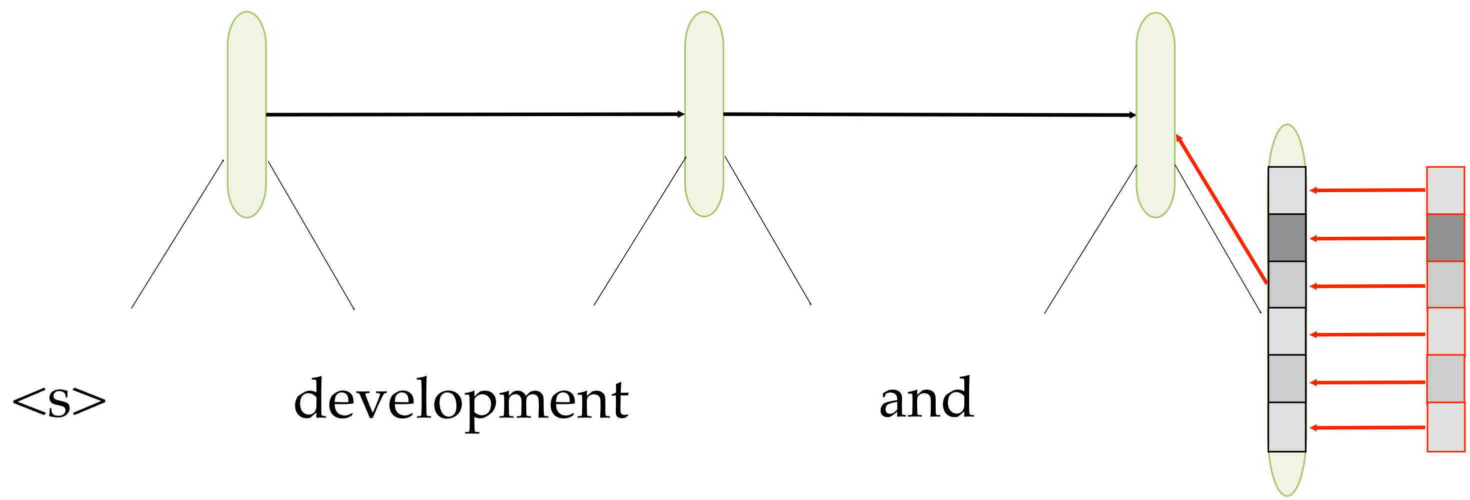
$$J = -\sum_i t^i \, log \; y^i$$

Goal: Make correct outputs most likely

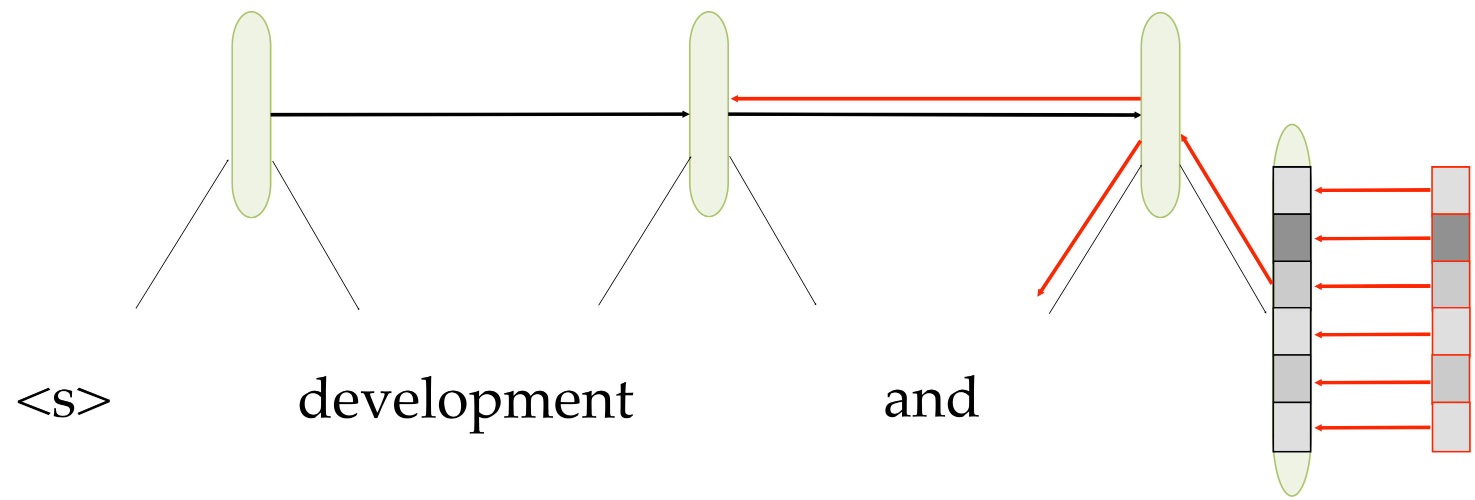
Back propagation through time



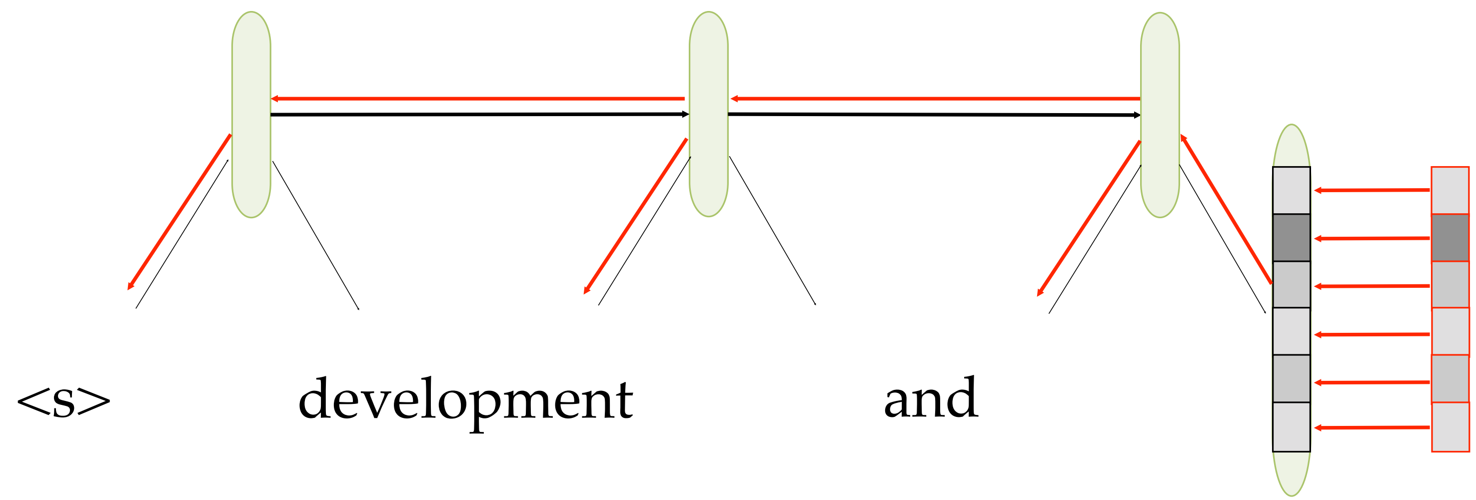
Back propagation through time



Back propagation through time



Back propagation through time



Optimization

Auli & Gao, ACL 2014 

- Likelihood training very common
- Optimizing for evaluation metrics difficult, but empirically successful (Och 2003, Smith 2006, Chiang 2009, Gimpel 2010, Hopkins 2011)

Optimization

Auli & Gao, ACL 2014 

- Likelihood training very common
- Optimizing for evaluation metrics difficult, but empirically successful (Och 2003, Smith 2006, Chiang 2009, Gimpel 2010, Hopkins 2011)
- **Next:** Task-specific training of neural nets for translation

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\text{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \text{BP}$$

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\text{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \text{BP}$$

precision scores

brevity penalty

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\text{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \text{BP}$$

precision scores

brevity penalty

Human: development and progress of the region

System: advance and progress of region

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\text{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \text{BP}$$

precision scores

brevity penalty

Human: development $\boxed{\text{and}}$ $\boxed{\text{progress}}$ $\boxed{\text{of}}$ the $\boxed{\text{region}}$

System: advance $\boxed{\text{and}}$ $\boxed{\text{progress}}$ $\boxed{\text{of}}$ $\boxed{\text{region}}$

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\mathrm{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \mathrm{BP}$$

precision scores

brevity penalty

Human: development $\boxed{\text{and}}\boxed{\text{progress}}\boxed{\text{of}}$ the region

System: advance $\boxed{\text{and}}\boxed{\text{progress}}\boxed{\text{of}}$ region

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

$$\mathrm{BLEU} = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log p_n\right) \mathrm{BP}$$

precision scores

brevity penalty

Human: development $\boxed{\text{and progress of}}$ the region

System: advance $\boxed{\text{and progress of}}$ region

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

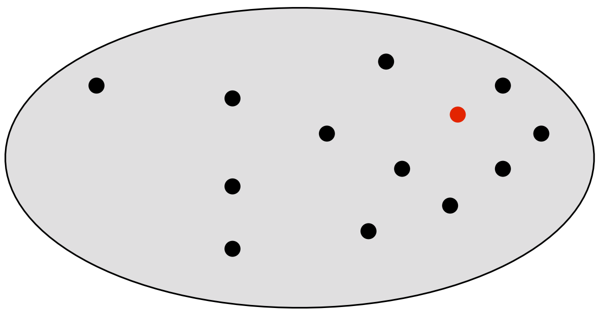
L: $\max_{\theta} p(\tilde{e}|f;\theta)$

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

Desired translation output

L: $$\max_{\theta} p(\tilde{e}|f;\theta)$$

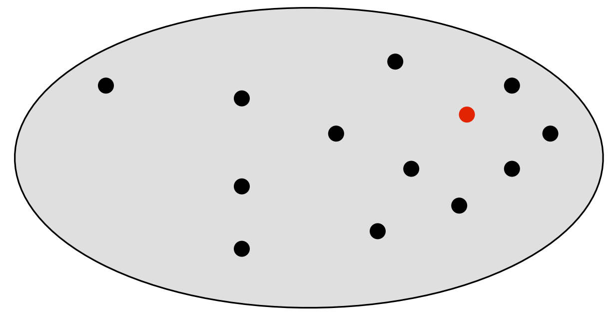


Expected BLEU Training

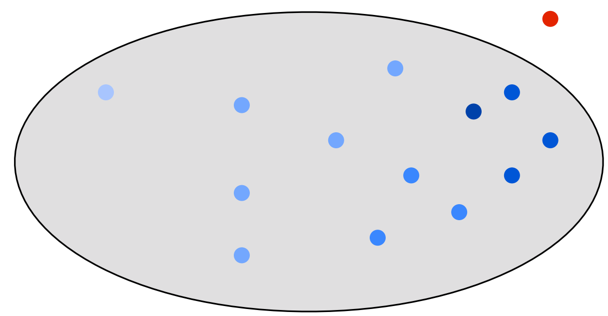
(Smith 2006, He 2012, Gao 2014)

Desired translation output

L: $$\max_{\theta} p(\tilde{e}|f;\theta)$$



xBLEU: $$\max_{\theta} \sum_{e \in E(f)} \mathrm{sBLEU}(e, \tilde{e}) p(e|f;\theta)$$

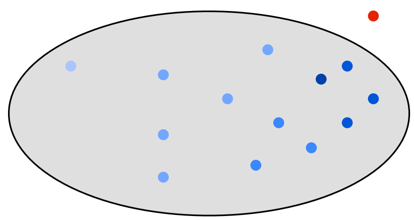


Generated outputs
e.g. n-best, lattice

Gain function

Model probability

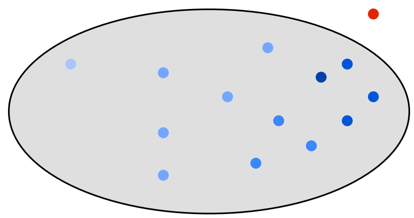
Expected BLEU Training



本 地区 的 发展 和 进步

Human: development and progress of the region

Expected BLEU Training

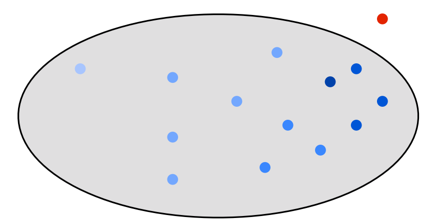


本 地区 的 发展 和 进步

Human: development and progress of the region

advance and progress of the region
development and progress of this province
progress of this region

Expected BLEU Training

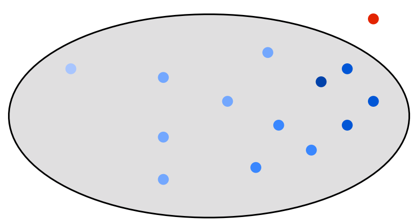


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU
advance and progress of the region	0.8
development and progress of this province	0.5
progress of this region	0.3

Expected BLEU Training

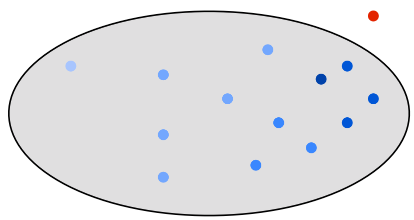


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e \mid f; \theta)$
advance and progress of the region	0.8	0.2
development and progress of this province	0.5	0.3
progress of this region	0.3	0.5

Expected BLEU Training

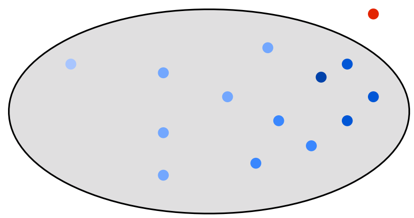


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e \mid f; \theta)$
advance and progress of the region	0.8	0.2
development and progress of this province	0.5	0.3
progress of this region	0.3	0.5

Expected BLEU Training

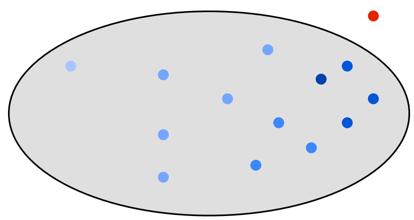


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f;\theta)$
advance and progress of the region	0.8	0.2
development and progress of this province	0.5	0.3
progress of this region	0.3	0.5

Expected BLEU Training



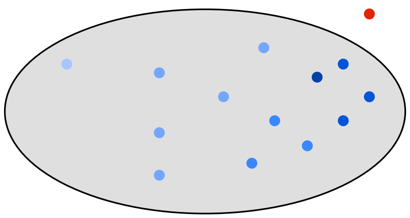
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f;\theta)$
advance and progress of the region	0.8	0.2
development and progress of this province	0.5	0.3
progress of this region	0.3	0.5

$$\text{xBLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e}) p(e|f;\theta) = \mathbf{0.5}$$

Expected BLEU Training



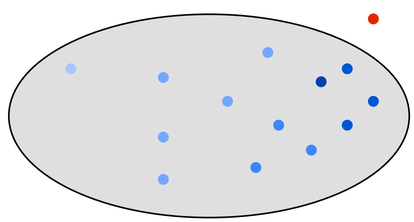
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f;\theta)$	δ_t
advance and progress of the region	0.8	0.2	0.3
development and progress of this province	0.5	0.3	0
progress of this region	0.3	0.5	-0.2

$$\text{xBLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e}) p(e|f;\theta) = 0.5$$

Expected BLEU Training



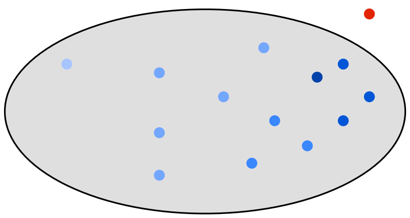
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f;\theta)$	δ_t	$p_{t+1}(e\|f;\theta)$
advance and progress of the region	0.8	0.2	0.3	0.5
development and progress of this province	0.5	0.3	0	0.3
progress of this region	0.3	0.5	-0.2	0.2

$$\text{xBLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e}) p(e|f;\theta) = \mathbf{0.5}$$

Expected BLEU Training



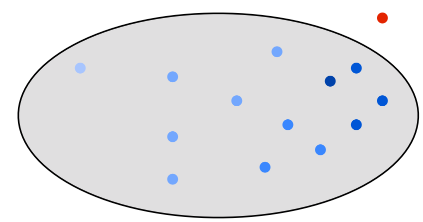
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f;\theta)$	δ_t	$p_{t+1}(e\|f;\theta)$
advance and progress of the region	0.8	0.2	0.3	0.5
development and progress of this province	0.5	0.3	0	0.3
progress of this region	0.3	0.5	-0.2	0.2

$$\mathrm{xBLEU} = \sum_{e \in E(f)} \mathrm{sBLEU}(e, \tilde{e}) p(e|f;\theta) = \mathbf{0.5}$$

Expected BLEU Training



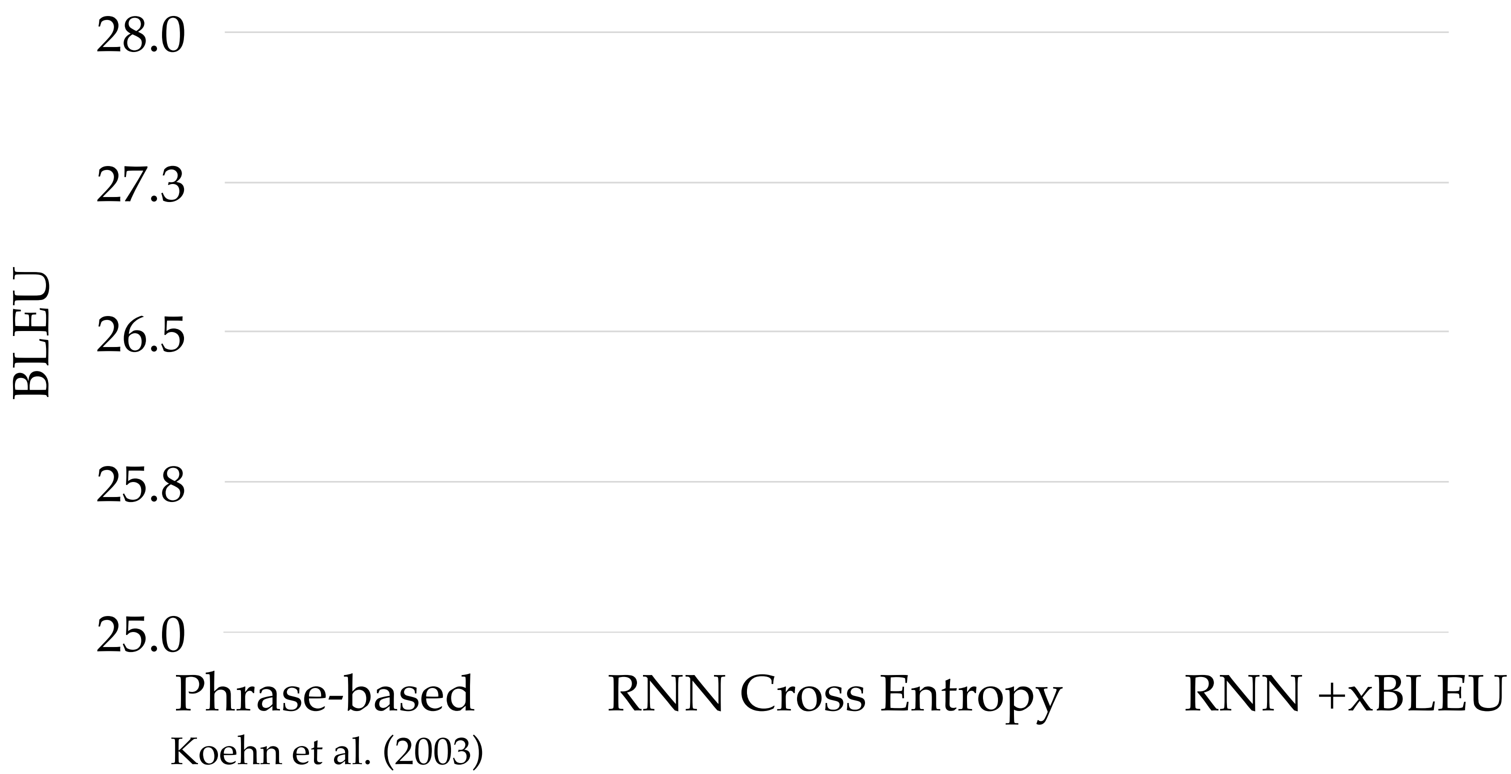
本 地区 的 发展 和 进步

Human: development and progress of the region

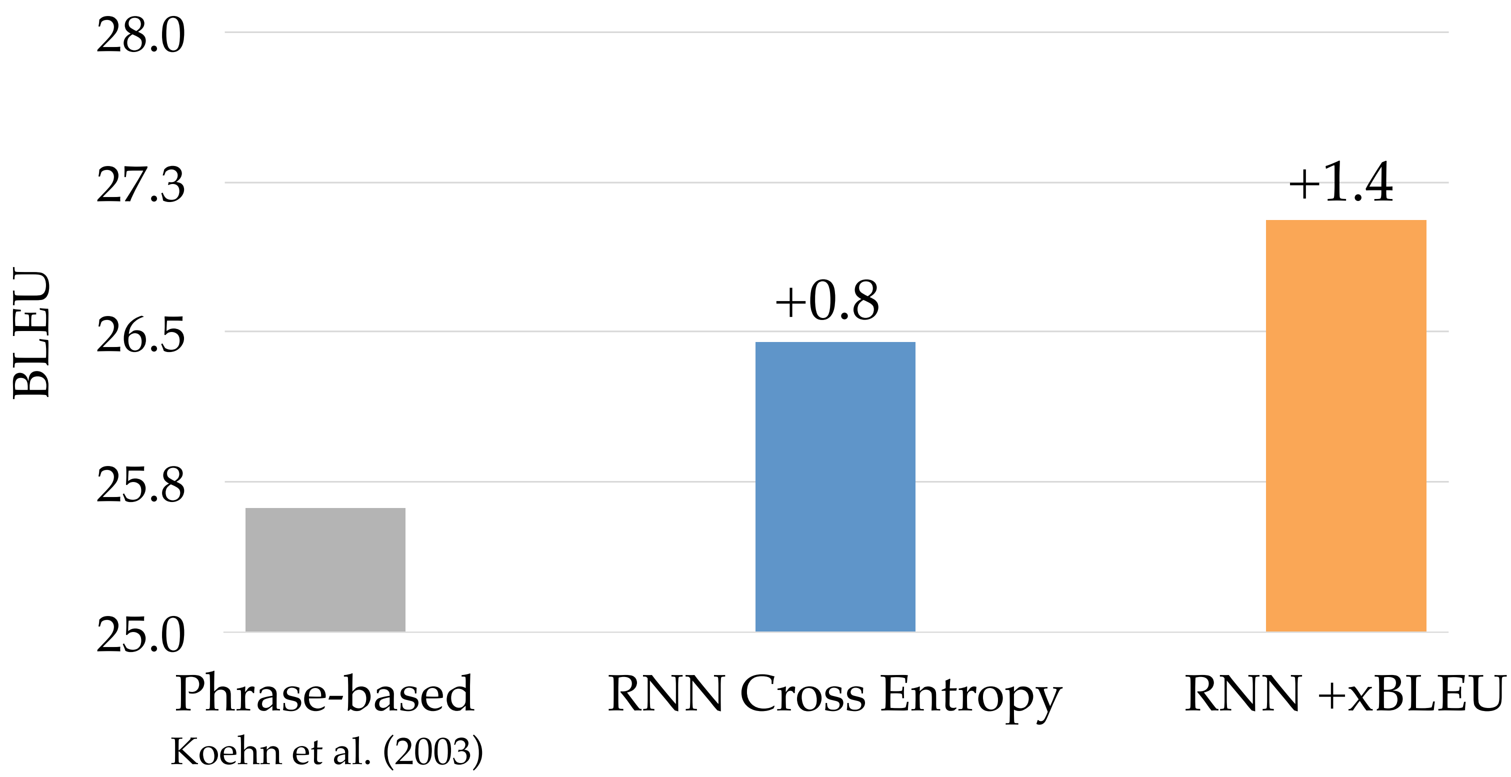
	sBLEU	$p_t(e \mid f;\theta)$	δ_t	$p_{t+1}(e \mid f;\theta)$
advance and progress of the region	0.8	0.2	0.3	0.5
development and progress of this province	0.5	0.3	0	0.3
progress of this region	0.3	0.5	-0.2	0.2

$$\text{xBLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e}) p(e \mid f; \theta) = 0.5 \rightarrow \mathbf{0.6}$$

Results


BLEU
28.0
27.3
26.5
25.8
25.0
Phrase-based
Koehn et al. (2003)
RNN Cross Entropy
RNN +xBLEU

Results



28.0
27.3
26.5
25.8
25.0
BLEU
+0.8
+1.4
Phrase-based
Koehn et al. (2003)
RNN Cross Entropy
RNN +xBLEU

Scaling linear reordering models

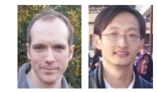
xBLEU training of millions of linear features

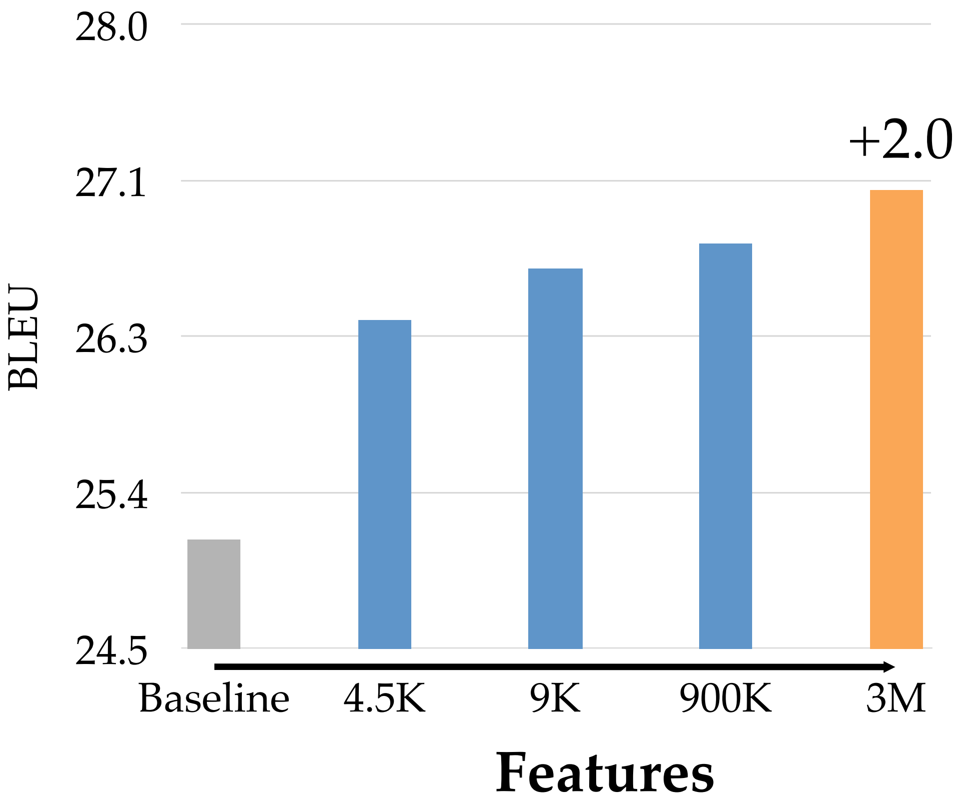
Auli et al., EMNLP 2014  

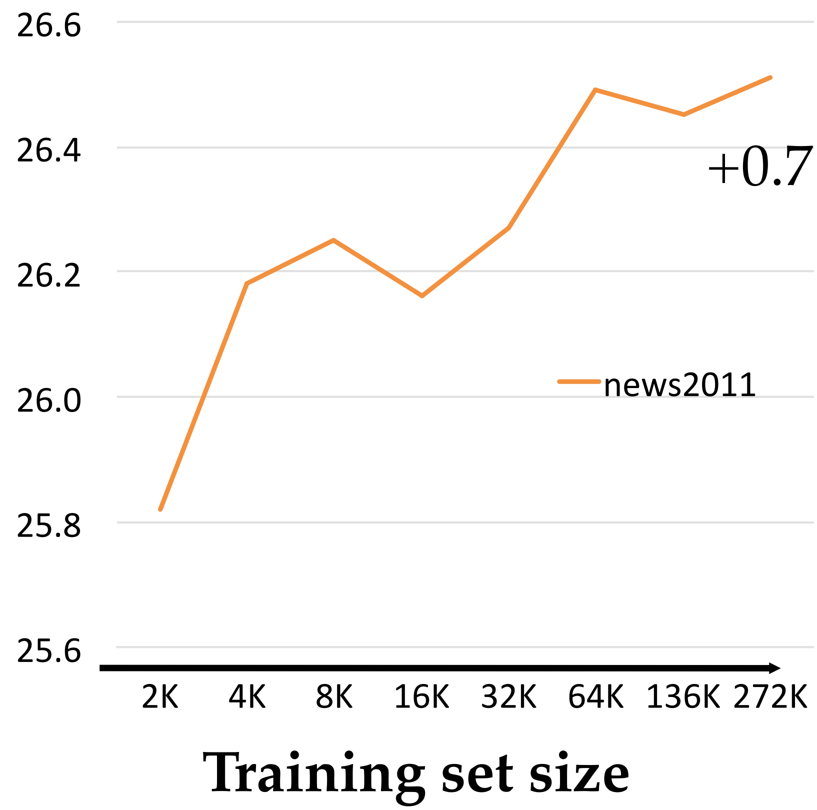
Scaling linear reordering models

xBLEU training of millions of linear features

Auli et al., EMNLP 2014

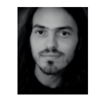






My other neural network projects

- Social media response generation with RNNs
  building neural net-based conversational agents based on twitter conversations




- Semi-supervised phrase table expansion with word embeddings
  using distributional word and phrase representations and
  by mapping between distributional source and target spaces with RBVs

- CCG parsing & tagging with RNNs

Semantic CCG Parsing

Zettlemoyer (2005, 2007), Bos (2008), Kwiatkowski (2010, 2013) Krishnamurthy (2012), Lewis (2013a,b)

Marcel — proved — completeness

$NP :$ *marcel*

$(S\backslash NP)/NP :$ $\lambda y.\lambda x.proved(x,y)$

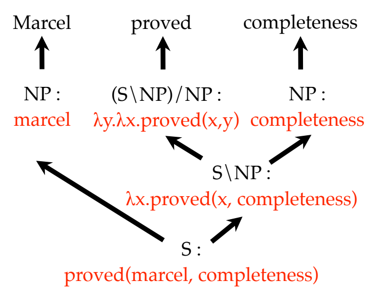
$NP :$ *completeness*

$S\backslash NP :$ $\lambda x.proved(x, completeness)$

$S :$ *proved(marcel, completeness)*

Combinatory Categorial Grammar (CCG; Steedman 2000)

How is this useful?



Answer

User query

(Semantic) parsing

Knowledge Base

How is this useful?

Parse failures and lexical ambiguity are a major source of errors in semantic parsing (Kwiatkowski 2013)

nswer

User query

(Semantic) parsing

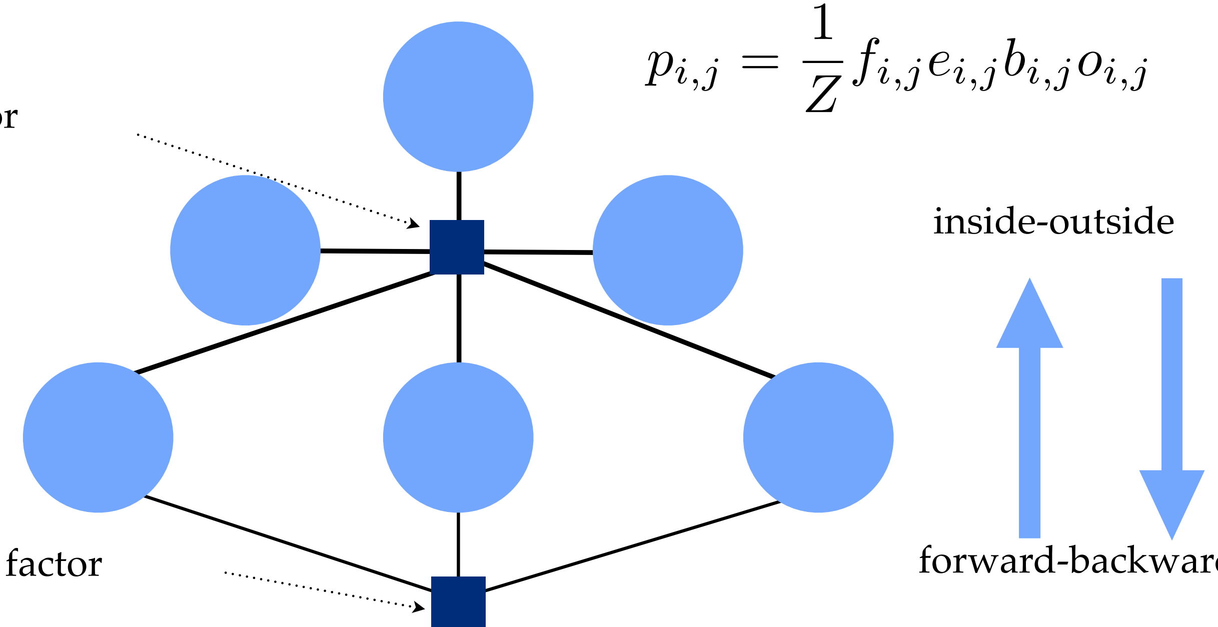
Knowledge Base

Integrated Parsing & Tagging

with belief propagation, dual decomposition and softmax-margin training

Auli & Lopez ACL 2011a,b
Auli & Lopez EMNLP 2011

$$p_{i,j} = \frac{1}{Z} f_{i,j} e_{i,j} b_{i,j} o_{i,j}$$

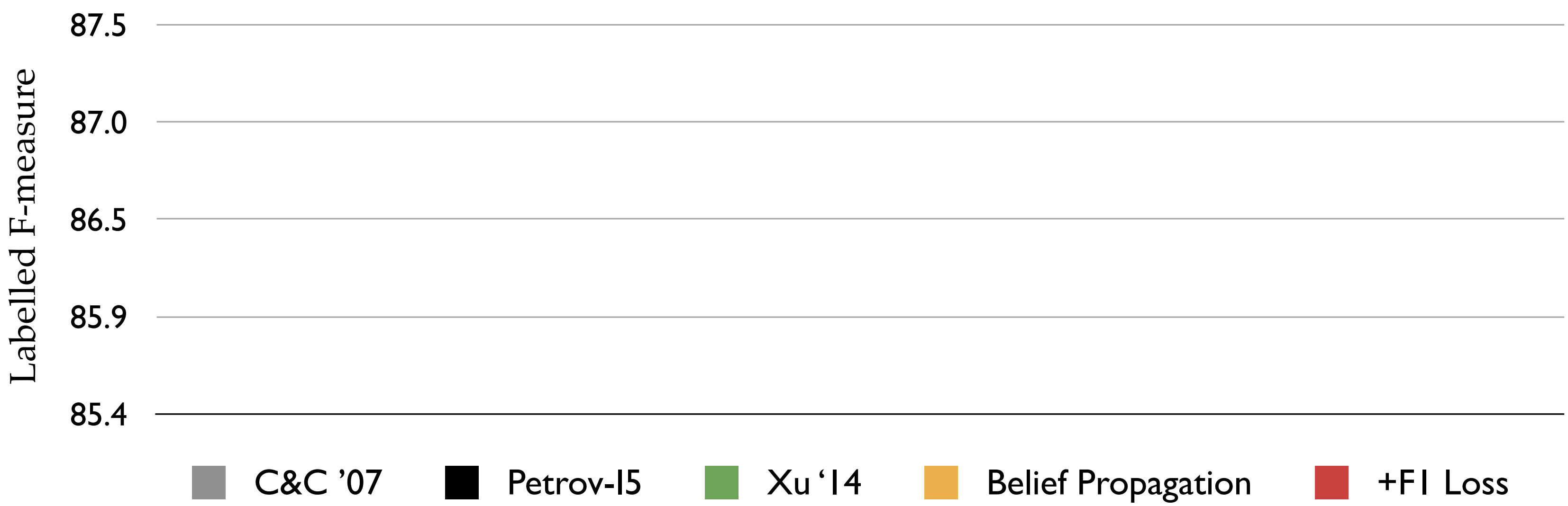


Integrated Parsing & Tagging

Auli & Lopez EMNLP 2011

- F-measure loss for parsing sub-model (+DecF_1).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

best CCG parsing results to date



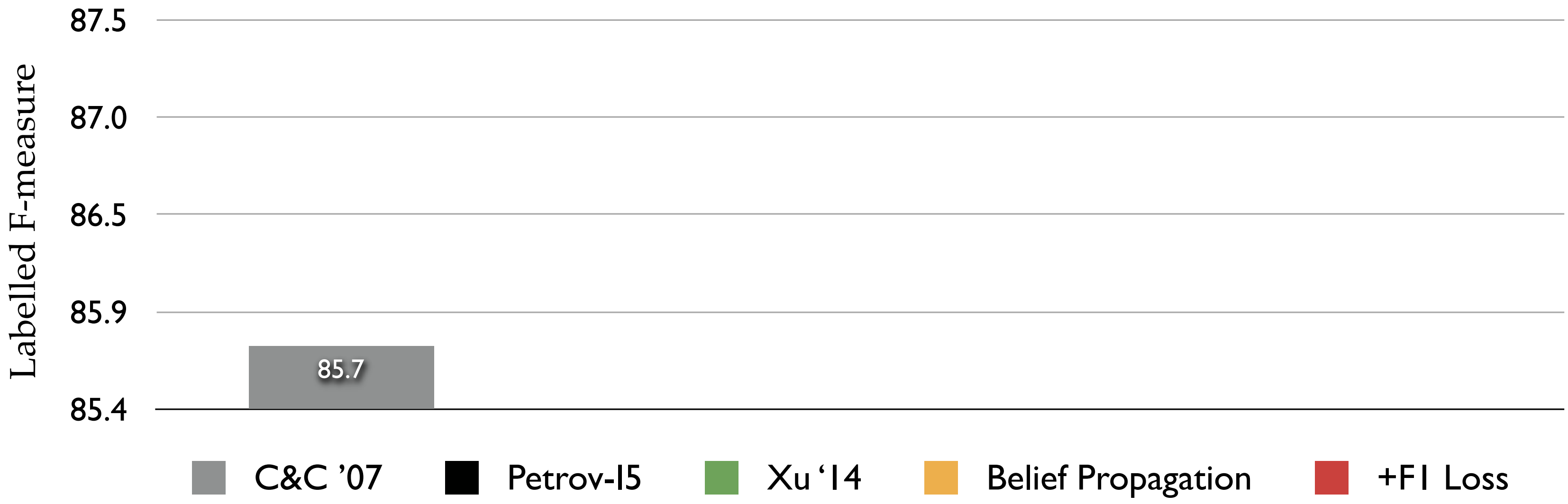
Fowler & Penn (2010)

Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+DecF$_1$).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

Auli & Lopez EMNLP 2011

best CCG parsing results to date



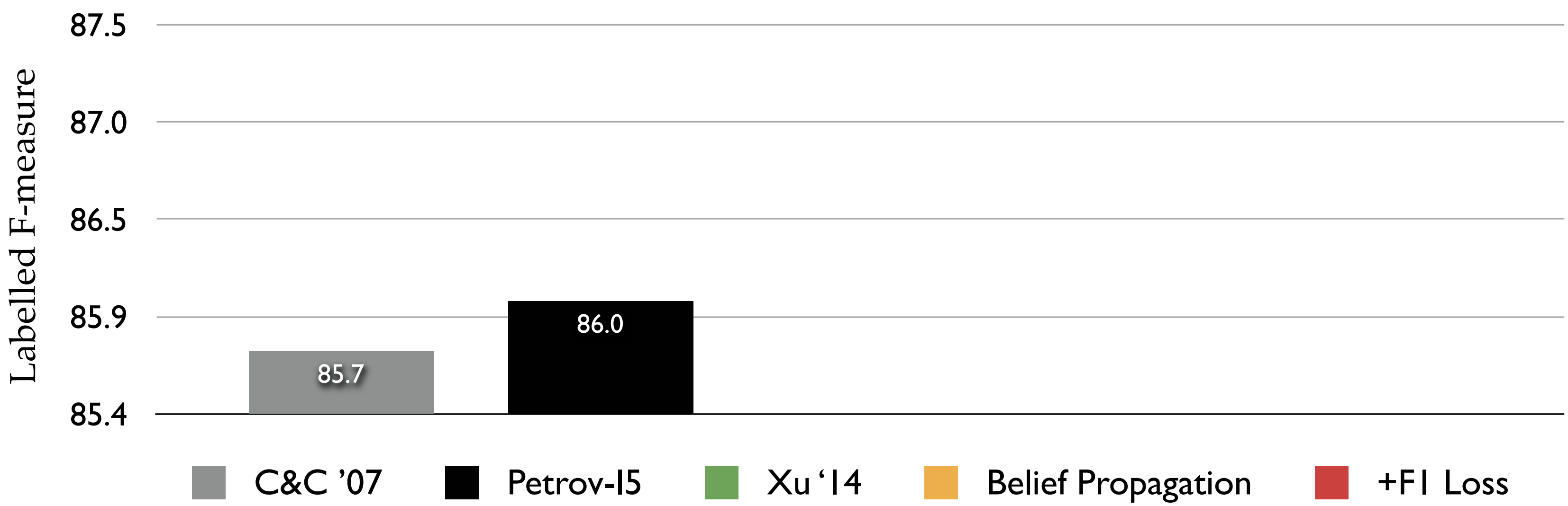
Fowler & Penn (2010)

Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+$DecF_1$).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

Auli & Lopez EMNLP 2011

best CCG parsing results to date



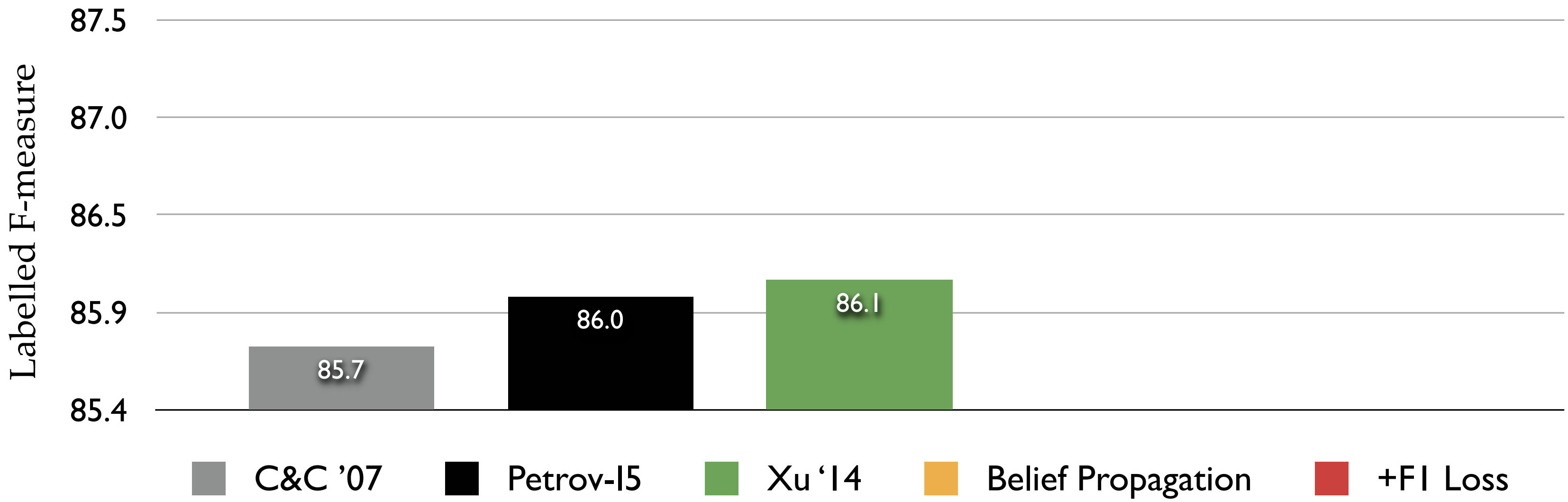
Fowler & Penn (2010)

Integrated Parsing & Tagging

Auli & Lopez EMNLP 2011

- F-measure loss for parsing sub-model (+DecF$_1$).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

best CCG parsing results to date



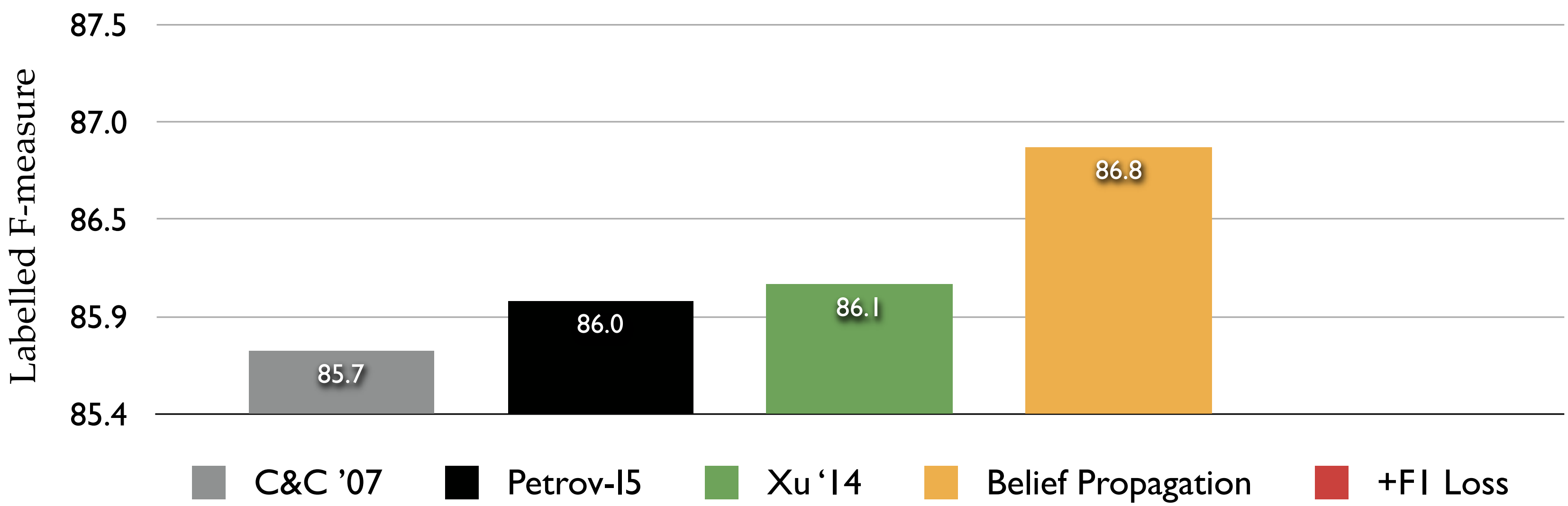
Fowler & Penn (2010)

Integrated Parsing & Tagging

Auli & Lopez EMNLP 2011

- F-measure loss for parsing sub-model (+DecF$_1$).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

best CCG parsing results to date



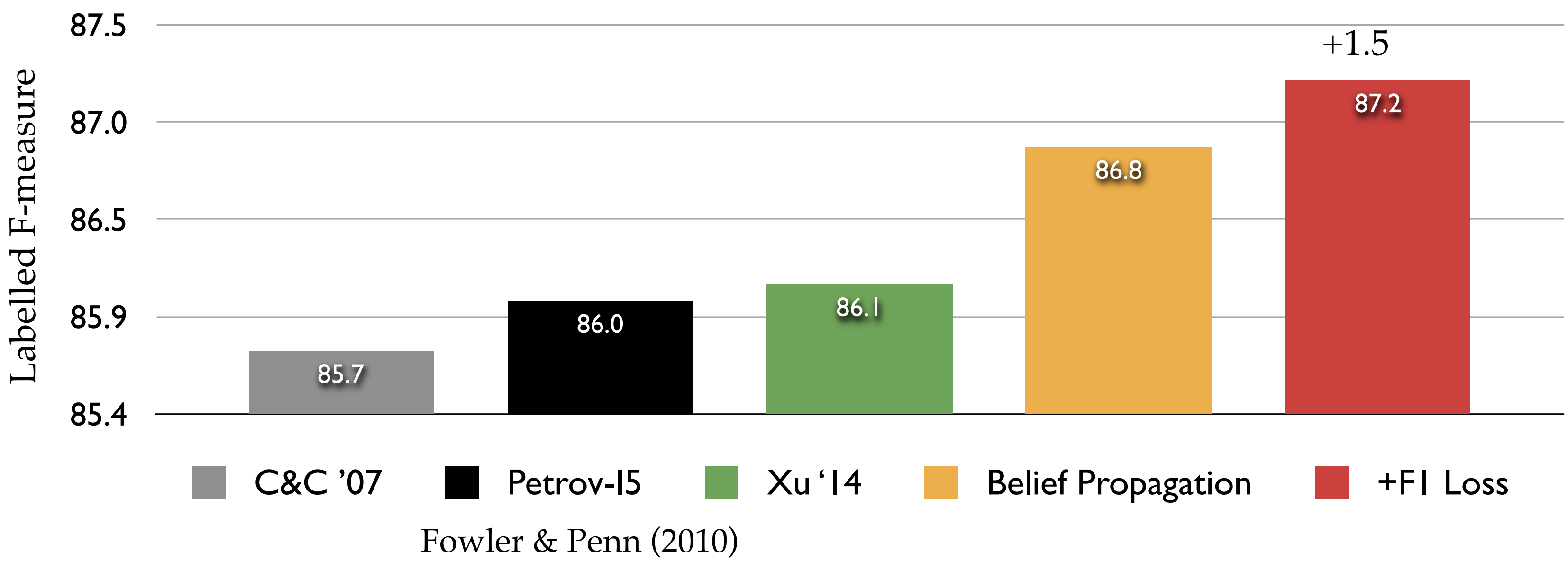
Fowler & Penn (2010)

Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+DecF$_1$).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

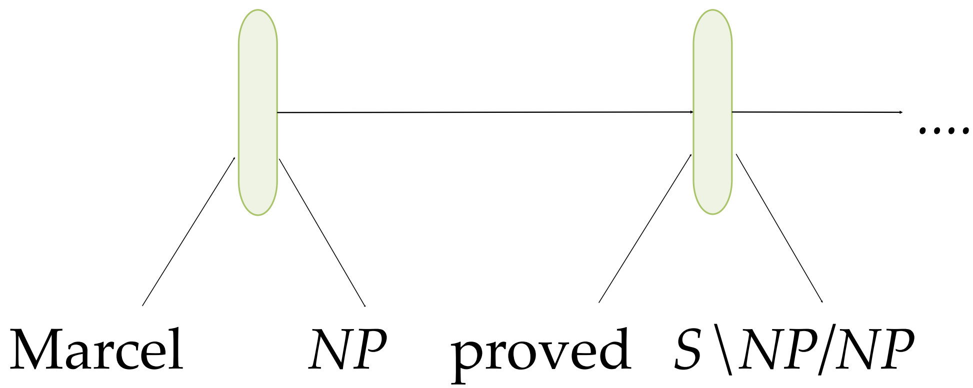
Auli & Lopez EMNLP 2011

best CCG parsing results to date



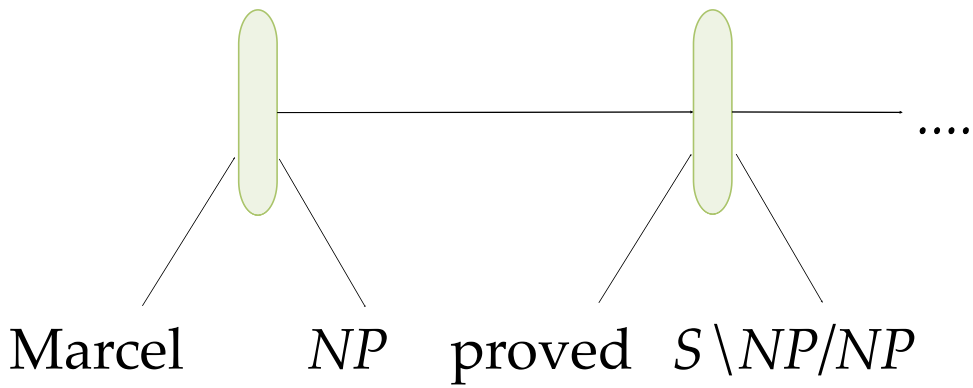
Recurrent nets for CCG supertagging & parsing

with Wenduan Xu 



Recurrent nets for CCG supertagging & parsing

with Wenduan Xu



	tagging	parsing
CRF (Clark & Curran '07)	91.5	85.3
FFN (Lewis & Steedman, '14)	91.5	86.0
RNN	92.3	86.5

Summary

- Two RNN translation models
- Neural nets help most when discrete models sparse
- Task-specific objective gives best performance
- Next: Better modeling of source-side, e.g., bi-directional RNNs, different architectures