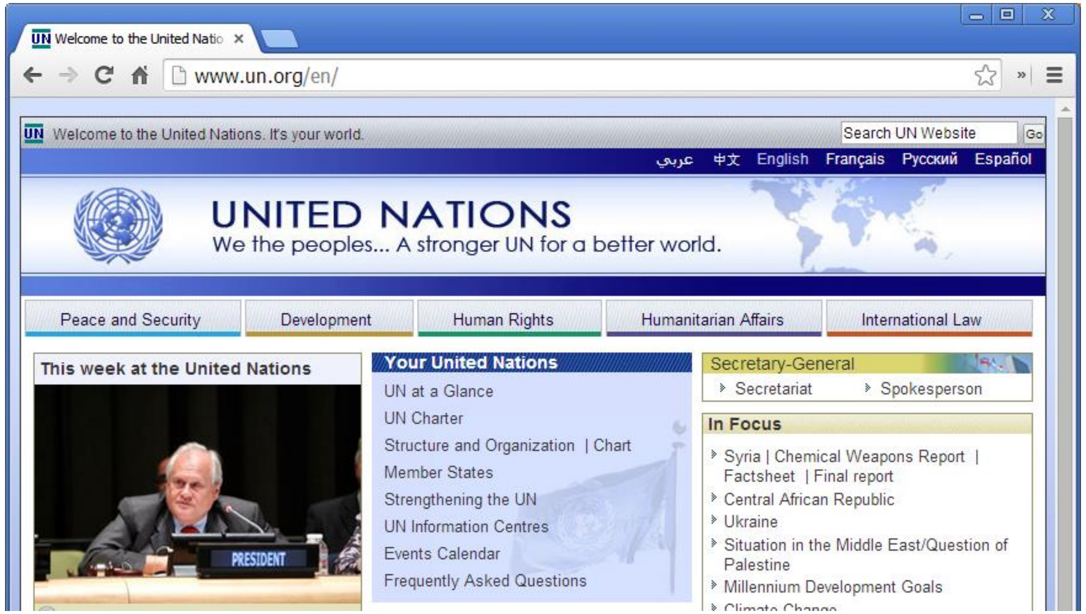
Learning to translate with neural networks

Michael Auli

Microsoft Research

What happened in MT over the past 10 years?



Welcome to the United Natio ×

www.un.org/en/

Welcome to the United Nations. It's your world.

Search UN Website Go

عربي 中文 English Français Русский Español

UNITED NATIONS

We the peoples... A stronger UN for a better world.

Peace and Security | Development | Human Rights | Humanitarian Affairs | International Law

This week at the United Nations

Your United Nations

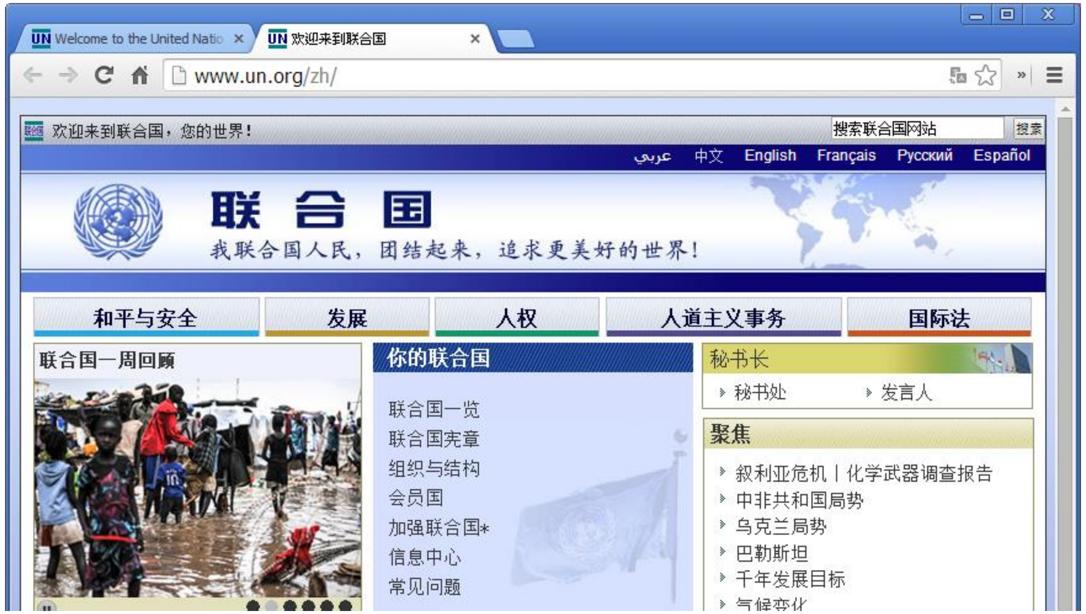
- UN at a Glance
- UN Charter
- Structure and Organization | Chart
- Member States
- Strengthening the UN
- UN Information Centres
- Events Calendar
- Frequently Asked Questions

Secretary-General

- Secretariat
- Spokesperson

In Focus

- Syria | Chemical Weapons Report | Factsheet | Final report
- Central African Republic
- Ukraine
- Situation in the Middle East/Question of Palestine
- Millennium Development Goals
- Climate Change



Welcome to the United Natio × 欢迎来到联合国 ×

www.un.org/zh/

欢迎来到联合国，您的世界！

搜索联合国网站 搜索

عربي 中文 English Français Русский Español

联 合 国

我联合国人民，团结起来，追求更美好的世界！

和平与安全 | 发展 | 人权 | 人道主义事务 | 国际法

联合国一周回顾

你的联合国

- 联合国一览
- 联合国宪章
- 组织与结构
- 会员国
- 加强联合国*
- 信息中心
- 常见问题

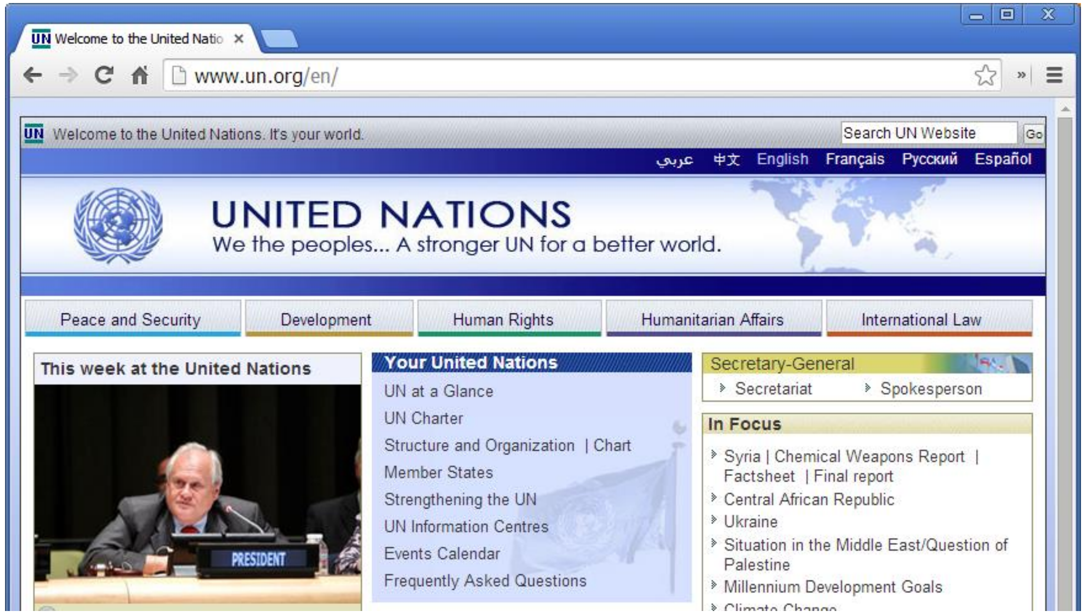
秘书长

- 秘书处
- 发言人

聚焦

- 叙利亚危机 | 化学武器调查报告
- 中非共和国局势
- 乌克兰局势
- 巴勒斯坦
- 千年发展目标
- 气候变化

What happened in MT over the past 10 years?



Welcome to the United Nations. It's your world.

UNITED NATIONS

We the peoples... A stronger UN for a better world.

This week at the United Nations

Your United Nations

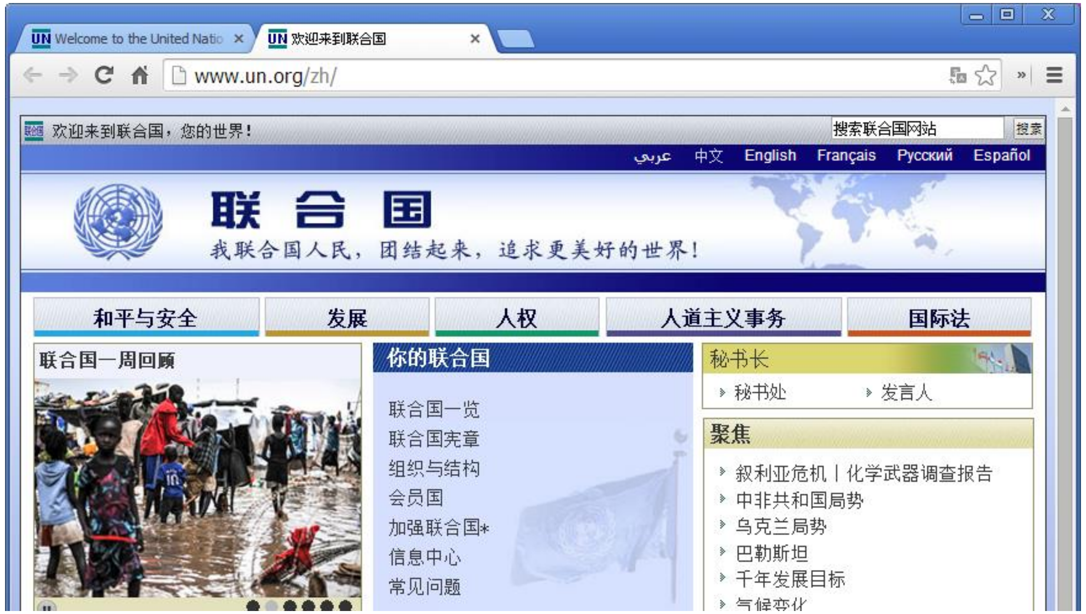
- UN at a Glance
- UN Charter
- Structure and Organization | Chart
- Member States
- Strengthening the UN
- UN Information Centres
- Events Calendar
- Frequently Asked Questions

Secretary-General

- Secretariat
- Spokesperson

In Focus

- Syria | Chemical Weapons Report | Factsheet | Final report
- Central African Republic
- Ukraine
- Situation in the Middle East/Question of Palestine
- Millennium Development Goals
- Climate Change



欢迎来到联合国，您的世界！

联合国

我联合国人民，团结起来，追求更美好的世界！

联合国一周回顾

你的联合国

- 联合国一览
- 联合国宪章
- 组织与结构
- 会员国
- 加强联合国*
- 信息中心
- 常见问题

秘书长

- 秘书处
- 发言人

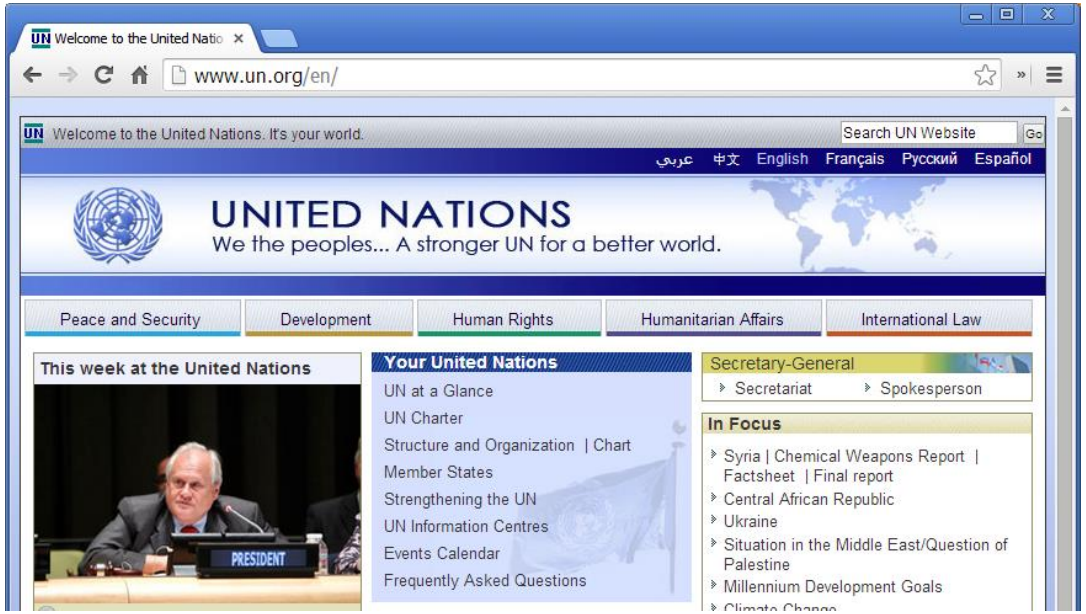
聚焦

- 叙利亚危机 | 化学武器调查报告
- 中非共和国局势
- 乌克兰局势
- 巴勒斯坦
- 千年发展目标
- 气候变化

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

(Lopez & Post, 2013)

What happened in MT over the past 10 years?



Welcome to the United Nations. It's your world. Search UN Website Go

عربي 中文 English Français Русский Español

UNITED NATIONS

We the peoples... A stronger UN for a better world.

Peace and Security | Development | Human Rights | Humanitarian Affairs | International Law

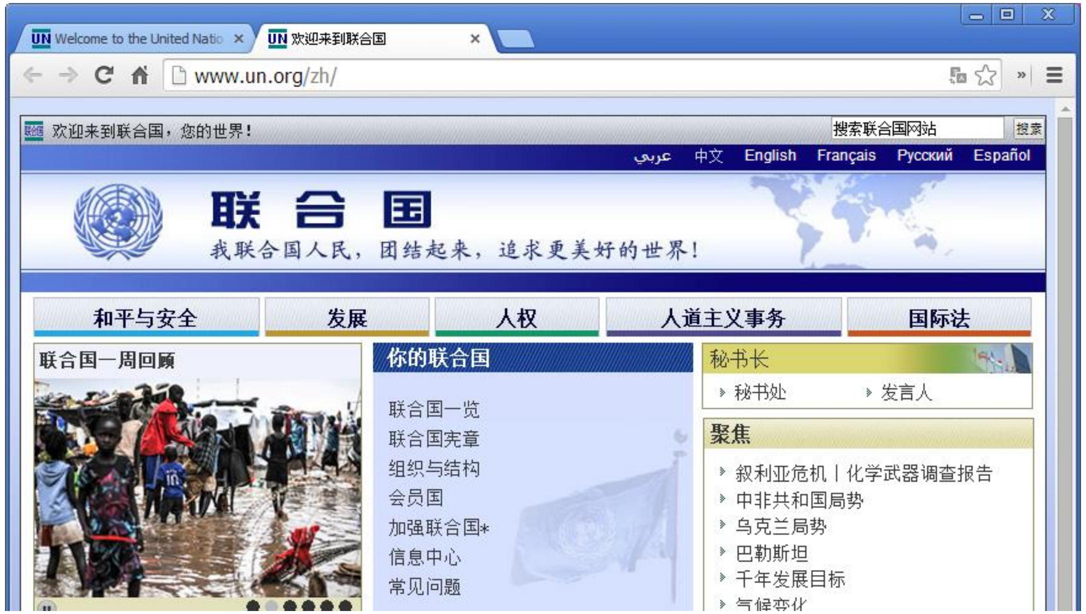
This week at the United Nations

PRESIDENT

Your United Nations
- UN at a Glance
- UN Charter
- Structure and Organization | Chart
- Member States
- Strengthening the UN
- UN Information Centres
- Events Calendar
- Frequently Asked Questions

Secretary-General
- Secretariat
- Spokesperson

In Focus
- Syria | Chemical Weapons Report | Factsheet | Final report
- Central African Republic
- Ukraine
- Situation in the Middle East/Question of Palestine
- Millennium Development Goals
- Climate Change



欢迎来到联合国，您的世界！ 搜索联合国网站 搜索

عربي 中文 English Français Русский Español

联 合 国

我联合国人民，团结起来，追求更美好的世界！

和平与安全 | 发展 | 人权 | 人道主义事务 | 国际法

联合国一周回顾

你的联合国
- 联合国一览
- 联合国宪章
- 组织与结构
- 会员国
- 加强联合国*
- 信息中心
- 常见问题

秘书长
- 秘书处
- 发言人

聚焦
- 叙利亚危机 | 化学武器调查报告
- 中非共和国局势
- 乌克兰局势
- 巴勒斯坦
- 千年发展目标
- 气候变化

“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

Phrase-based Translation

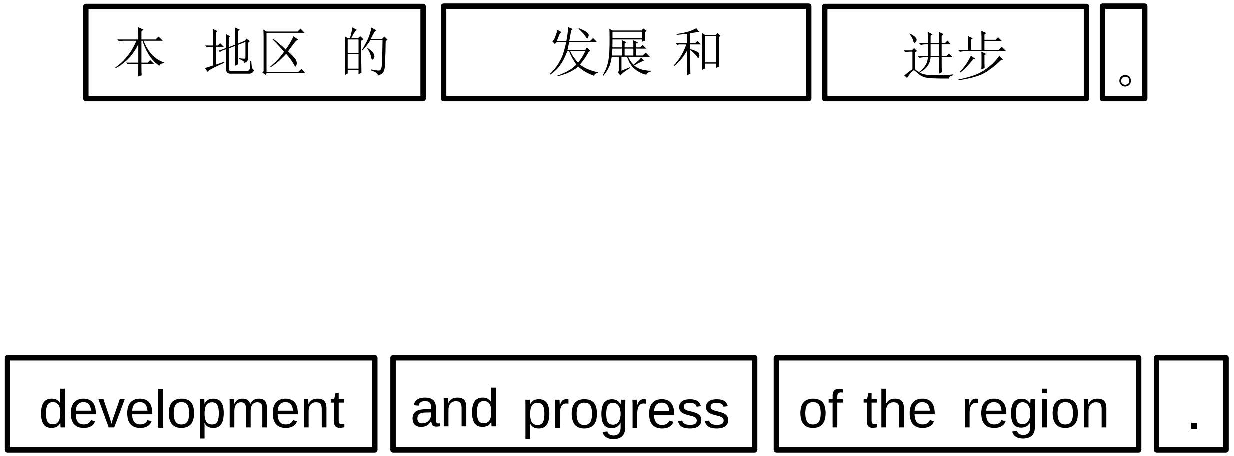
Koehn et al. (2003)

本 地区 的 发展 和 进步 。

development and progress of the region .

Phrase-based Translation

Koehn et al. (2003)

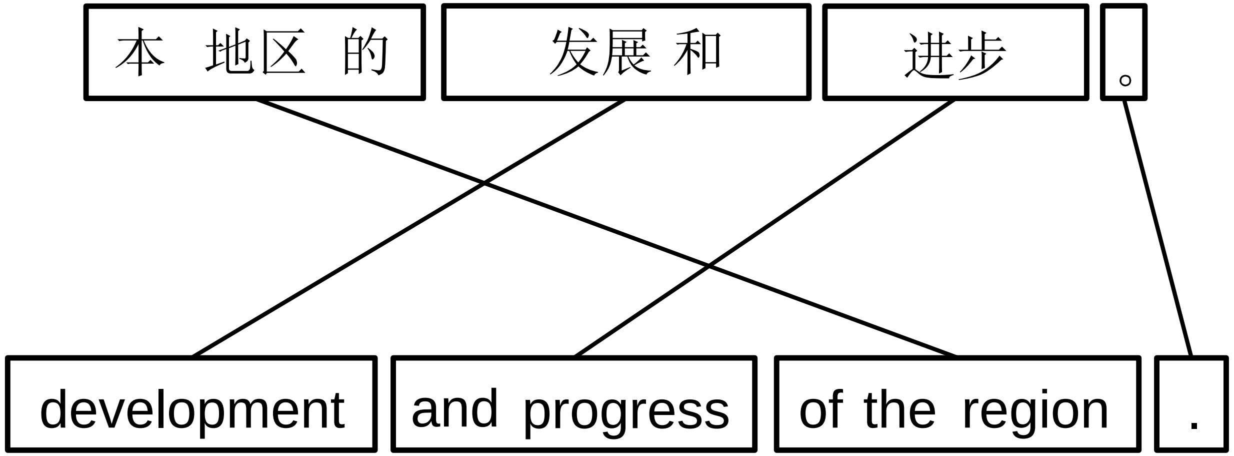


本 地区 的 | 发展 和 | 进步 | 。

development | and progress | of the region | .

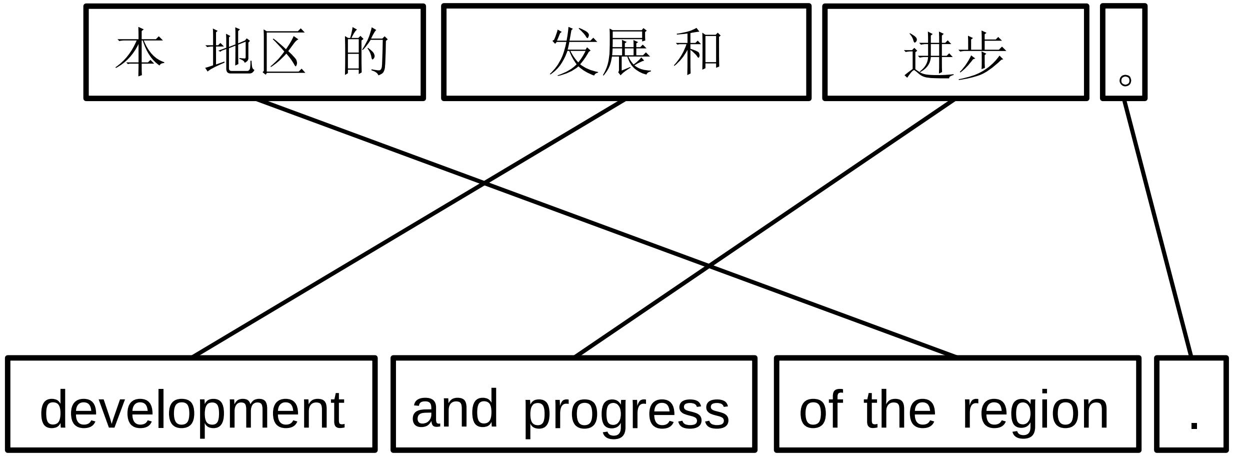
Phrase-based Translation

Koehn et al. (2003)



Phrase-based Translation

Koehn et al. (2003)



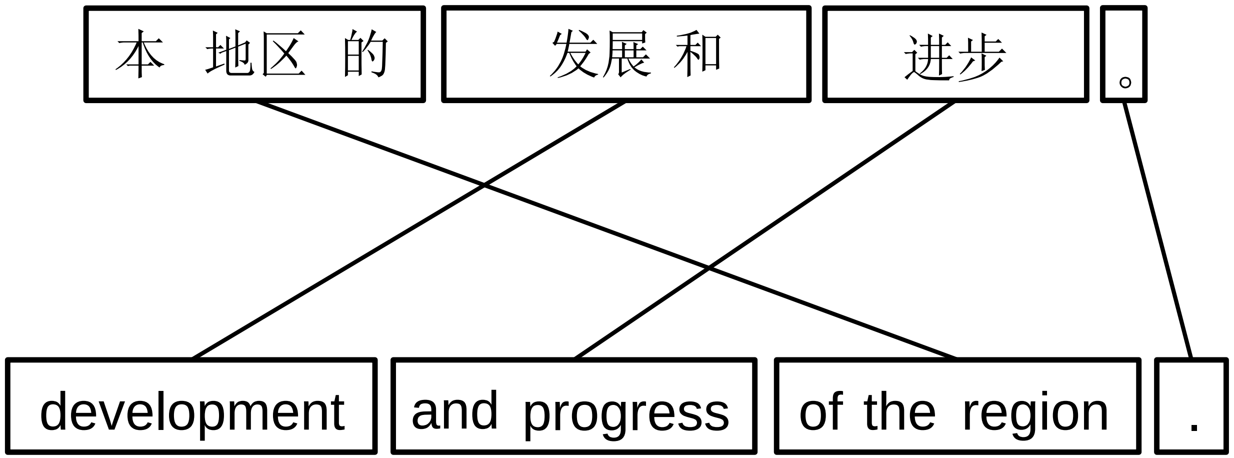
本 地区 的 → of the region

发展 → development

和 进步 → and progress

Phrase-based Translation

Koehn et al. (2003)


本 地区 的

发展 和

进步

。

development

and progress

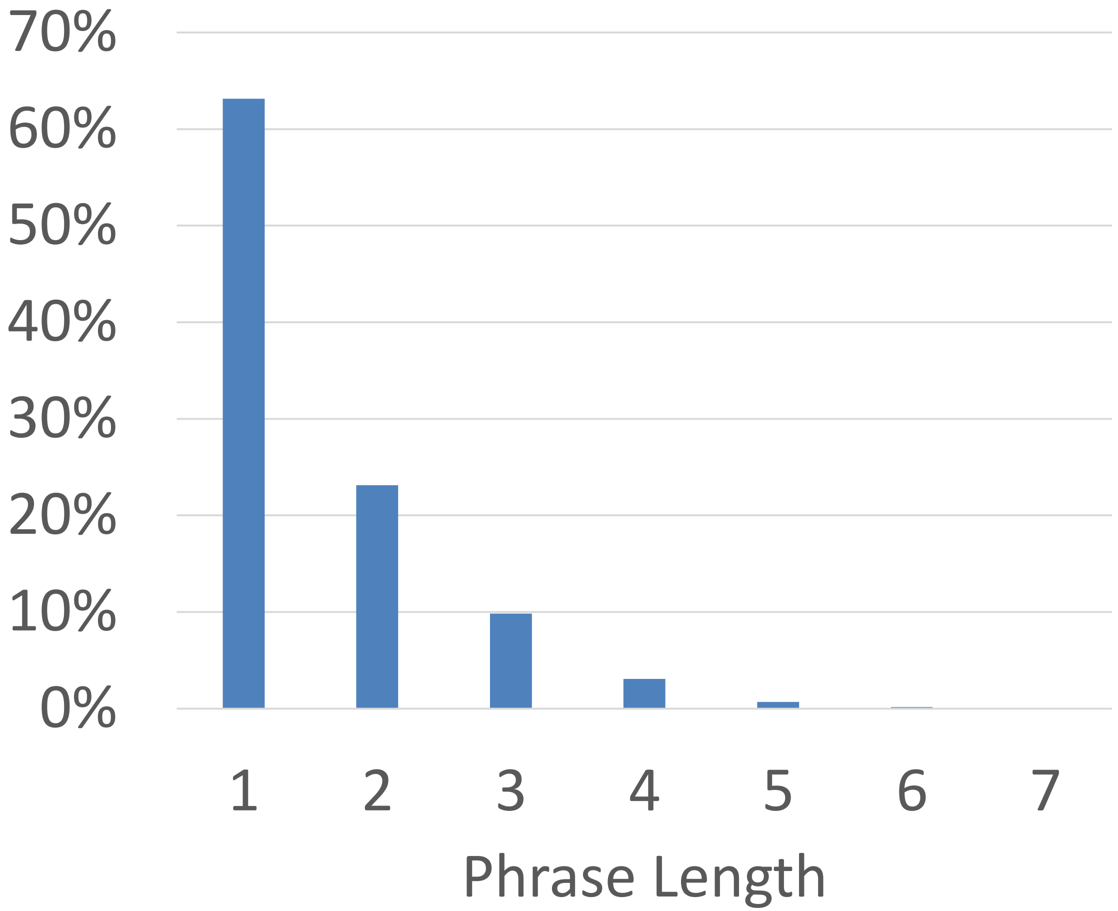
of the region

.

本 地区 的 → of the region

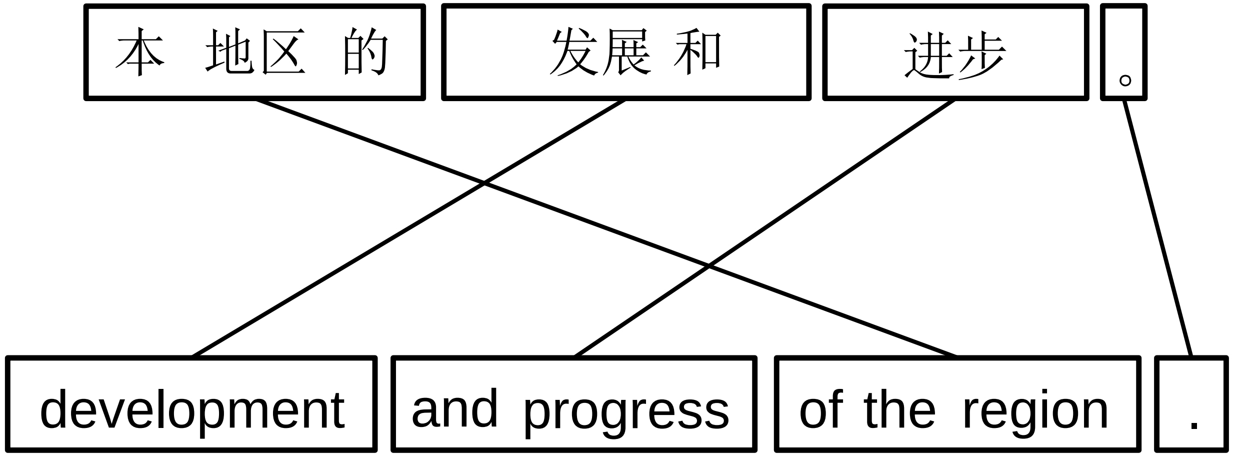
发展 → development

和 进步 → and progress



Phrase-based Translation

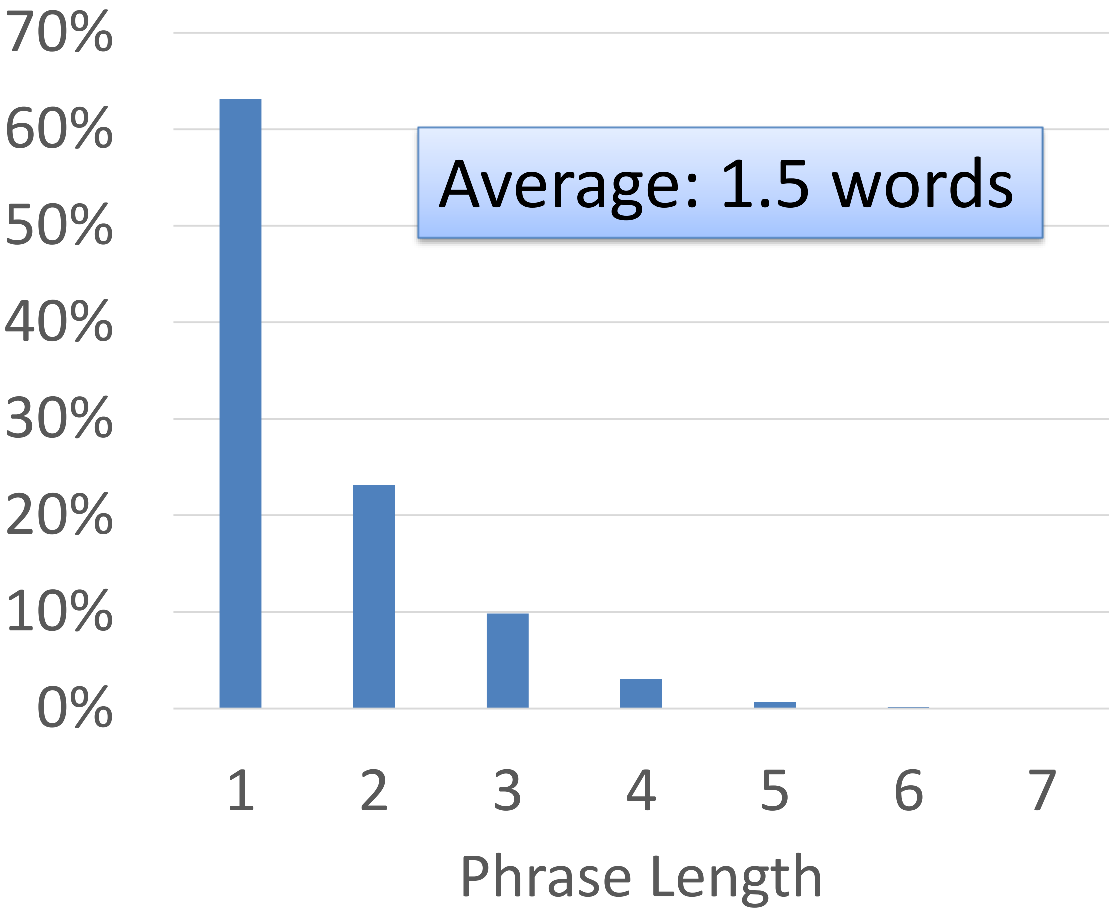
Koehn et al. (2003)



本 地区 的 → of the region

发展 → development

和 进步 → and progress



Average: 1.5 words

n-gram Language Modeling

Kneser & Ney (1996)

p(progress in the region) =

n-gram Language Modeling

Kneser & Ney (1996)

p(progress in the region) =

Train data:

…
development and progress of the region
…

n-gram Language Modeling

Kneser & Ney (1996)

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

Train data:

…
development and progress of
the region
…

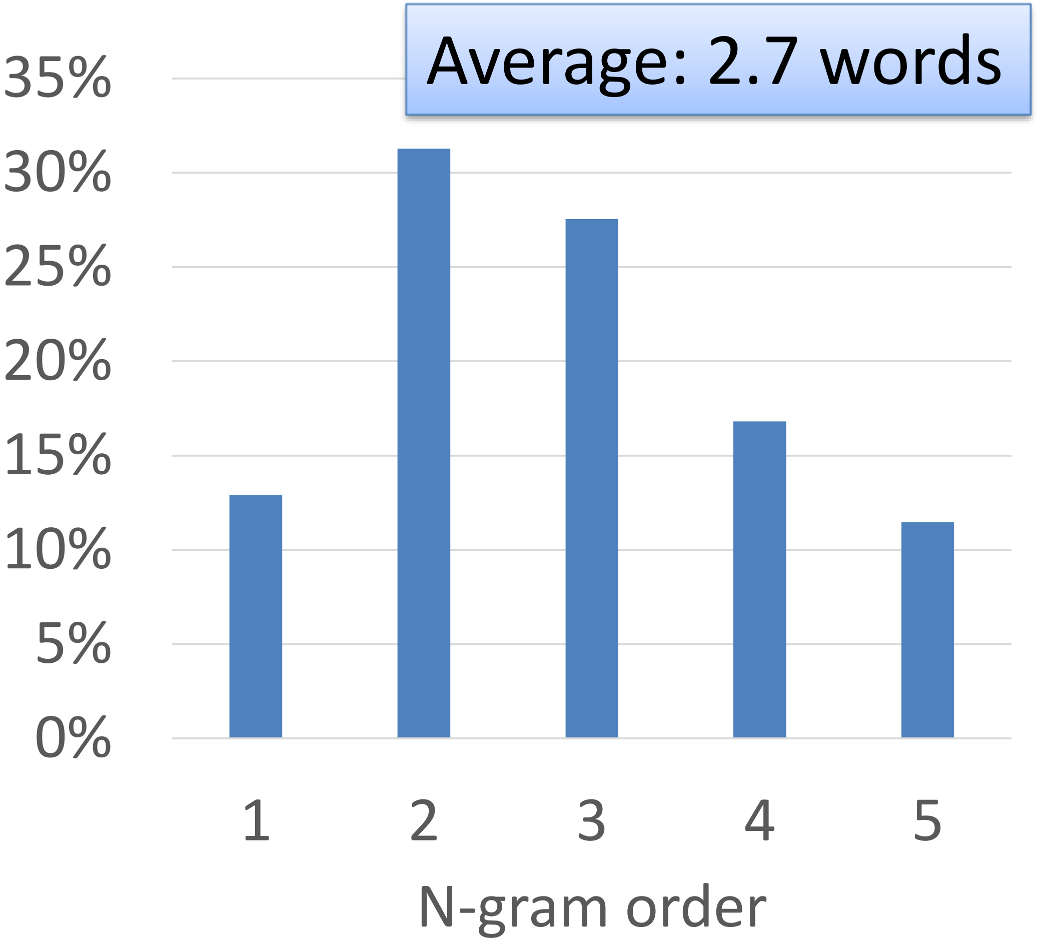
n-gram Language Modeling

Kneser & Ney (1996)

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

Train data:

…
development and progress of the region
…



Does not include out-of-vocabulary tokens

How can we improve on this?

- Or: how to capture relationships beyond 1.5 - 2.7 words
- Neural networks: From discrete to distributional representations
- Recurrent nets: From fixed length contexts to unbounded histories

Overview

- Recurrent neural network joint models (Auli et al., EMNLP 2013)
  Combined language and translation modeling

- Minimum translation modeling with recurrent nets (Hu et al., EACL 2014)
  Sequence models over bilingual units

- Training recurrent nets (Auli & Gao, ACL 2014)
  Expected BLEU training for neural network translation models

- Large-scale discriminative sparse ordering models (Auli et al., in submission)
  Training millions of linear ordering features with expected BLEU

Overview

- Recurrent neural network joint models (Auli et al., EMNLP 2013)
  Combined language and translation modeling

- Minimum translation modeling with recurrent nets (Hu et al., EACL 2014)
  Sequence models over bilingual units

- Training recurrent nets (Auli & Gao, ACL 2014)
  Expected BLEU training for neural network translation models

- Large-scale discriminative sparse ordering models (Auli et al., in submission)
  Training millions of linear ordering features with expected BLEU

Feed-forward Network

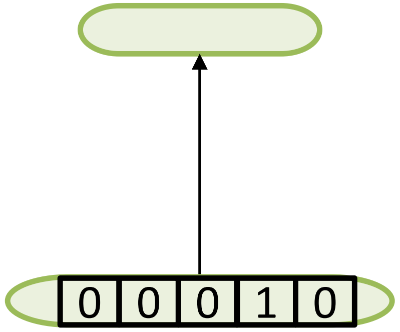
and

Feed-forward Network

0	0	0	1	0

and

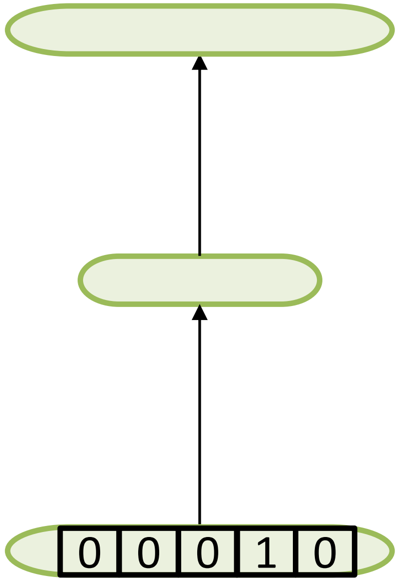
Feed-forward Network



0	0	0	1	0

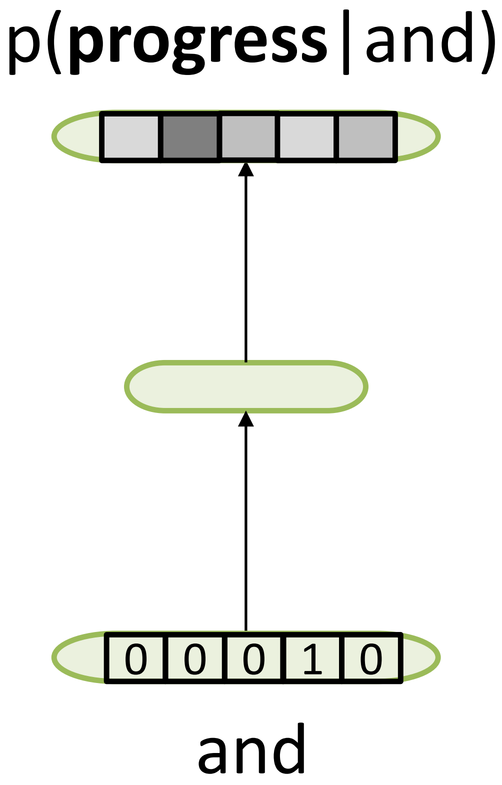
and

Feed-forward Network

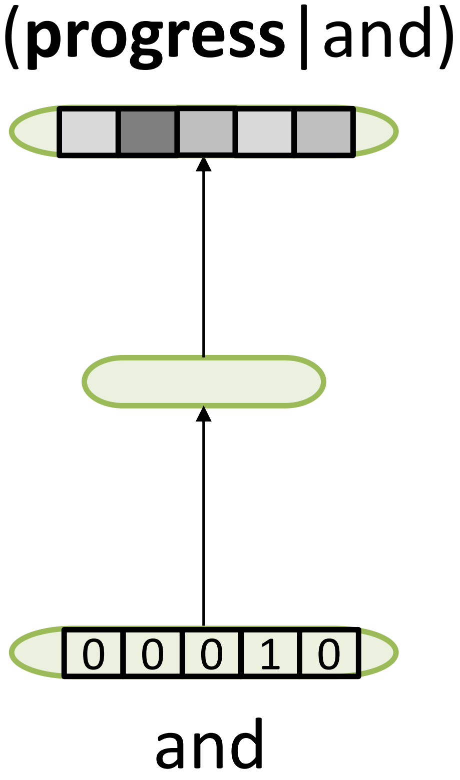


and

Feed-forward Network



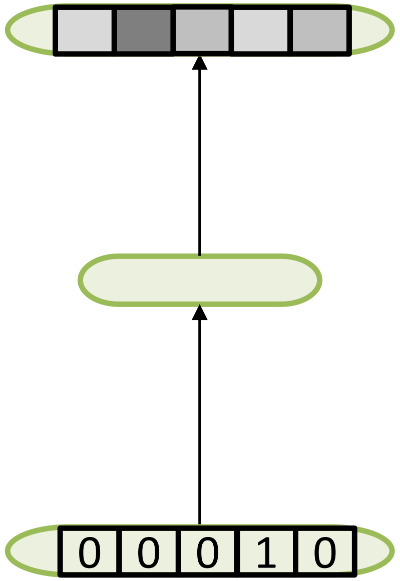
Feed-forward Network



Still based on limited context!

Recurrent Network

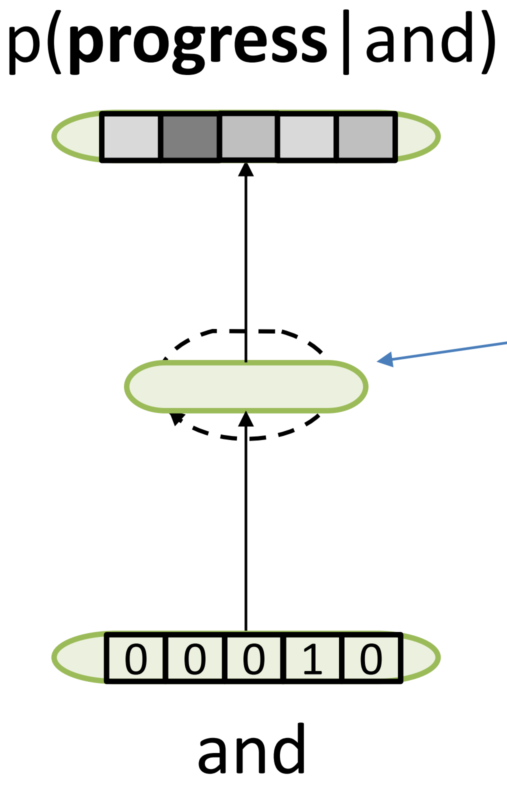
p(**progress**|and)



0	0	0	1	0

and

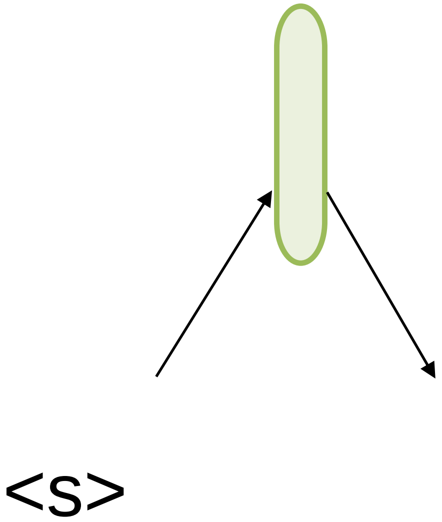
Recurrent Network



Dependence on
previous time steps

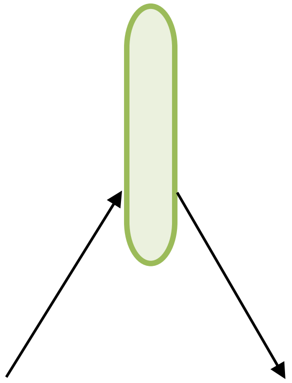
Recurrent Network

Recurrent Network



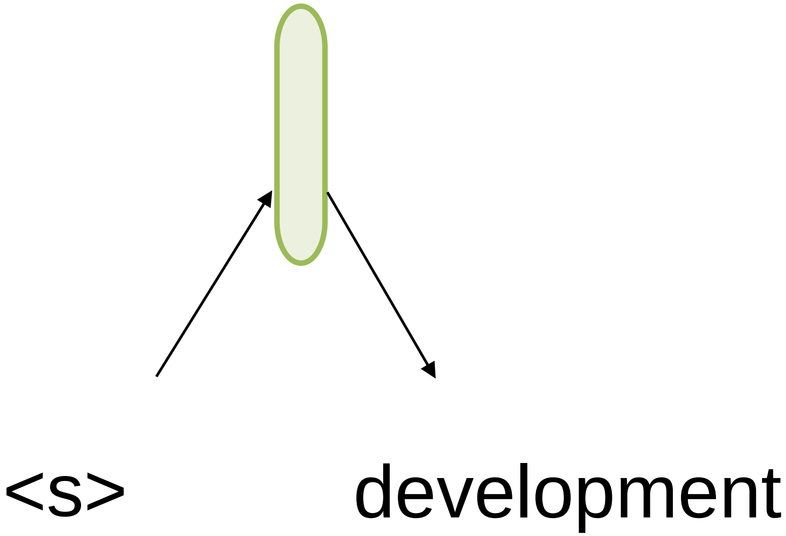
<s>

Recurrent Network



<s> p(development|<s>)

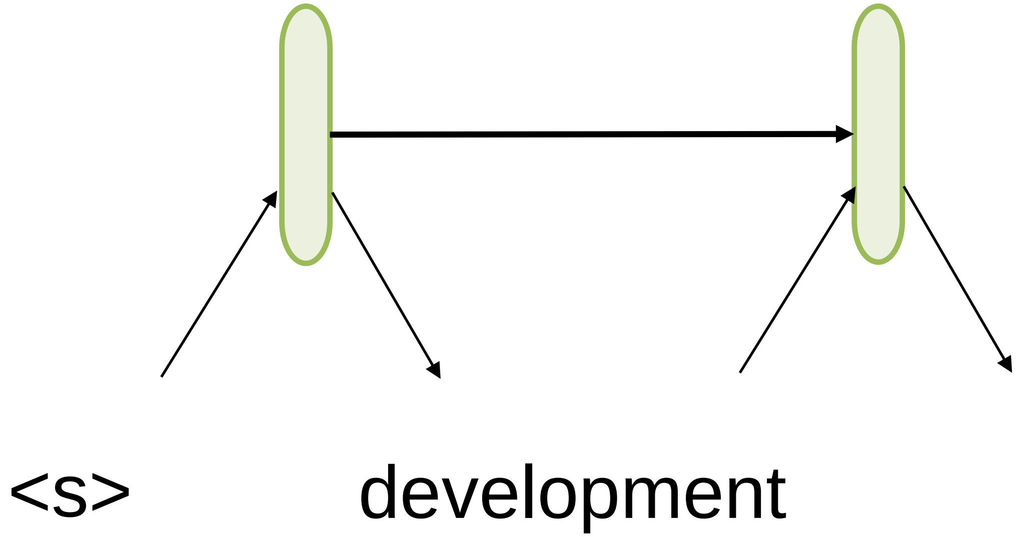
Recurrent Network



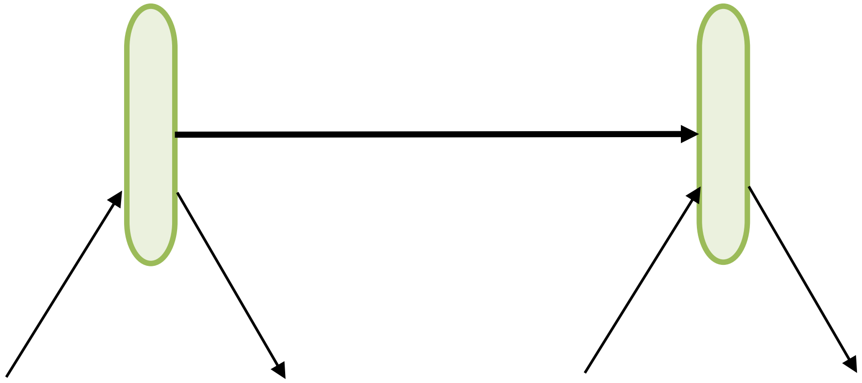
<s>

development

Recurrent Network

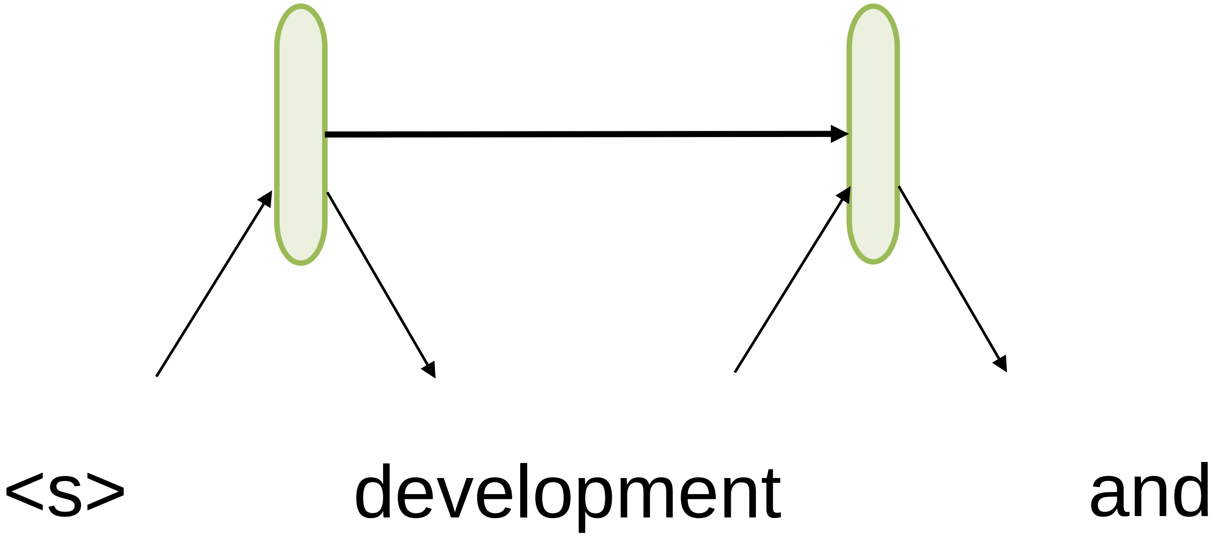


Recurrent Network



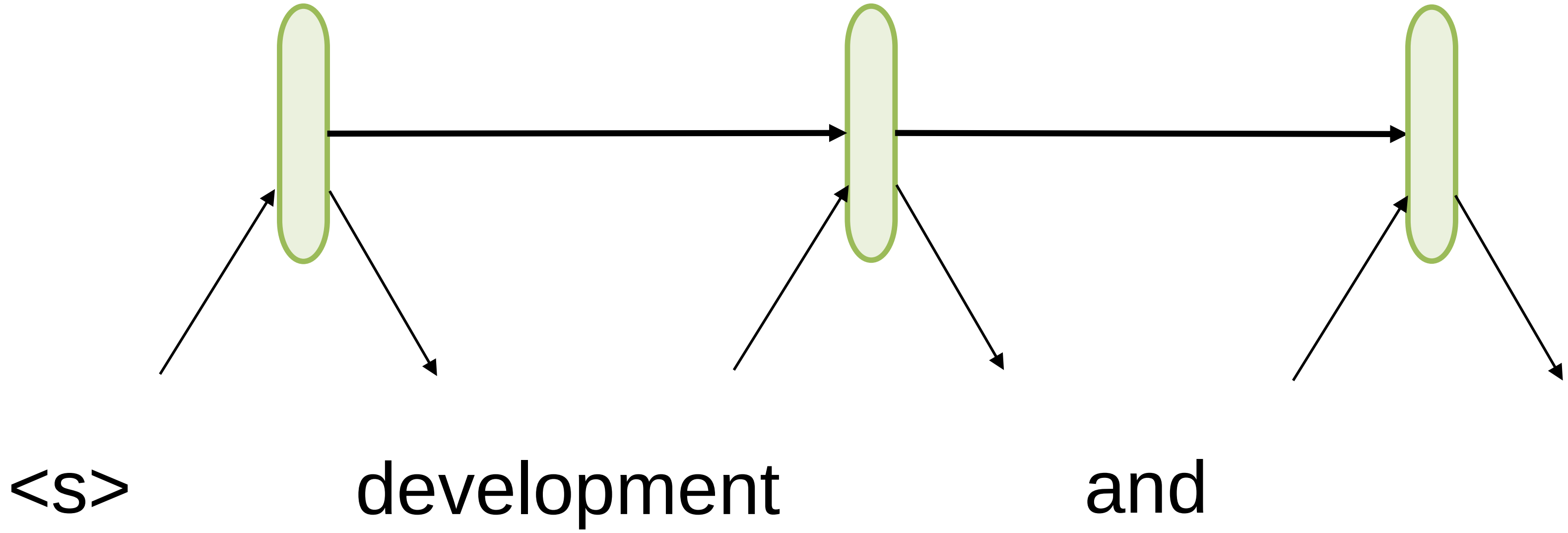
<s> development p(and|development, <s>)

Recurrent Network

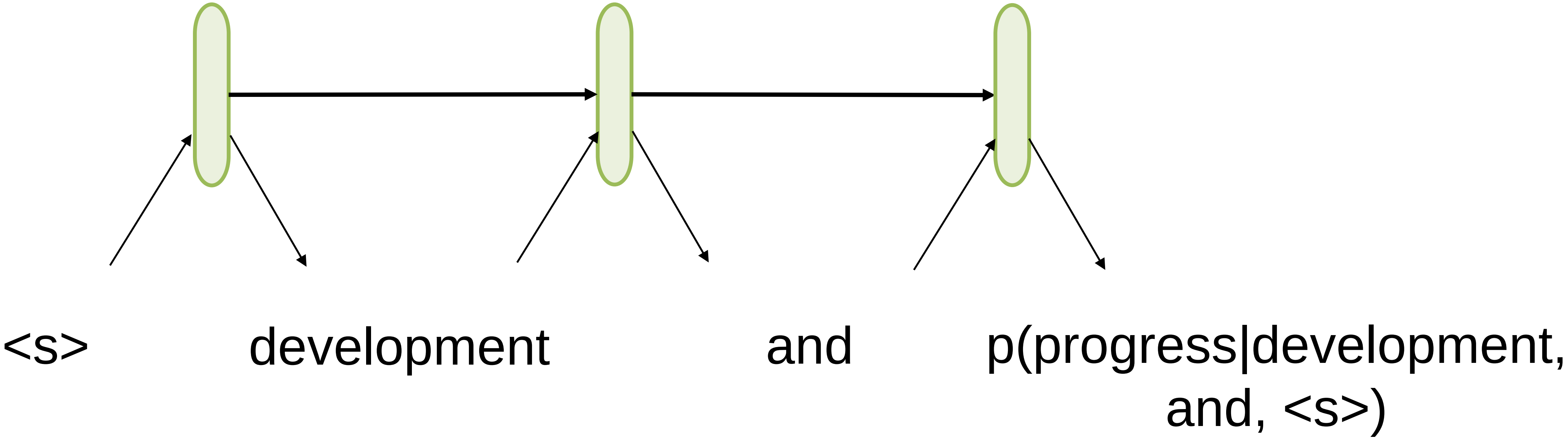


<s> development and

Recurrent Network



Recurrent Network



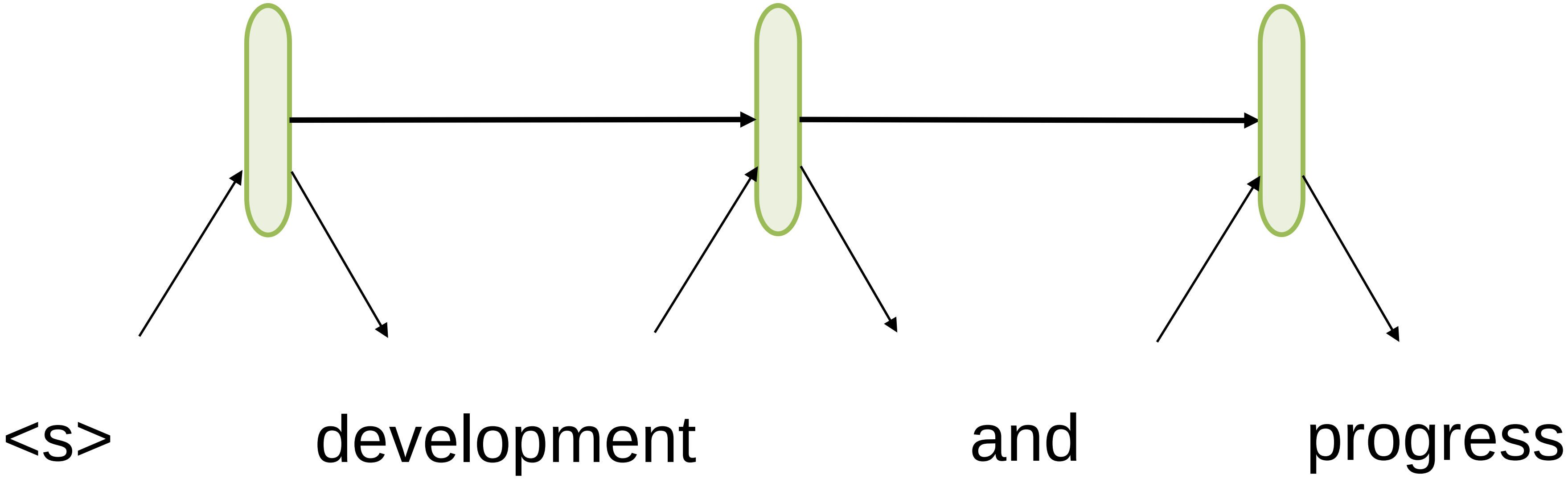
<s>

development

and

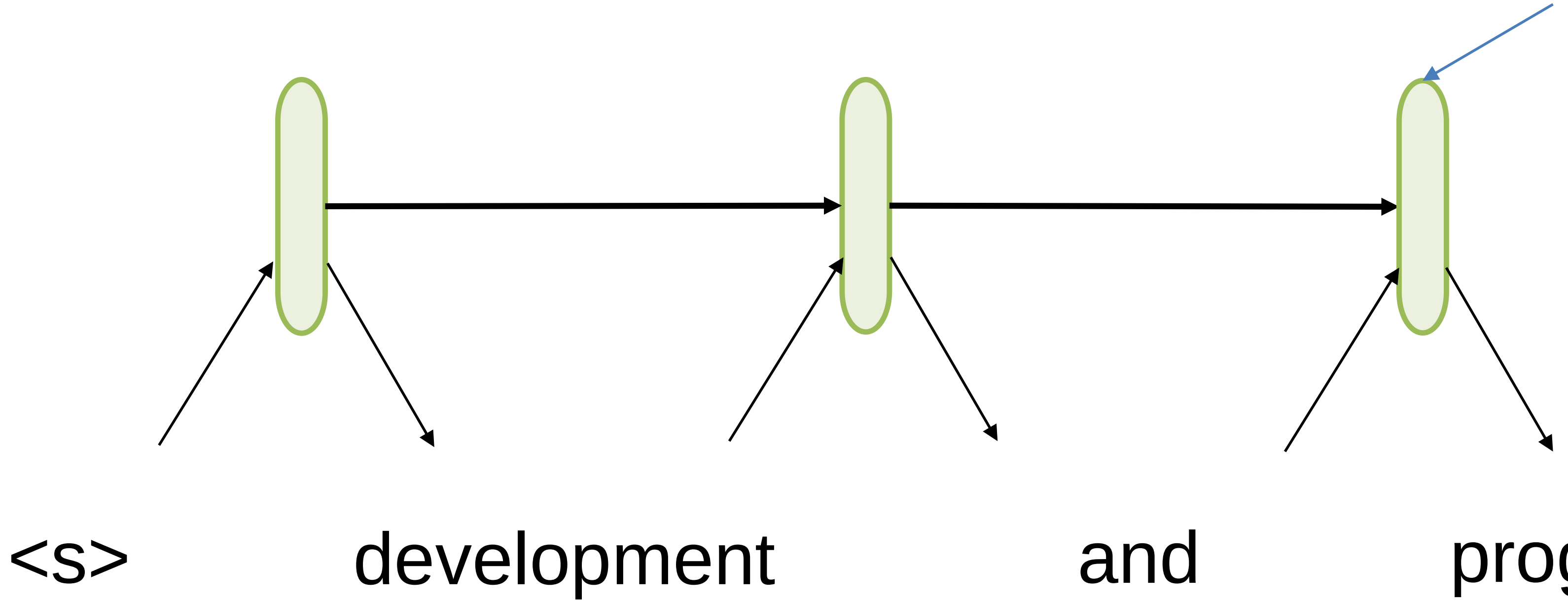
p(progress|development,
and, <s>)

Recurrent Network



<s> development and progress

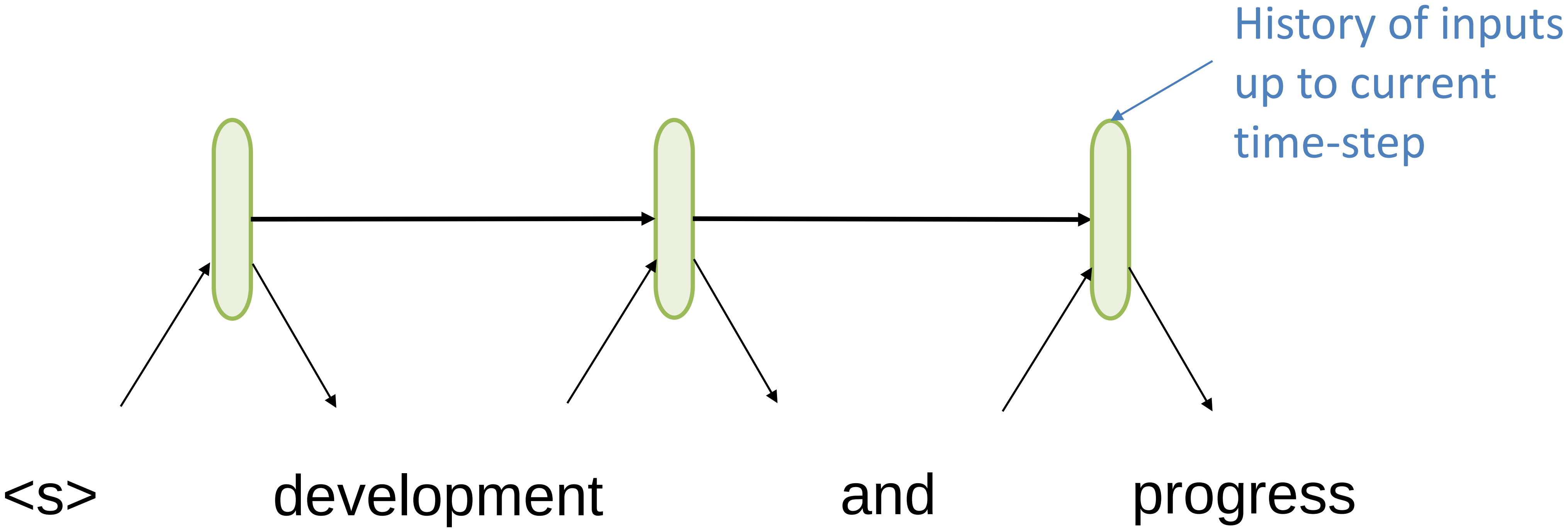
Recurrent Network



History of inputs up to current time-step

<s> development and progress

Recurrent Network



History of inputs up to current time-step

<s> development and progress

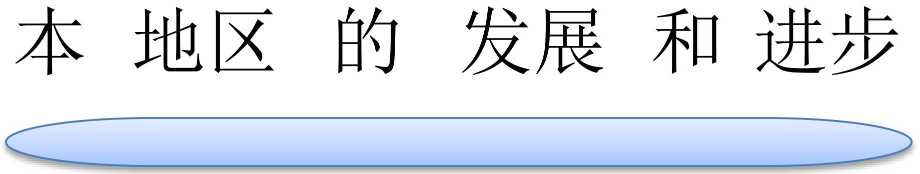
State of the art in language modeling (Mikolov 2011)

More accurate than feed-forward nets (Sundermeyer 2013)

Recurrent Network Joint Model

本 地区 的 发展 和 进步

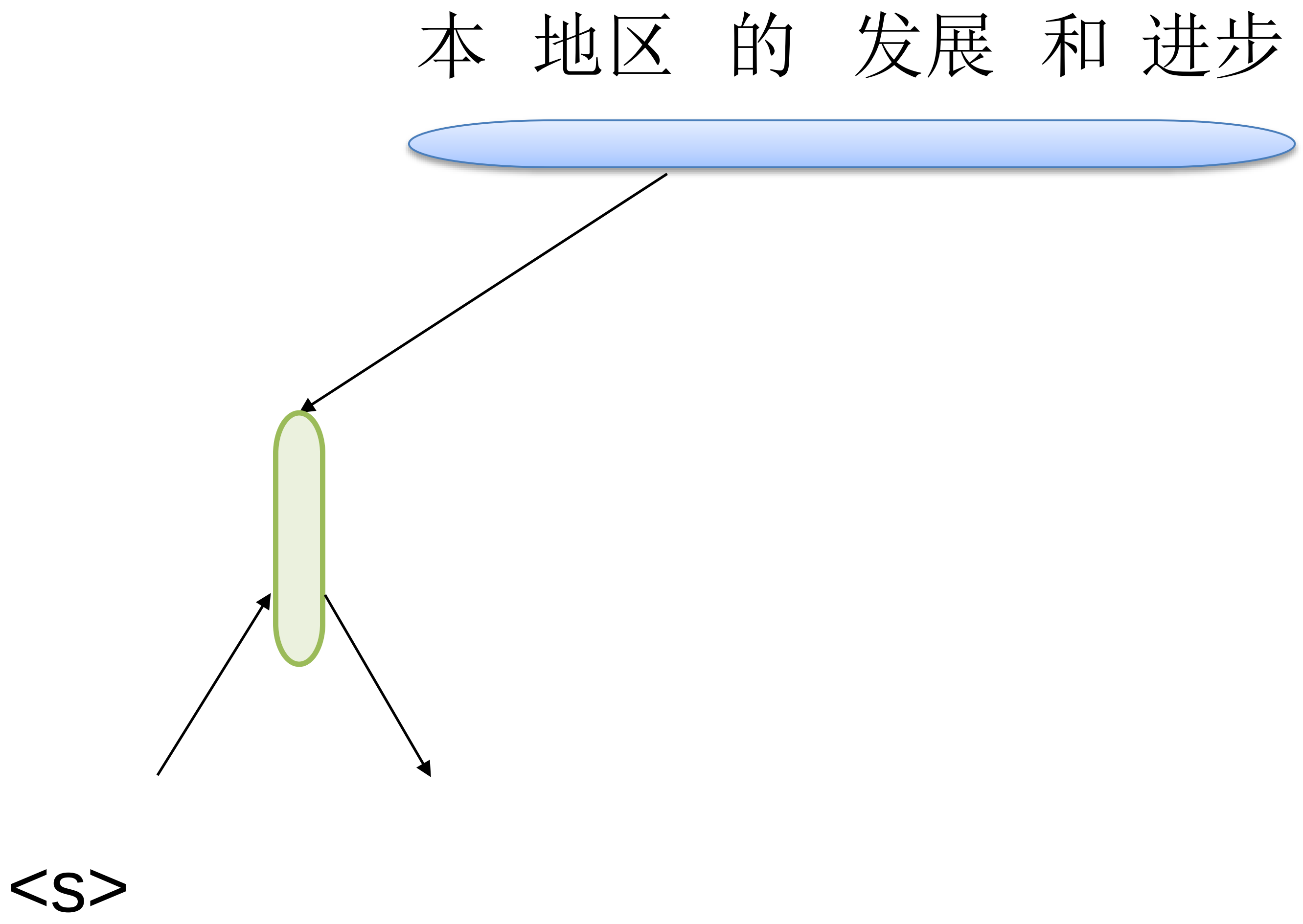
Recurrent Network Joint Model



本 地区 的 发展 和 进步

Entire source sentence representation

Recurrent Network Joint Model

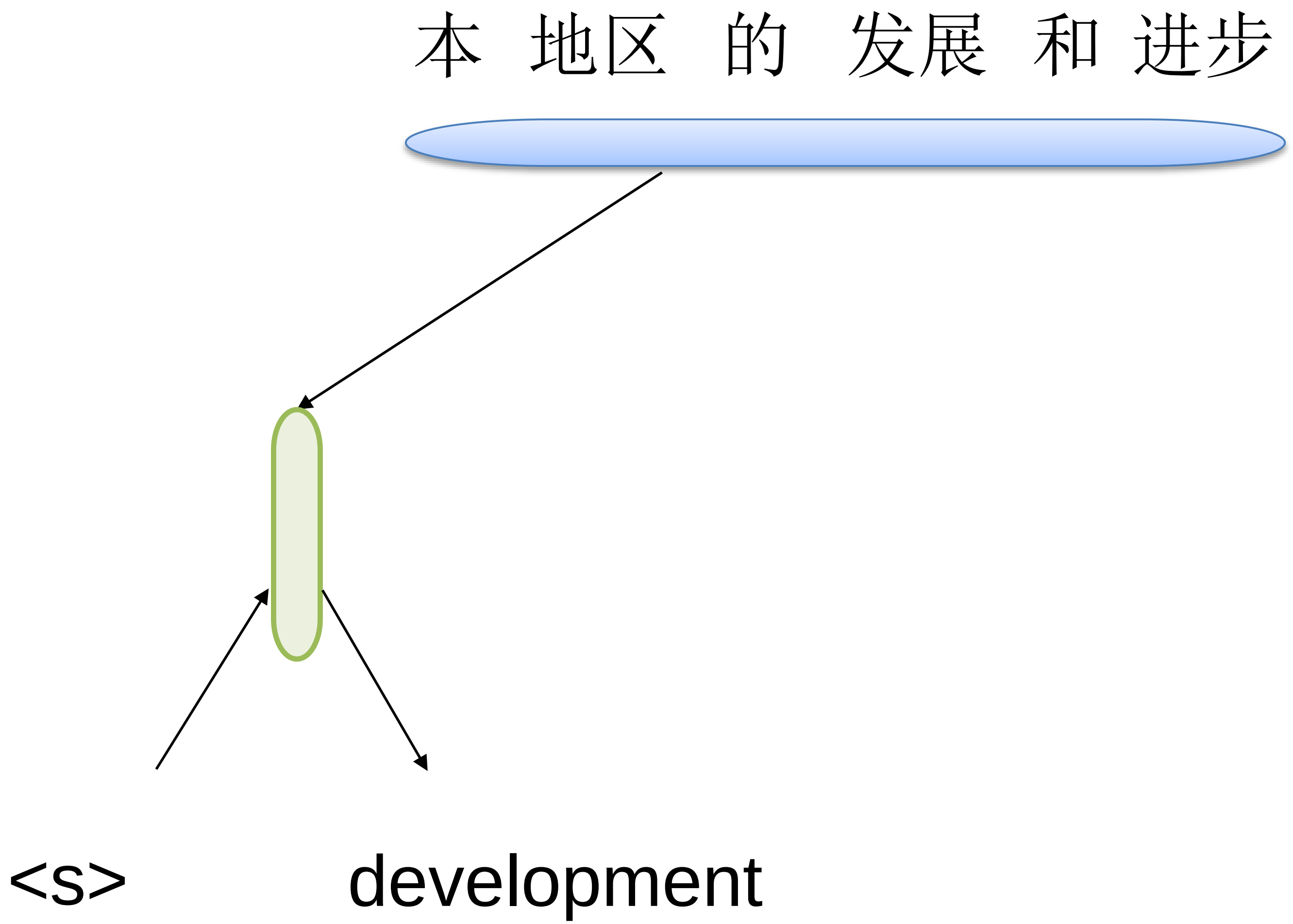


本 地区 的 发展 和 进步

<s>

Entire source sentence representation

Recurrent Network Joint Model



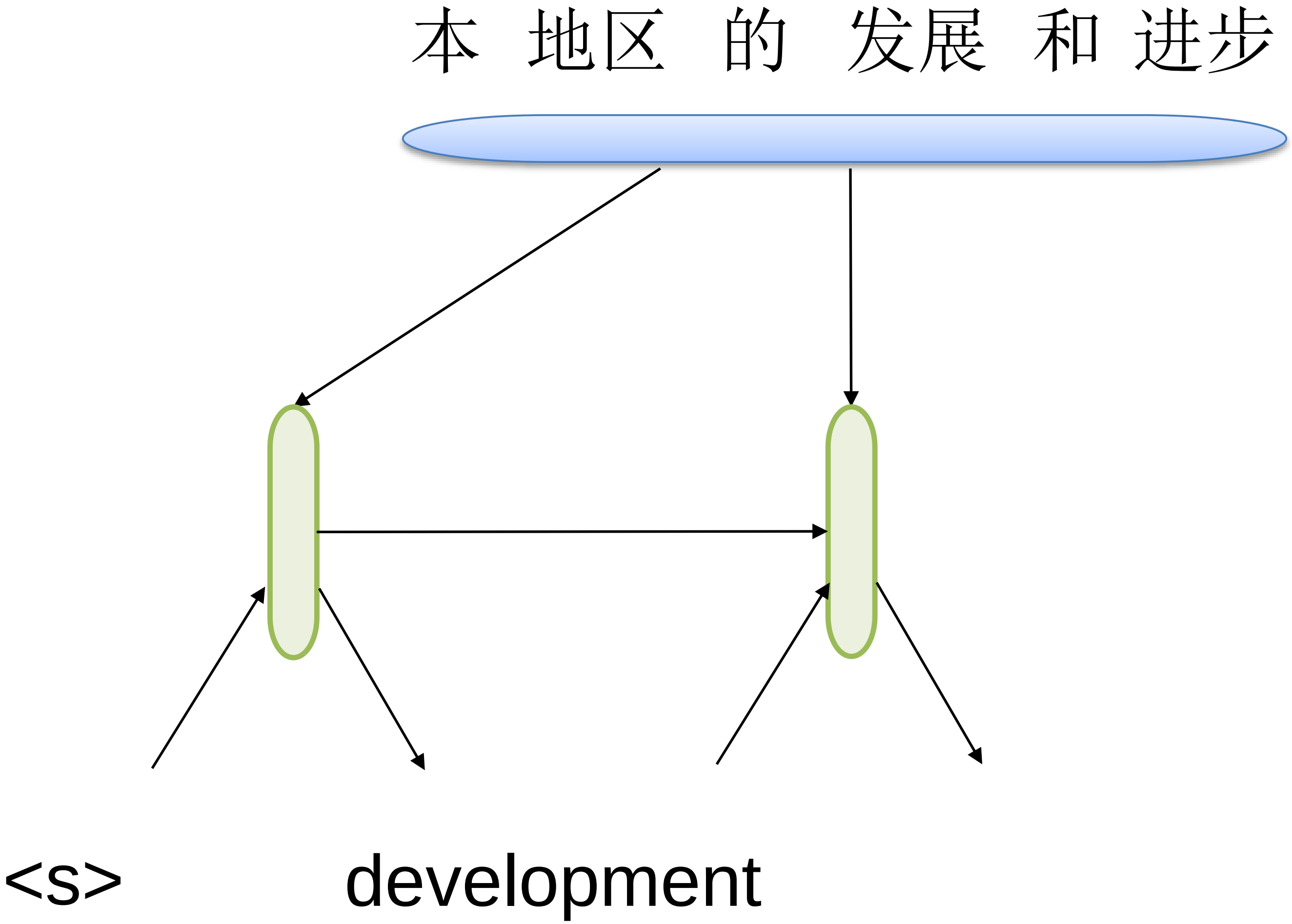
本 地区 的 发展 和 进步

<s>

development

Entire source sentence representation

Recurrent Network Joint Model



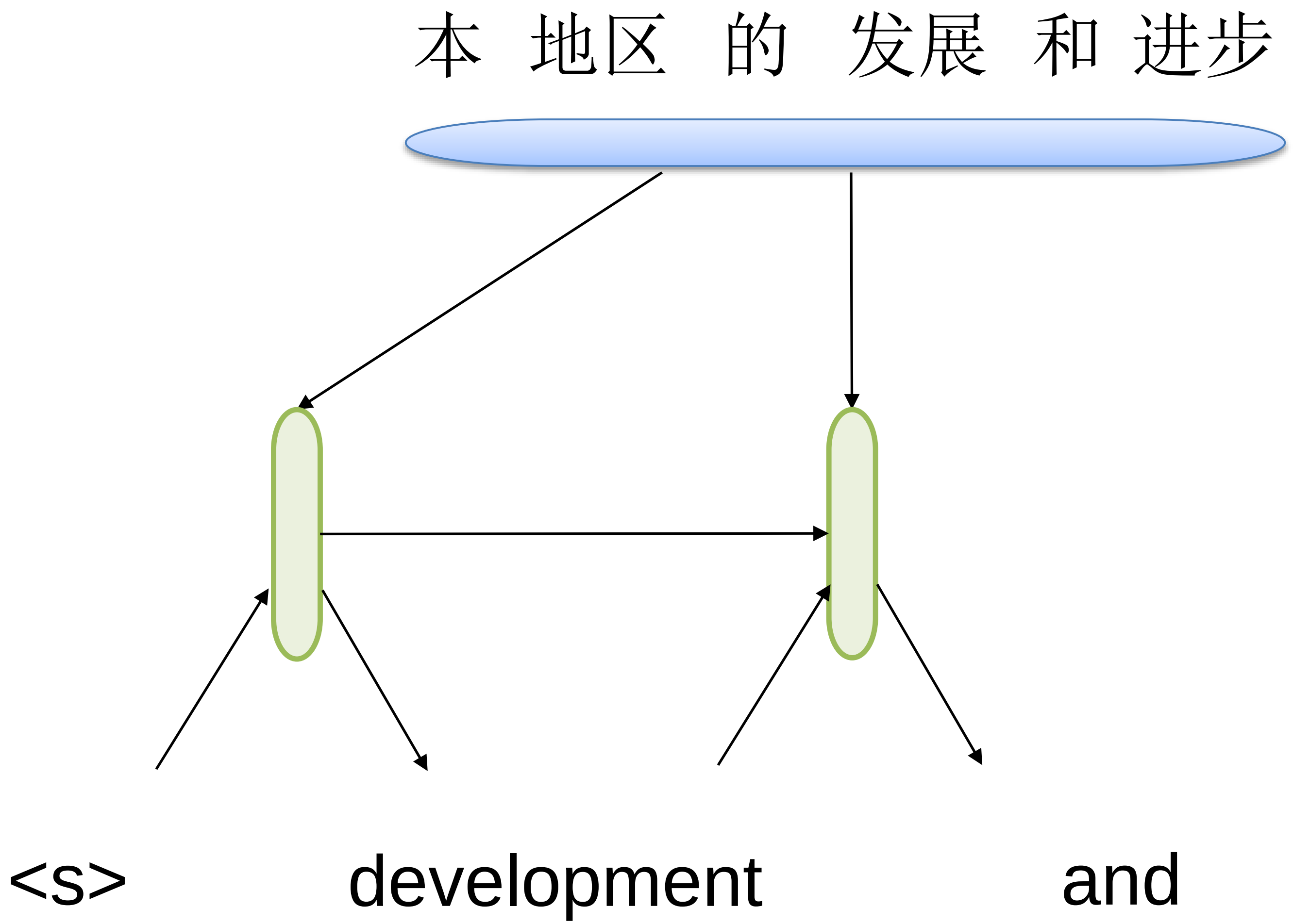
本 地区 的 发展 和 进步

<s>

development

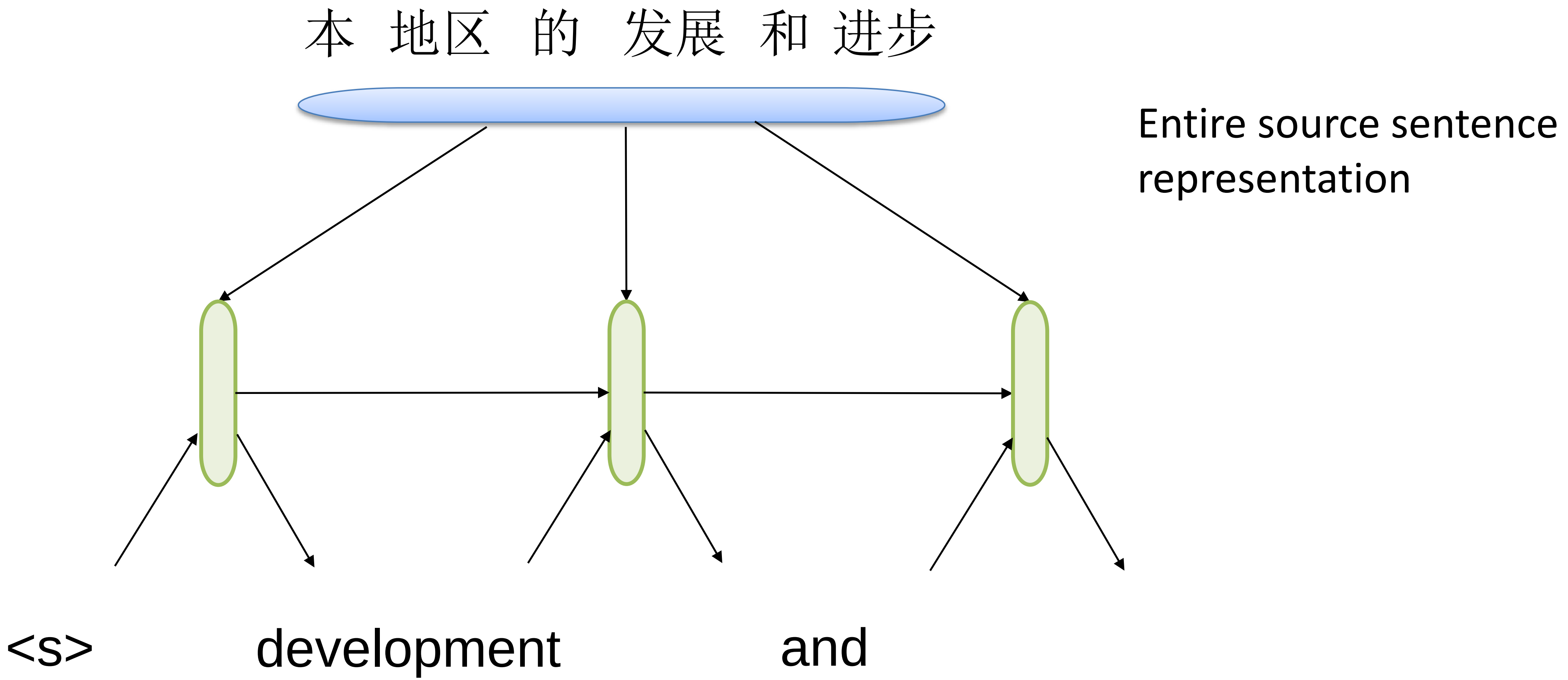
Entire source sentence representation

Recurrent Network Joint Model



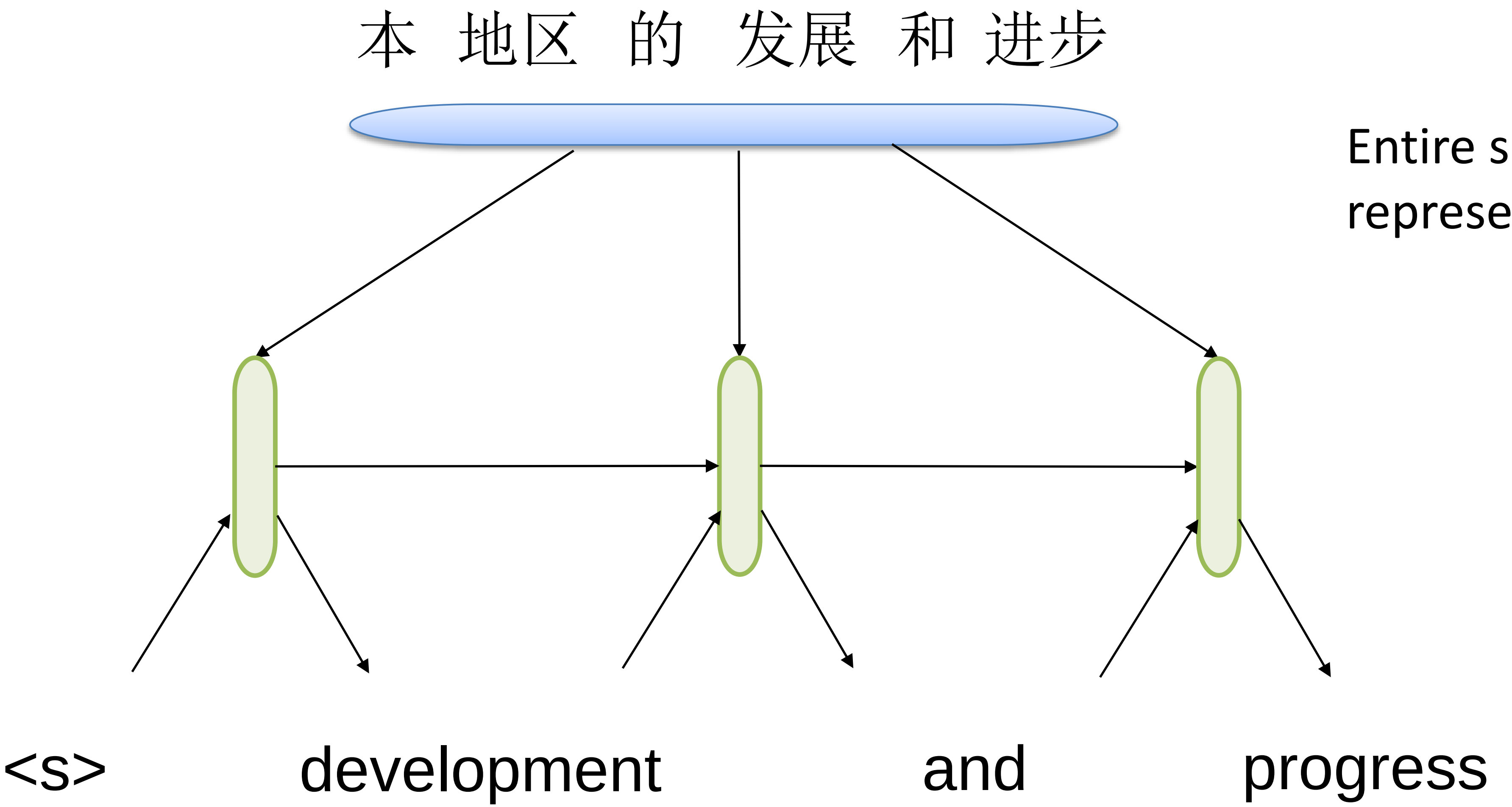
Entire source sentence representation

Recurrent Network Joint Model



Entire source sentence representation

Recurrent Network Joint Model



本 地区 的 发展 和 进步

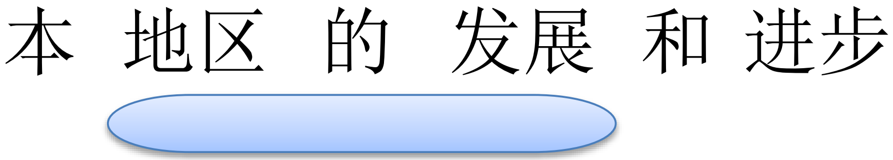
Entire source sentence representation

<s> development and progress

Recurrent Network Joint Model

本 地区 的 发展 和 进步

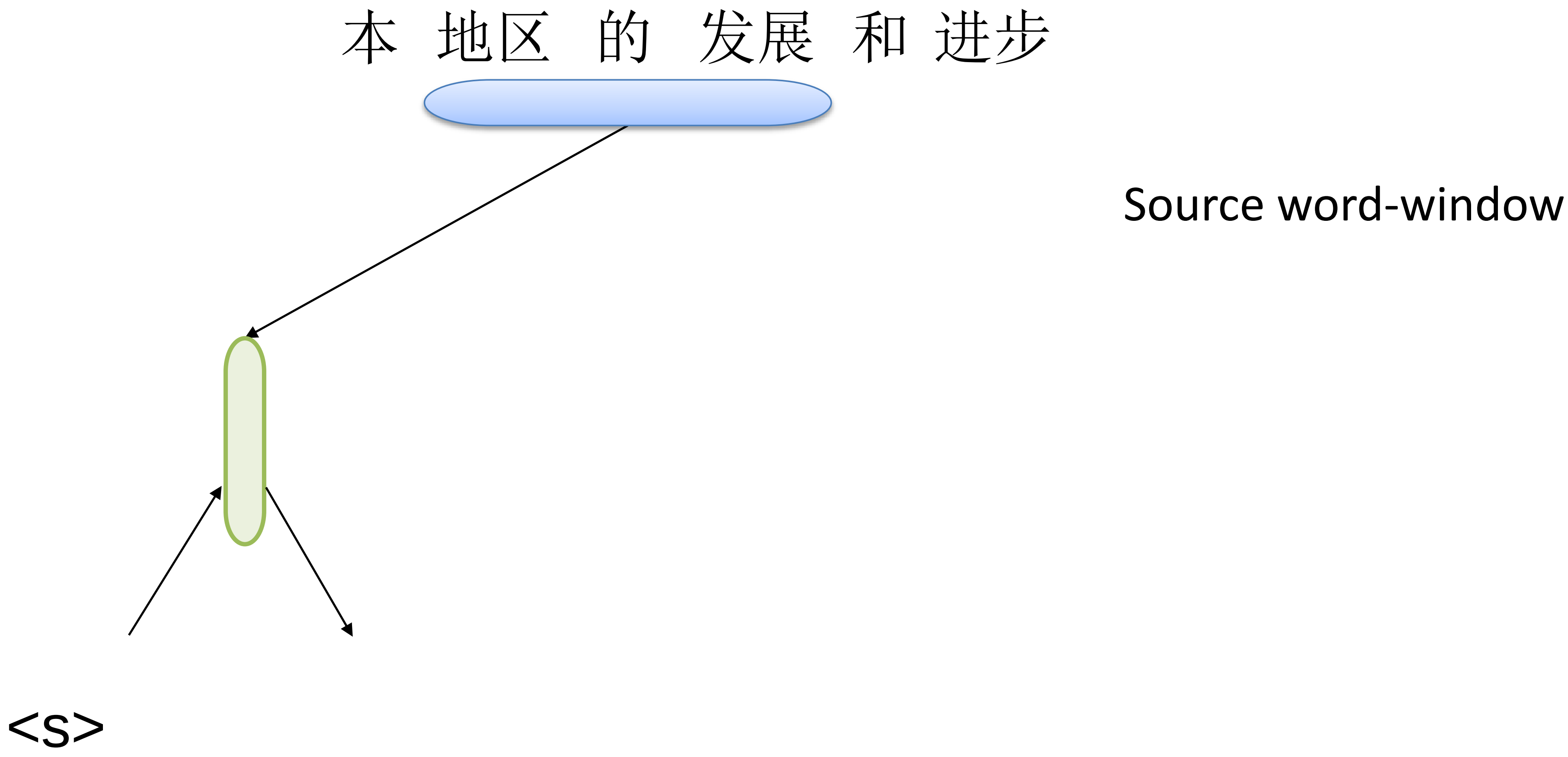
Recurrent Network Joint Model



本 地区 的 发展 和 进步

Source word-window

Recurrent Network Joint Model

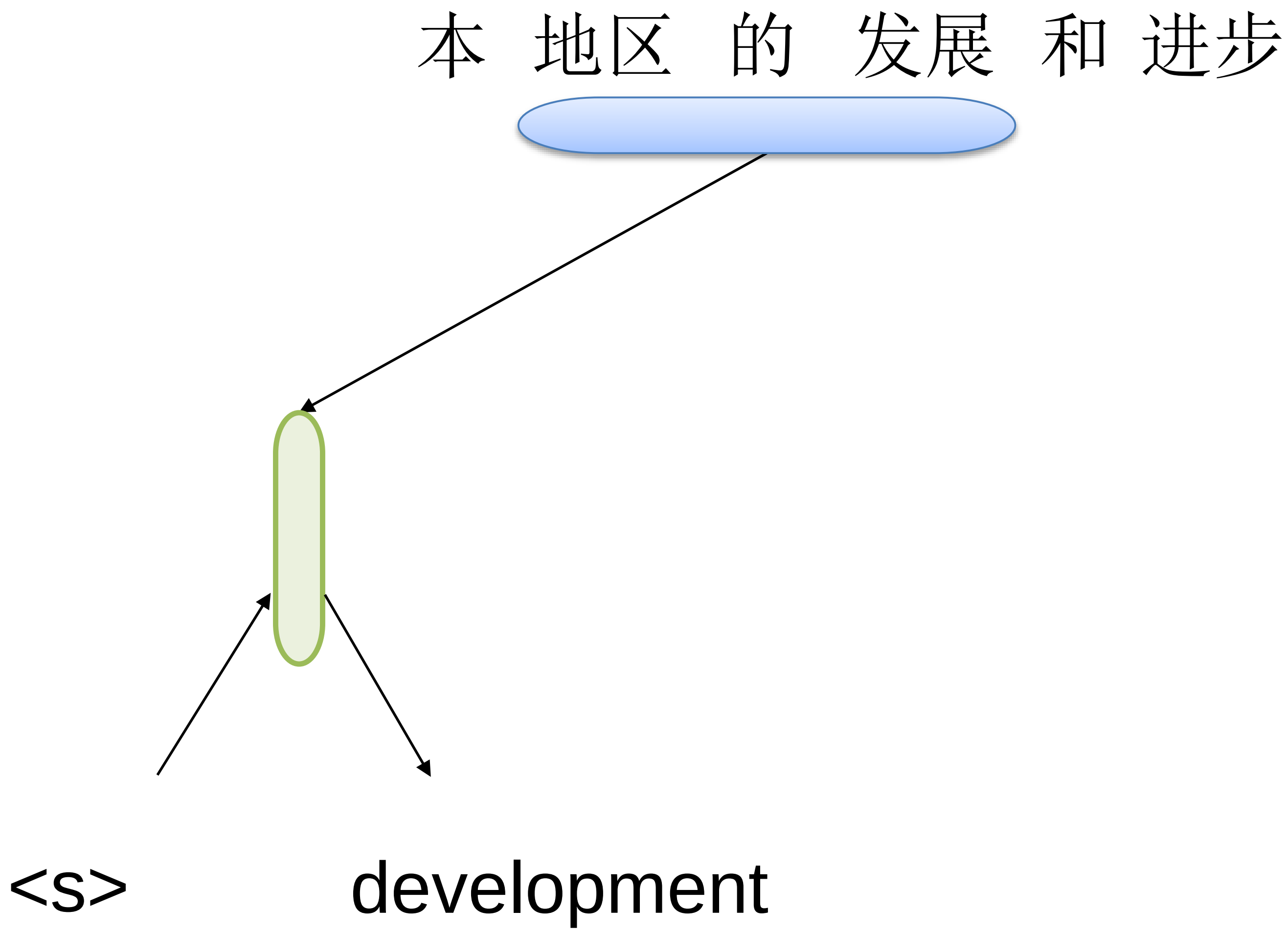


本 地区 的 发展 和 进步

Source word-window

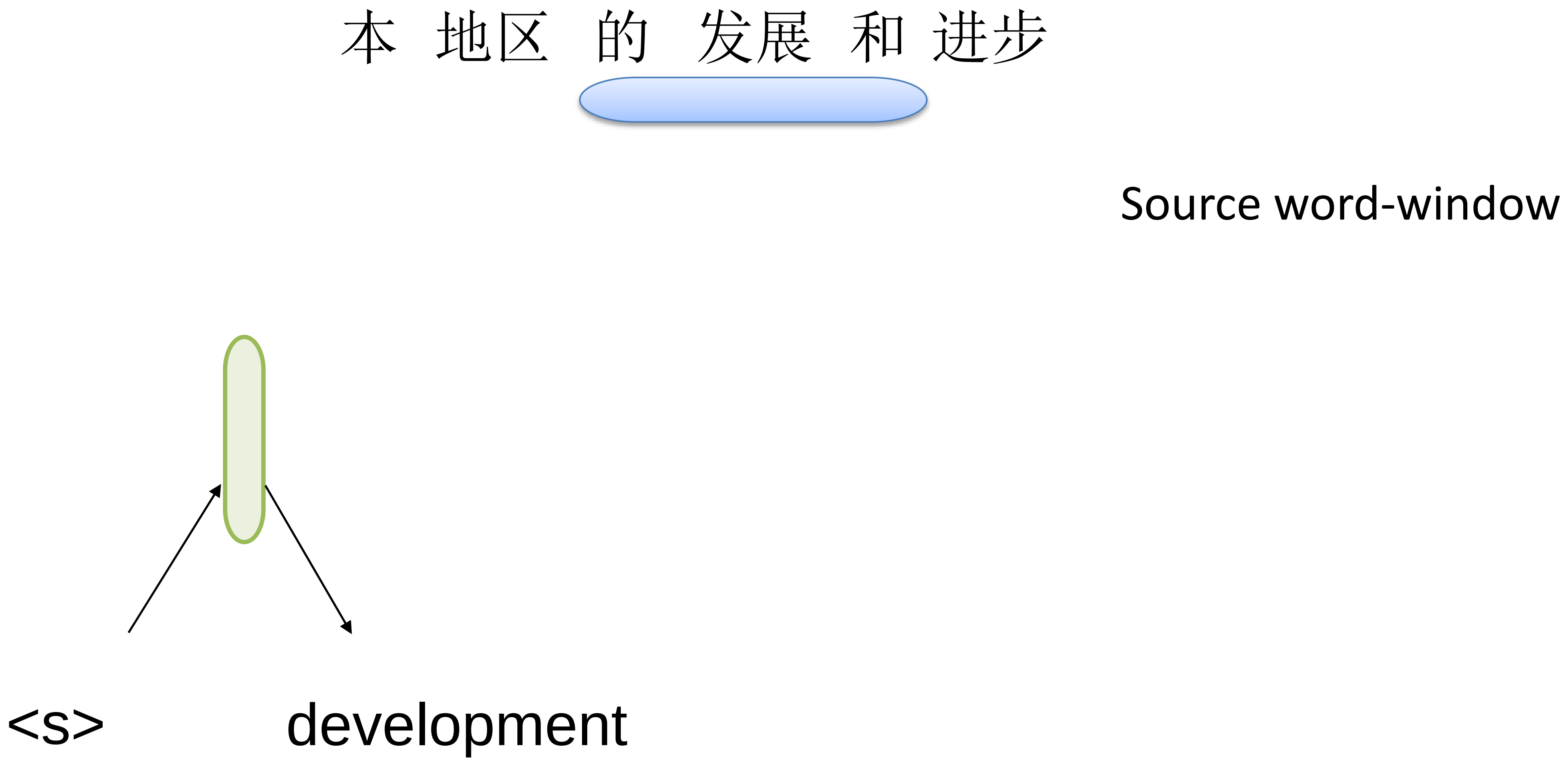
<s>

Recurrent Network Joint Model



Source word-window

Recurrent Network Joint Model



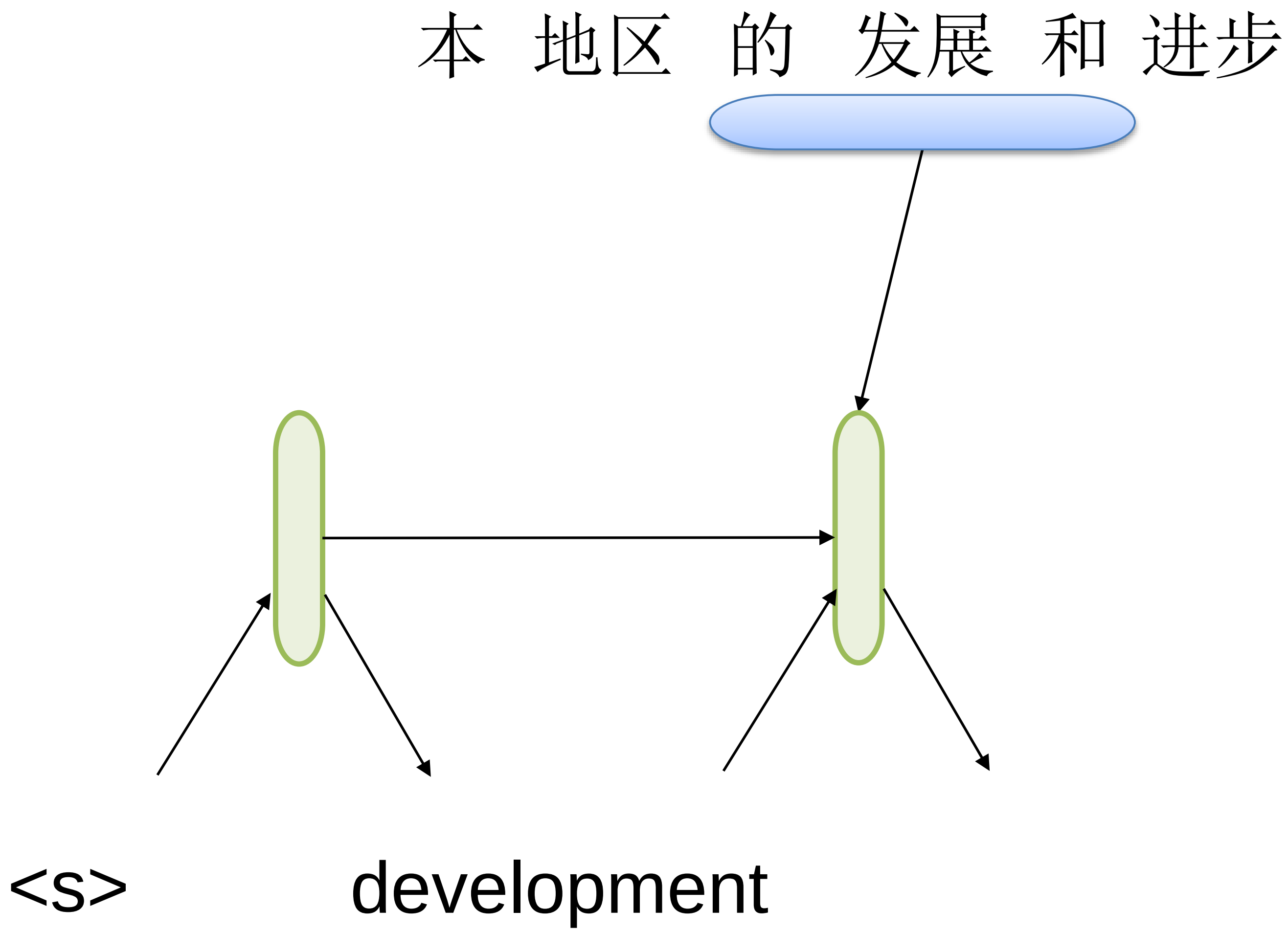
本 地区 的 发展 和 进步

Source word-window

<s>

development

Recurrent Network Joint Model



本 地区 的 发展 和 进步

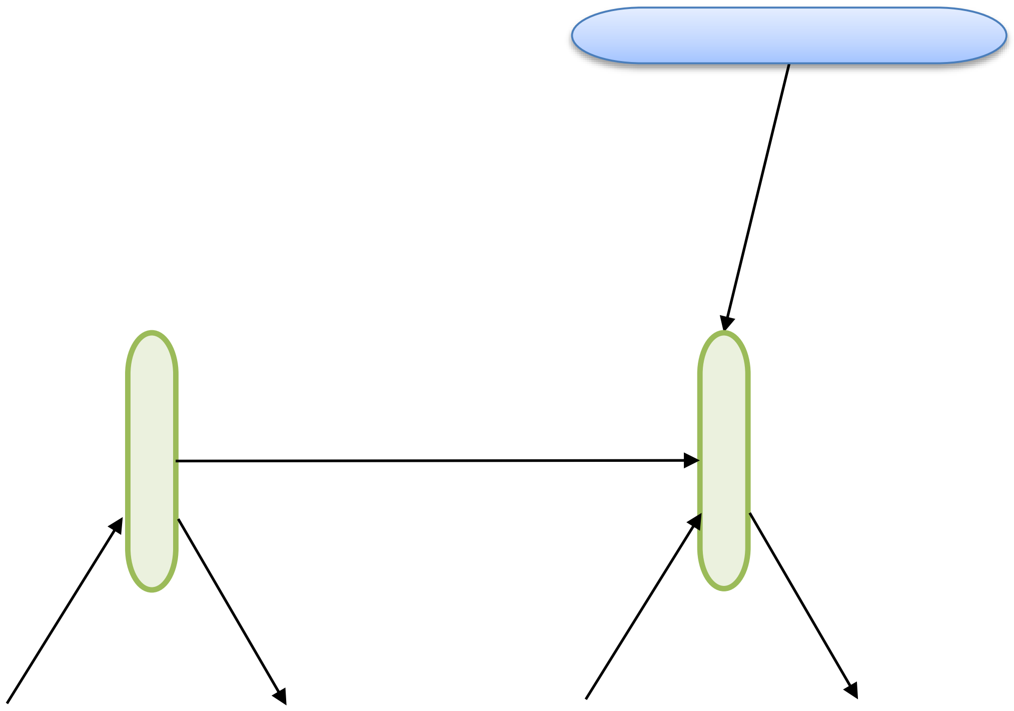
Source word-window

<s>

development

Recurrent Network Joint Model

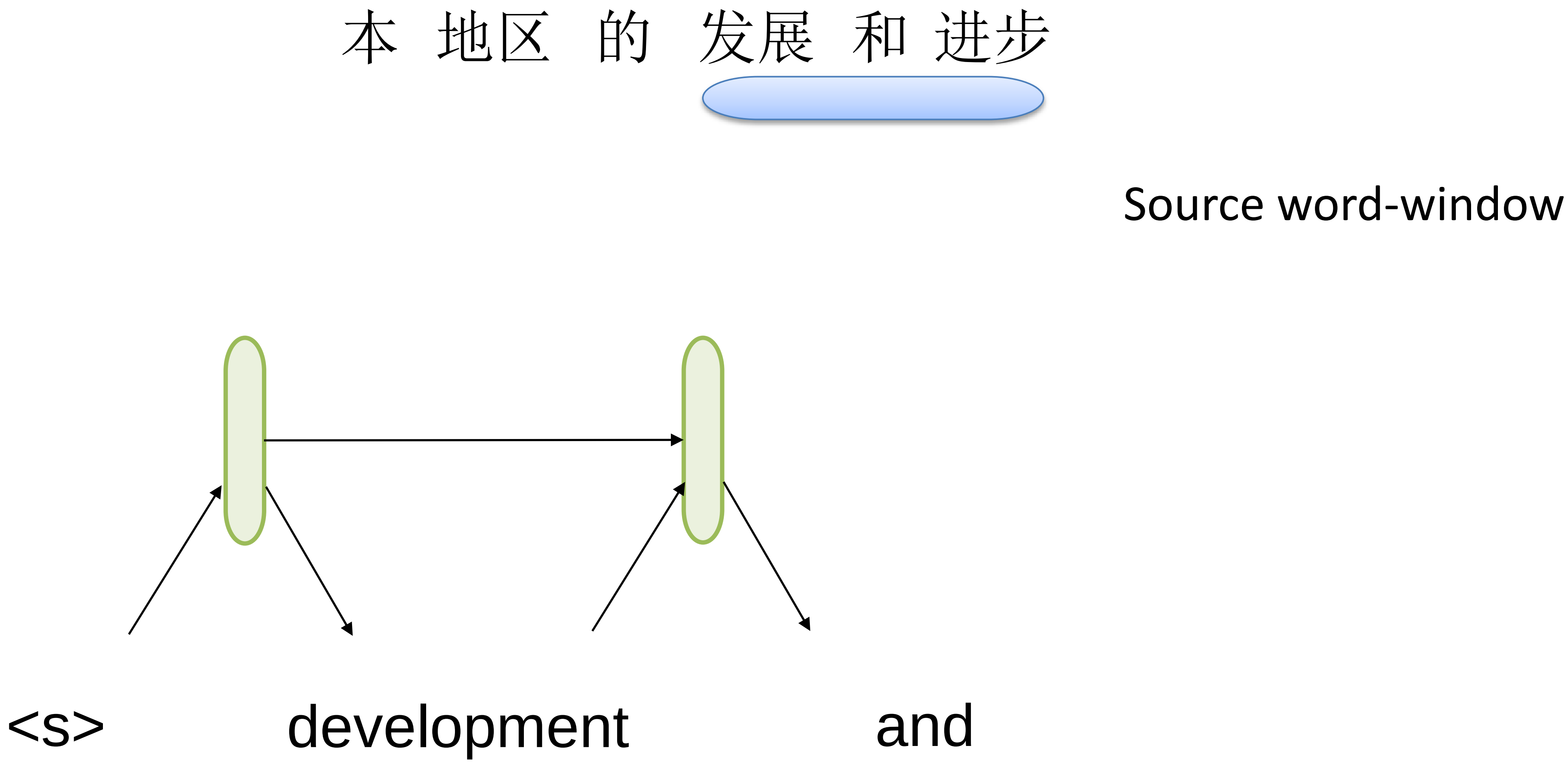
本 地区 的 发展 和 进步



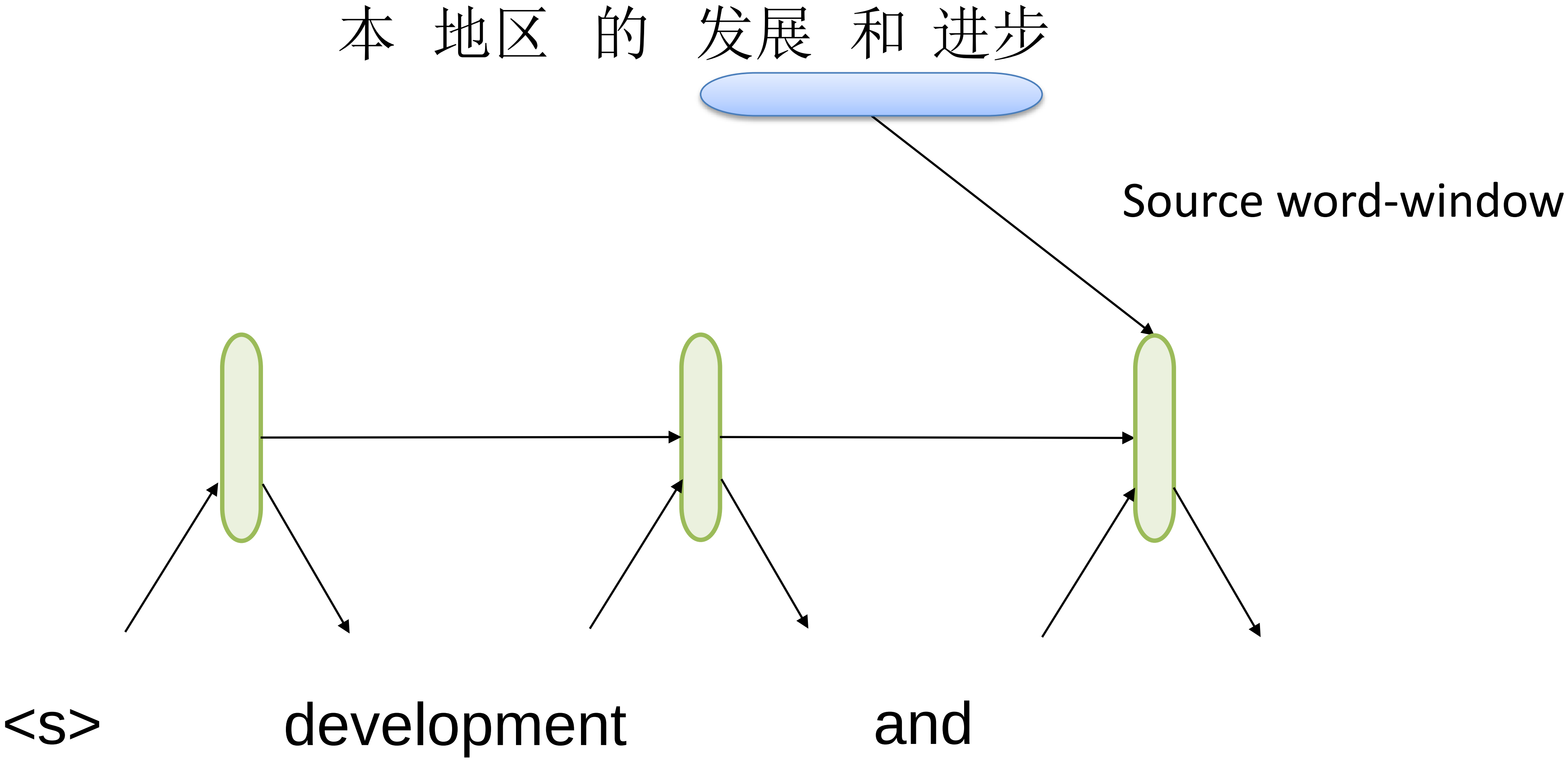
Source word-window

<s> development and

Recurrent Network Joint Model



Recurrent Network Joint Model

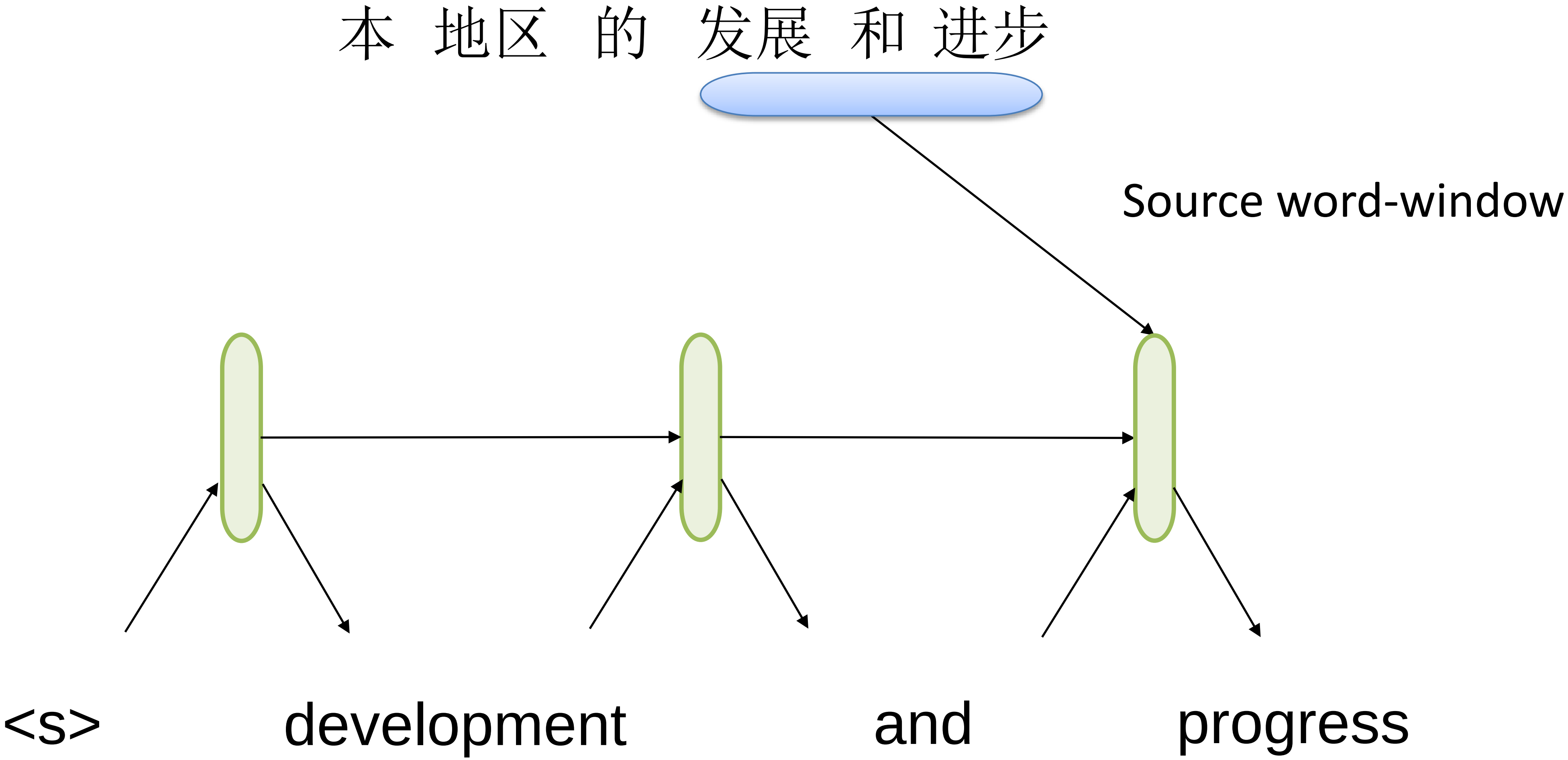


本 地区 的 发展 和 进步

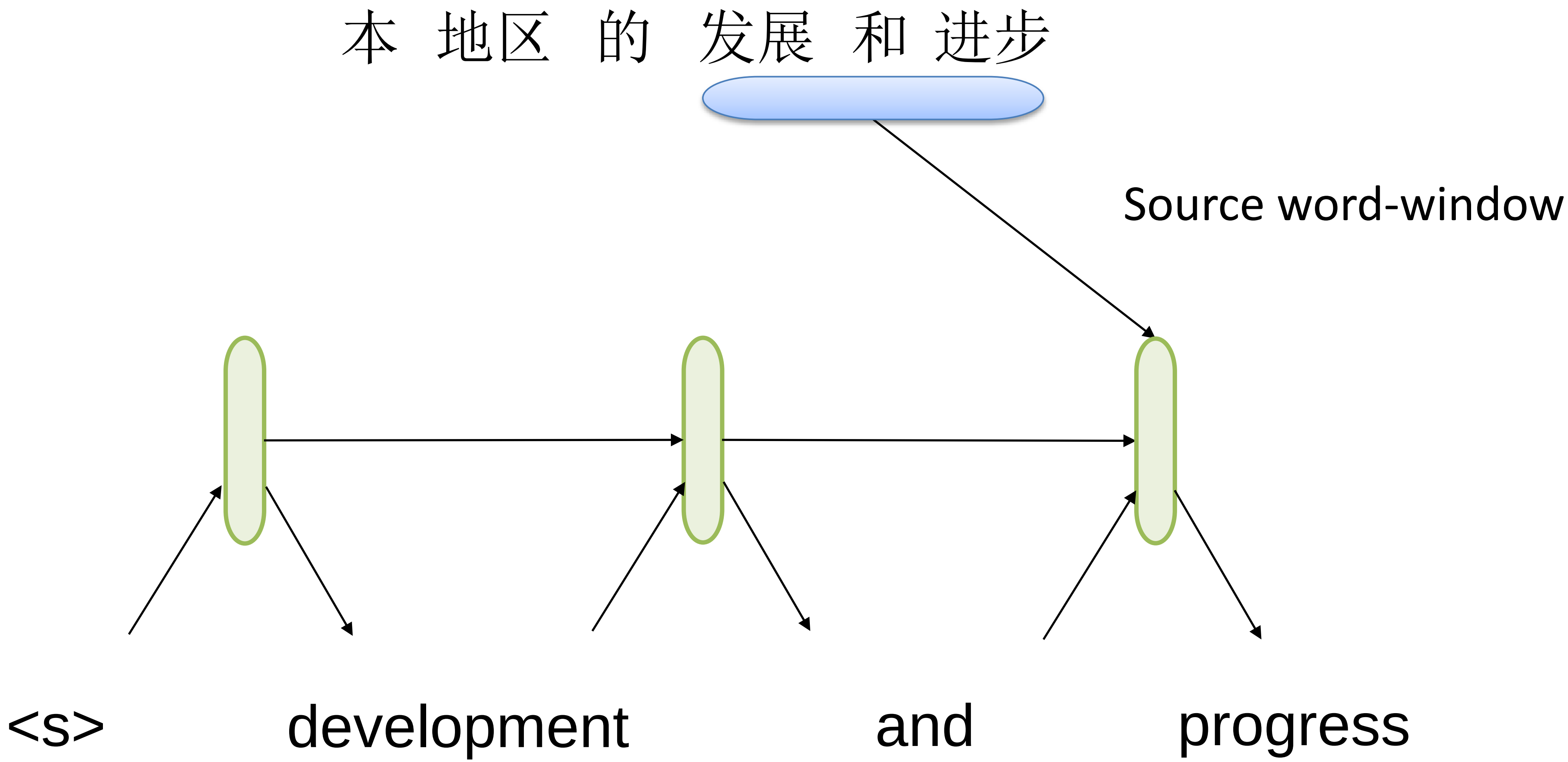
Source word-window

<s> development and

Recurrent Network Joint Model



Recurrent Network Joint Model



Feed-forward nets: Le (2012) & Devlin (2014)

Similar to Kalchbrenner (2013)

Does the joint model learn how to translate?

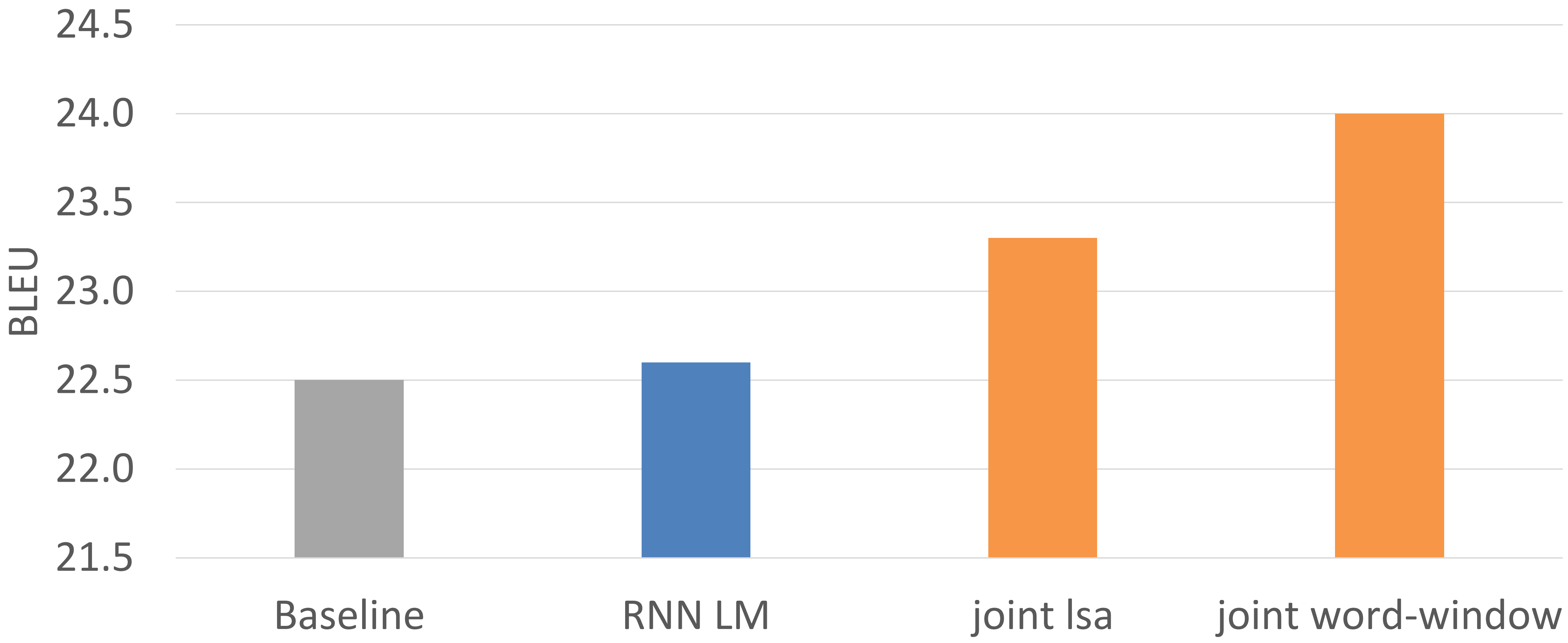
Experiment: Generate baseline n-best,
remove translation model,
rescore with RNN joint model

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

Does the joint model learn how to translate?

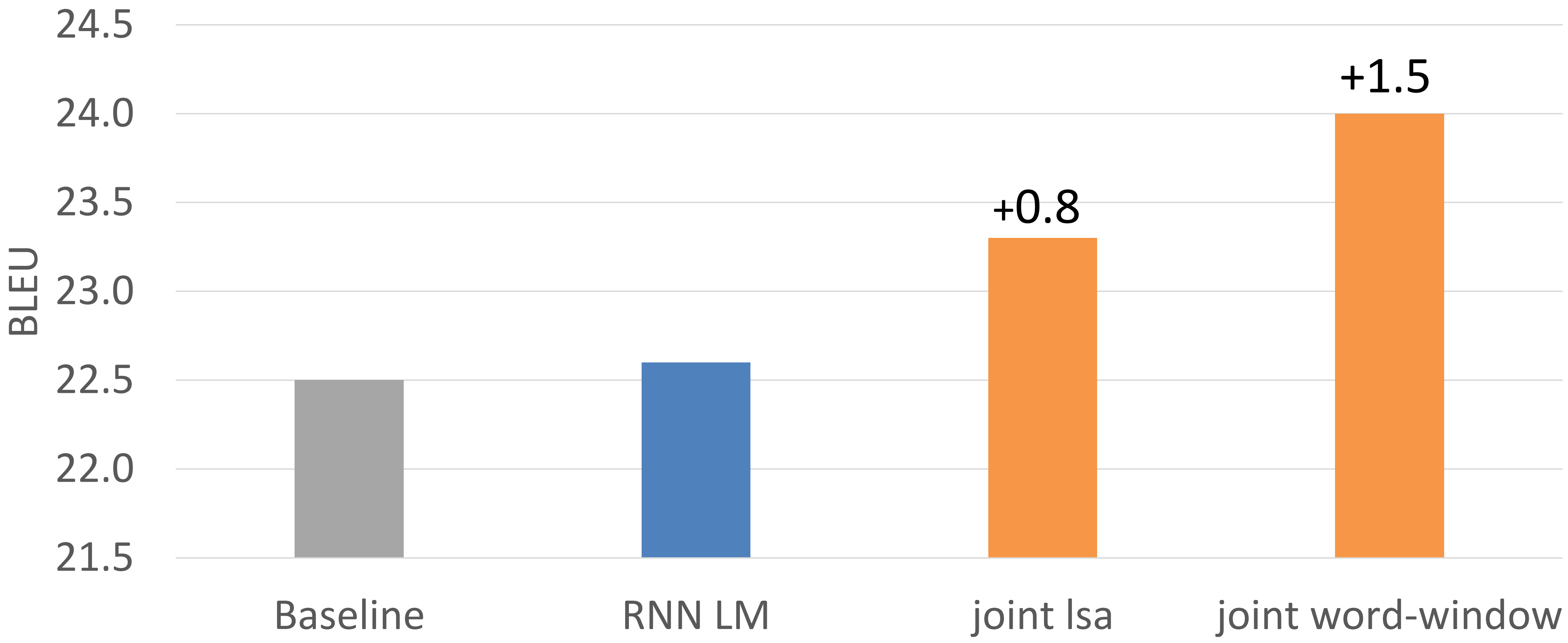
Average accuracy over news2010, syscomb2010, news2011

Does the joint model learn how to translate?



Average accuracy over news2010, syscomb2010, news2011

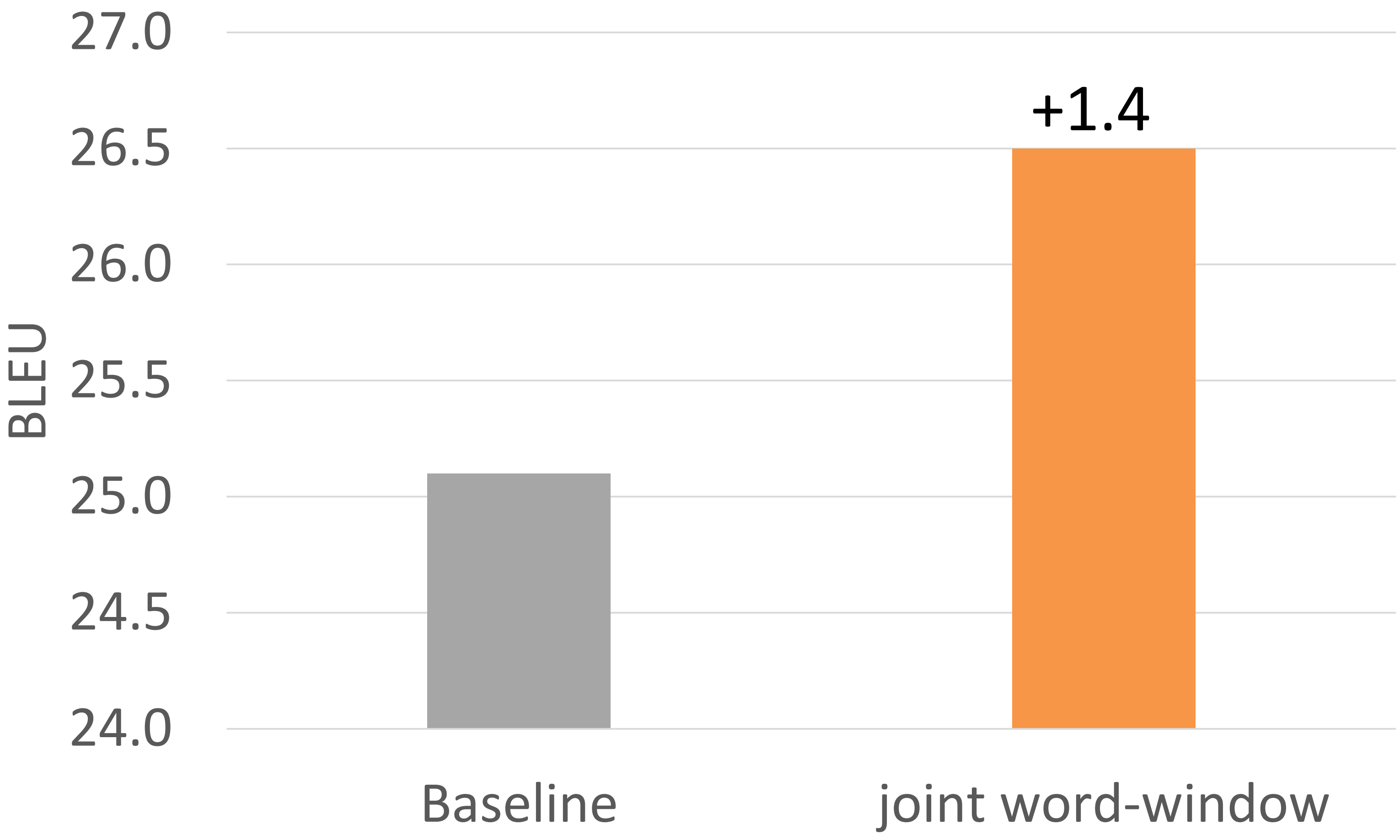
Does the joint model learn how to translate?



Average accuracy over news2010, syscomb2010, news2011

Improving a phrase-based baseline

Improving a phrase-based baseline



Qualitative Results

src: il aurait fallu 226 voix pour l’ approuver.

ref: its ratification would require 226 votes.

base: it should have been 226 votes to approve it.

rnn: it would have been 226 votes to approve.

Qualitative Results

src: il aurait fallu 226 voix pour l’ approuver.

ref: its ratification would require 226 votes.

base: it should have been 226 votes to approve it.

rnn: it would have been 226 votes to approve.

src: il reste à déterminer les vainqueurs.

ref: it is time to define the winners.

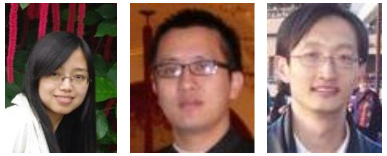
base: it remains to be seen the victors.

rnn: it remains to determine the victors.

Overview

- Recurrent neural network joint models (Auli et al., EMNLP 2013)
  Combined language and translation modeling

- Minimum translation modeling with recurrent nets (Hu et al., EACL 2014)
  Sequence models over bilingual units

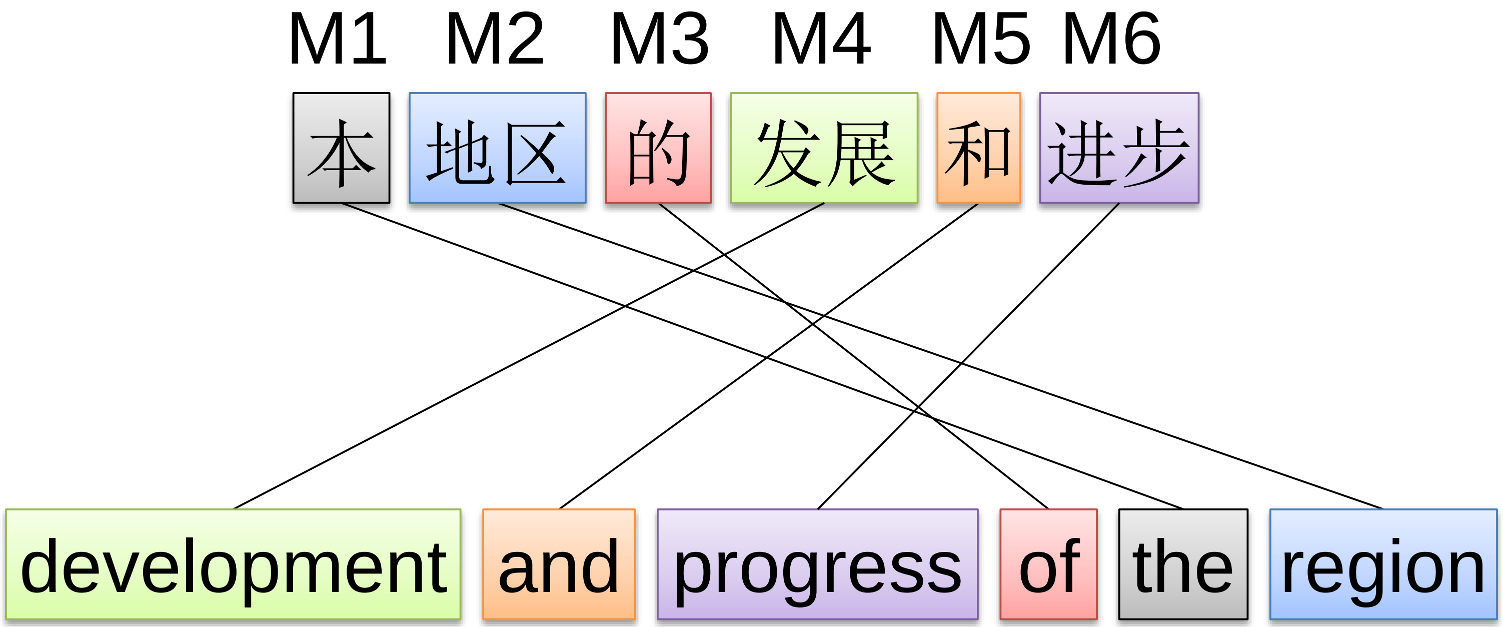


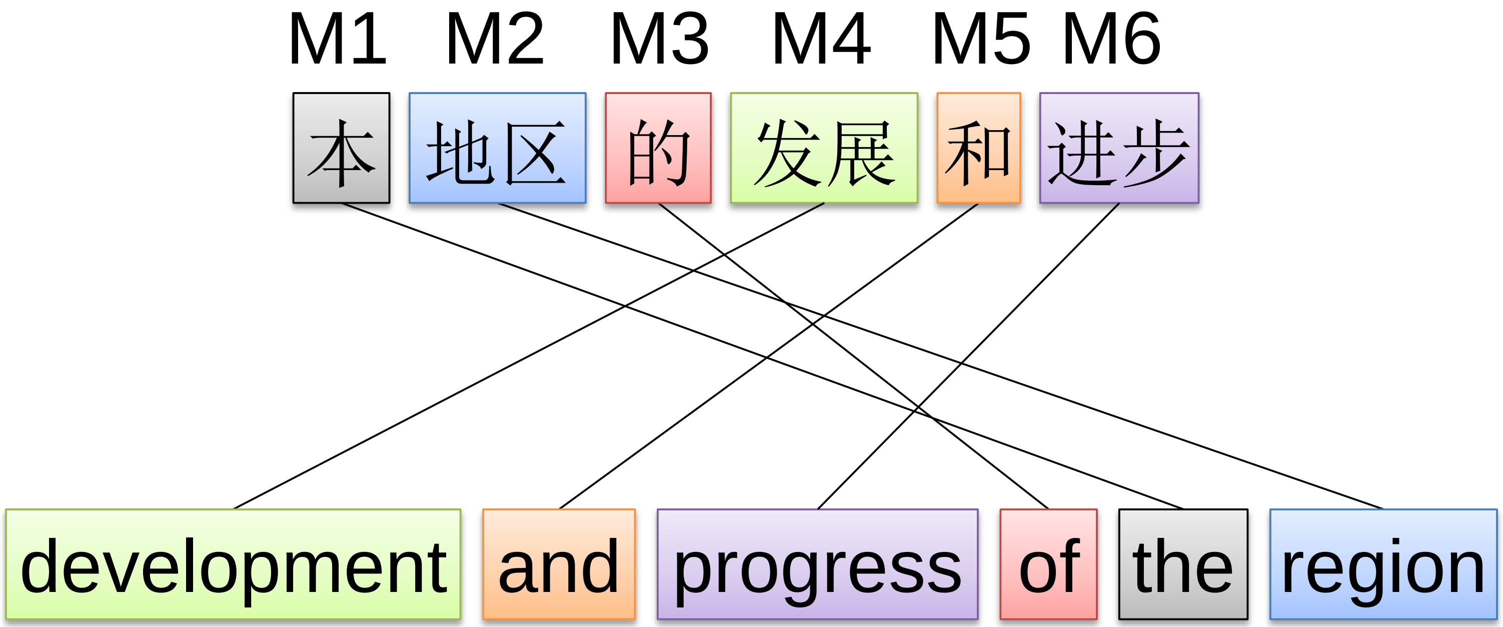
- Training recurrent nets (Auli & Gao, ACL 2014)
  Expected BLEU training for neural network translation models

- Large-scale discriminative sparse ordering models (Auli et al., in submission)
  Training millions of linear ordering features with expected BLEU

本 地区 的 发展 和 进步

development and progress of the region

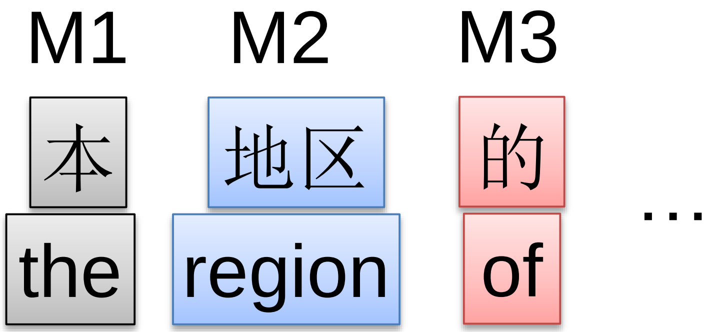

M1
M2
M3
M4
M5
M6
本
地区
的
发展
和
进步
development
and
progress
of
the
region



M1 M2 M3 M4 M5 M6

本 地区 的 发展 和 进步

development and progress of the region



M1 M2 M3

本 地区 的

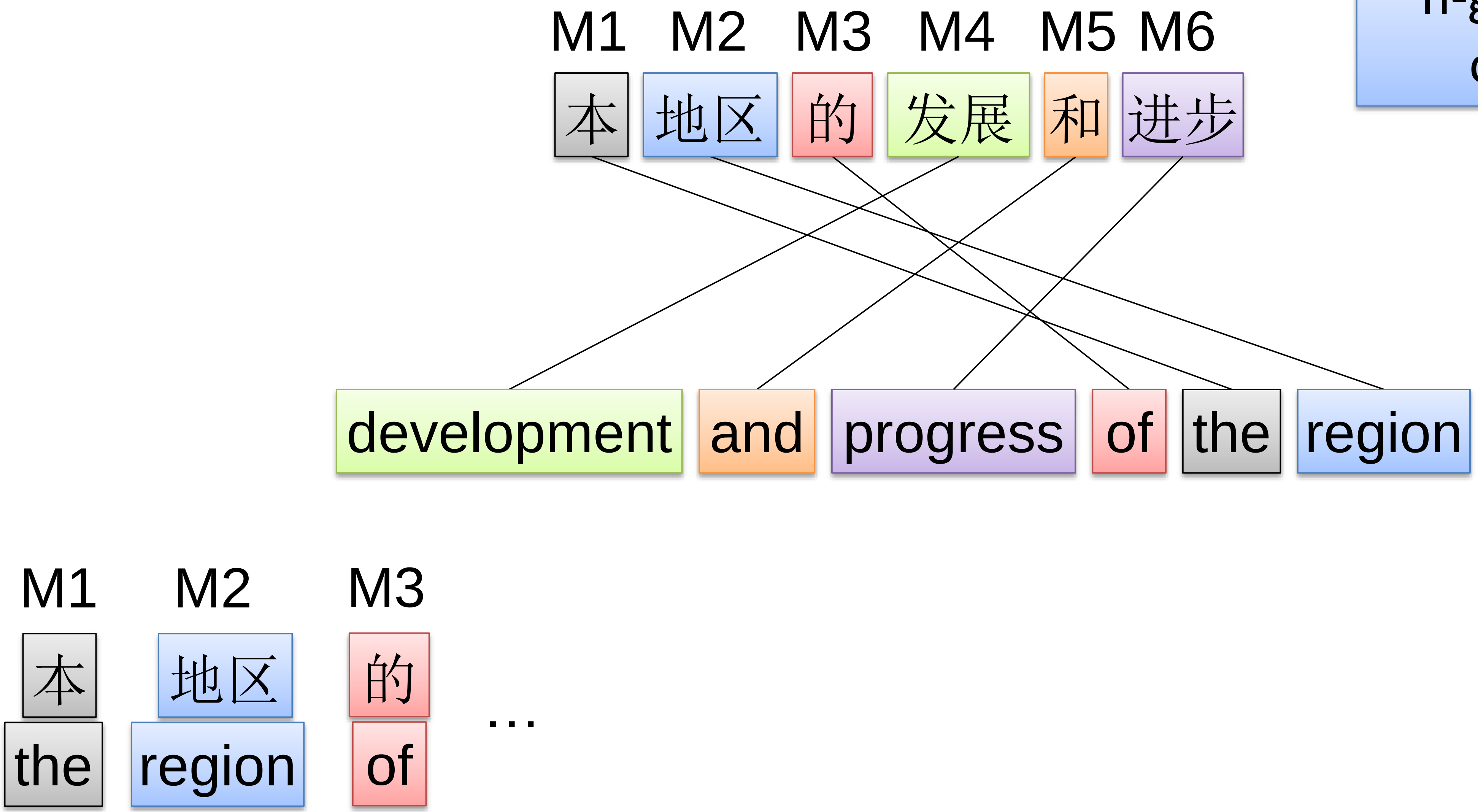
the region of

…

Banchs et al. (2005)

Quirk & Menezes (2006)

n-gram models over MTUs



M1 M2 M3 M4 M5 M6

本 地区 的 发展 和 进步

development and progress of the region

M1 M2 M3

本 地区 的

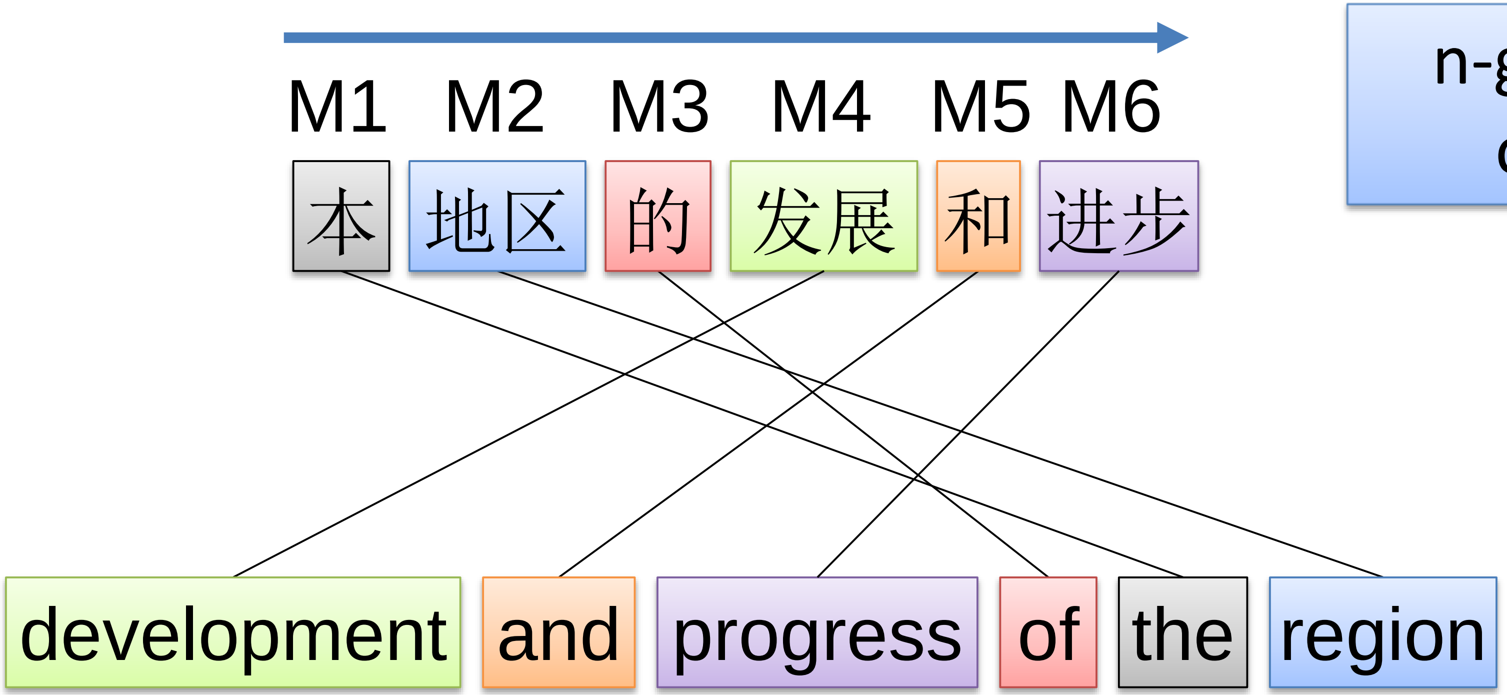
the region of

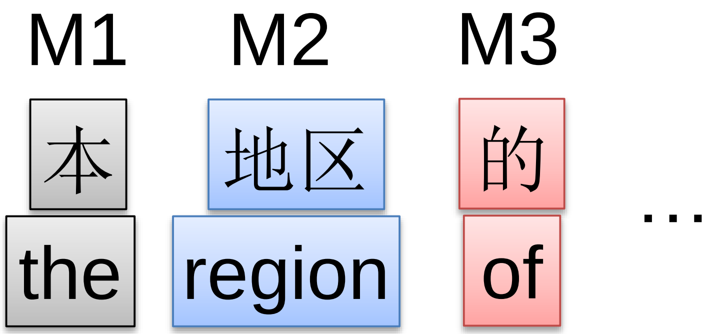
…

Banchs et al. (2005)

Quirk & Menezes (2006)

n-gram models
over MTUs



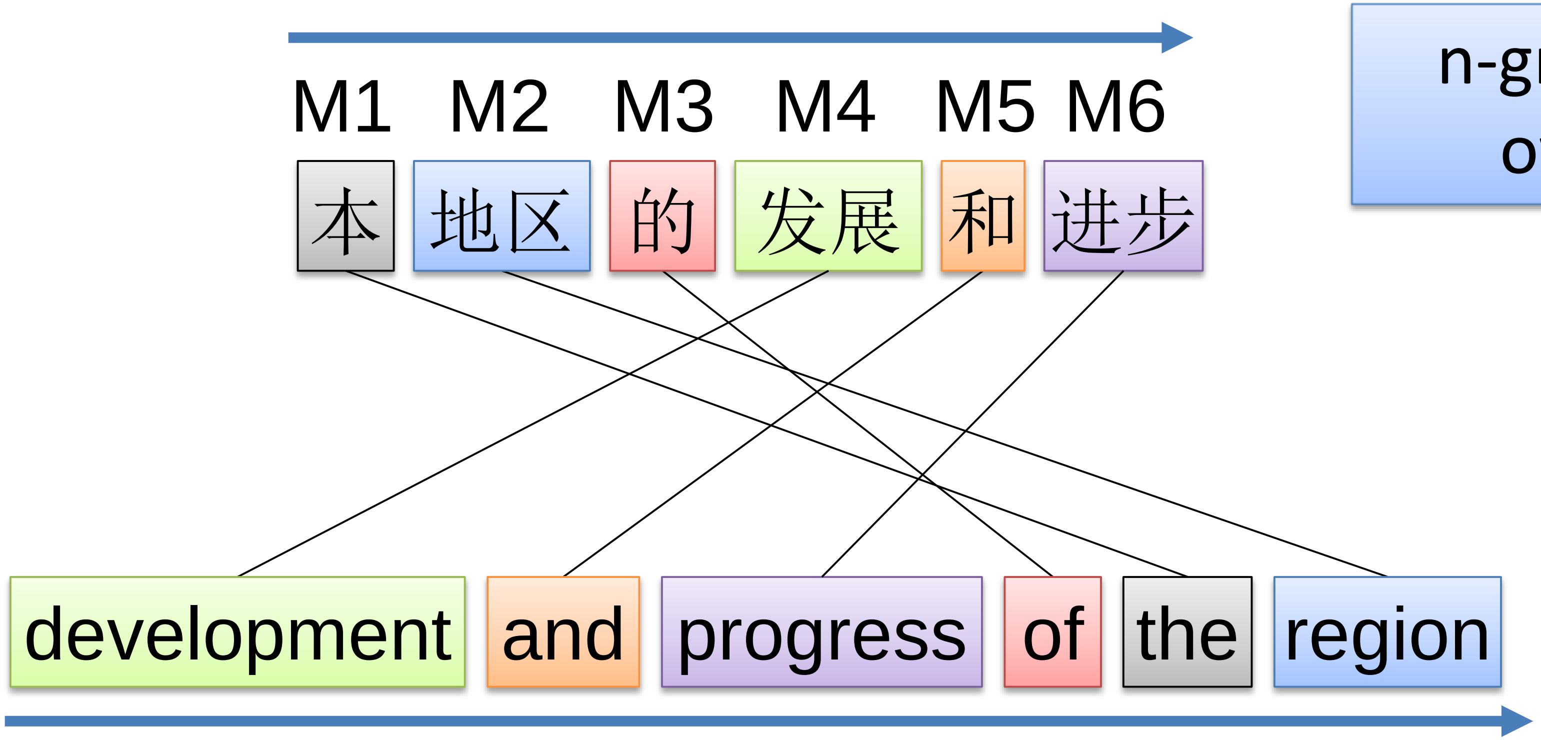


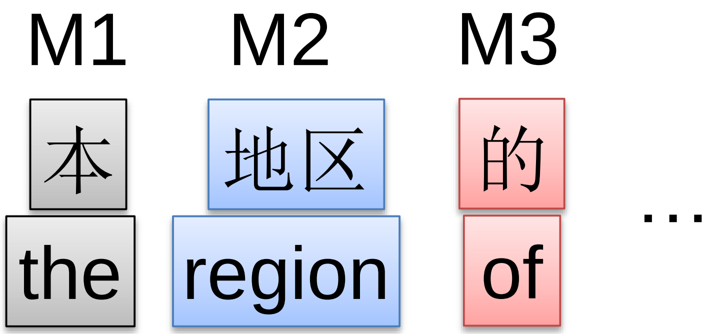
Source order: p(M1) p(M2|M1) p(M3|M1,M2) …

Banchs et al. (2005)

Quirk & Menezes (2006)

n-gram models over MTUs





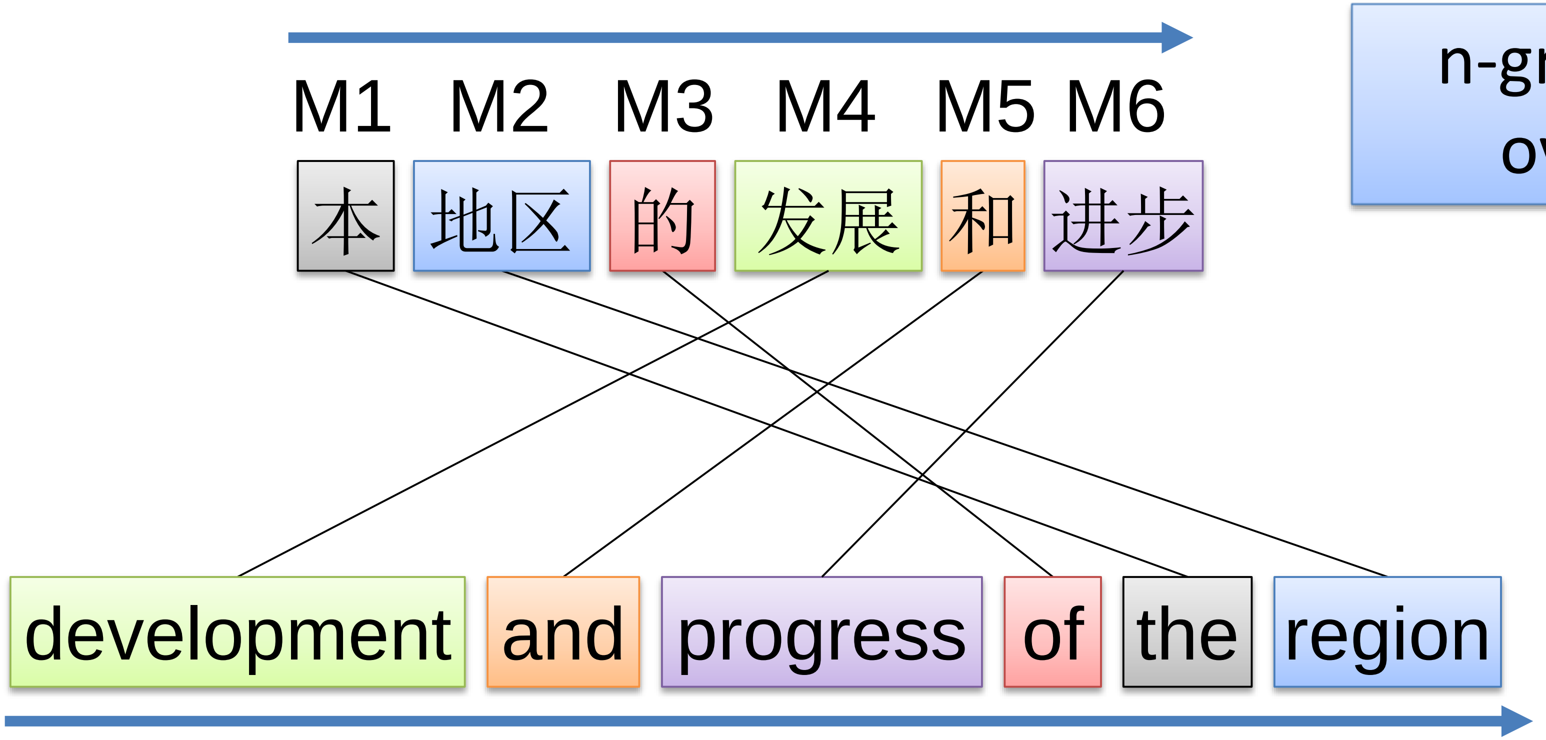
Source order: $p(M1)\ p(M2|M1)\ p(M3|M1,M2) \ldots$

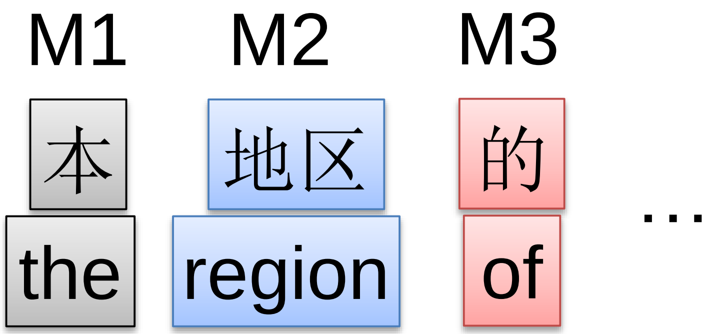
Target order: $p(M4)\ p(M5|M4)\ p(M6|M4,M5) \ldots$

Banchs et al. (2005)

Quirk & Menezes (2006)

n-gram models over MTUs

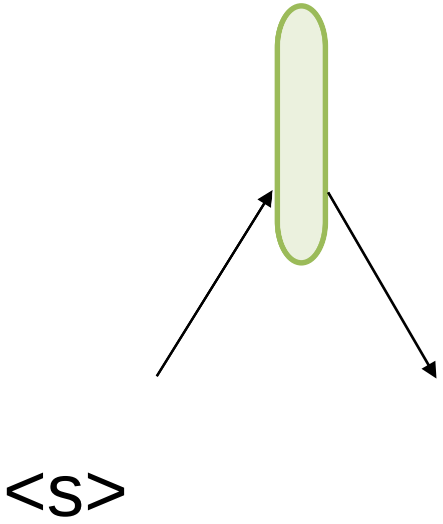




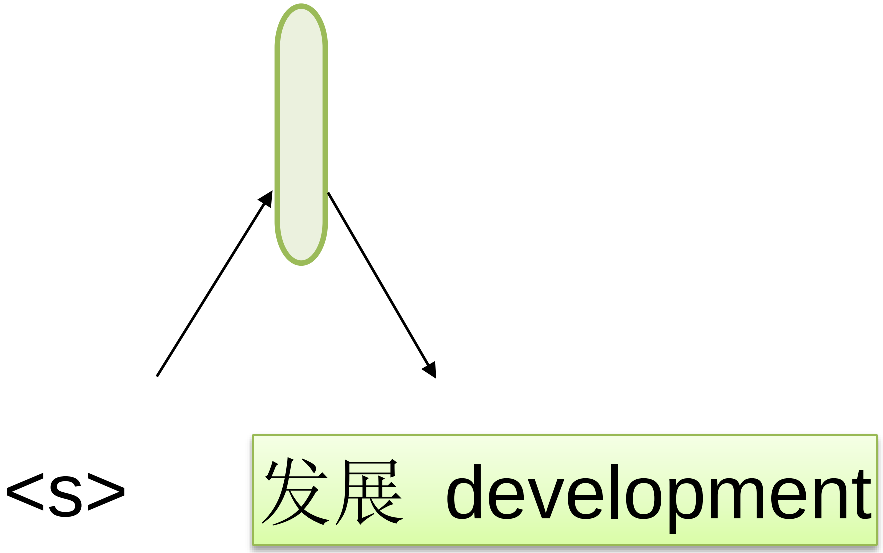
Source order: p(M1) p(M2|M1) p(M3|M1,M2) …

Target order: p(M4) p(M5|M4) p(M6|M4,M5) …

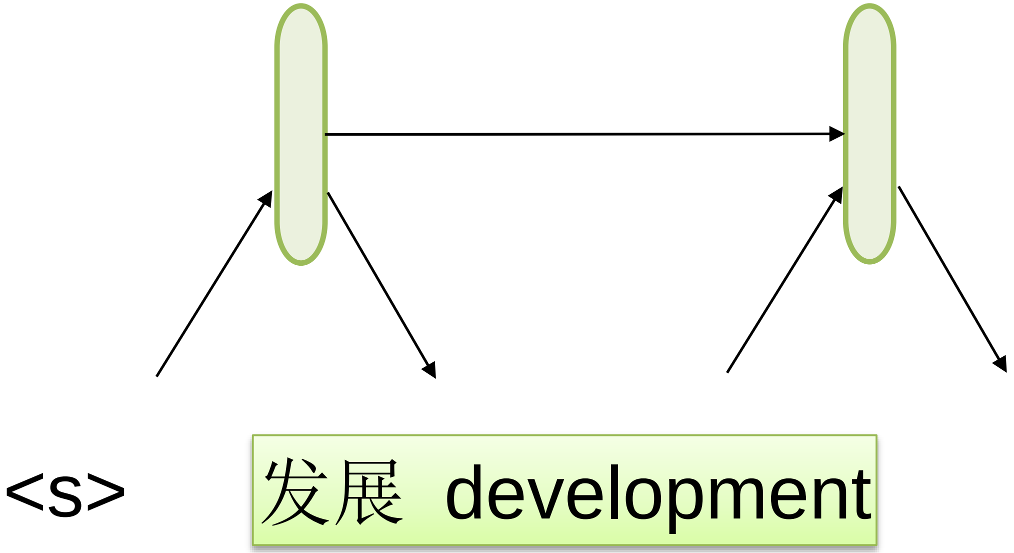
Model 1: Recurrent Atomic MTU



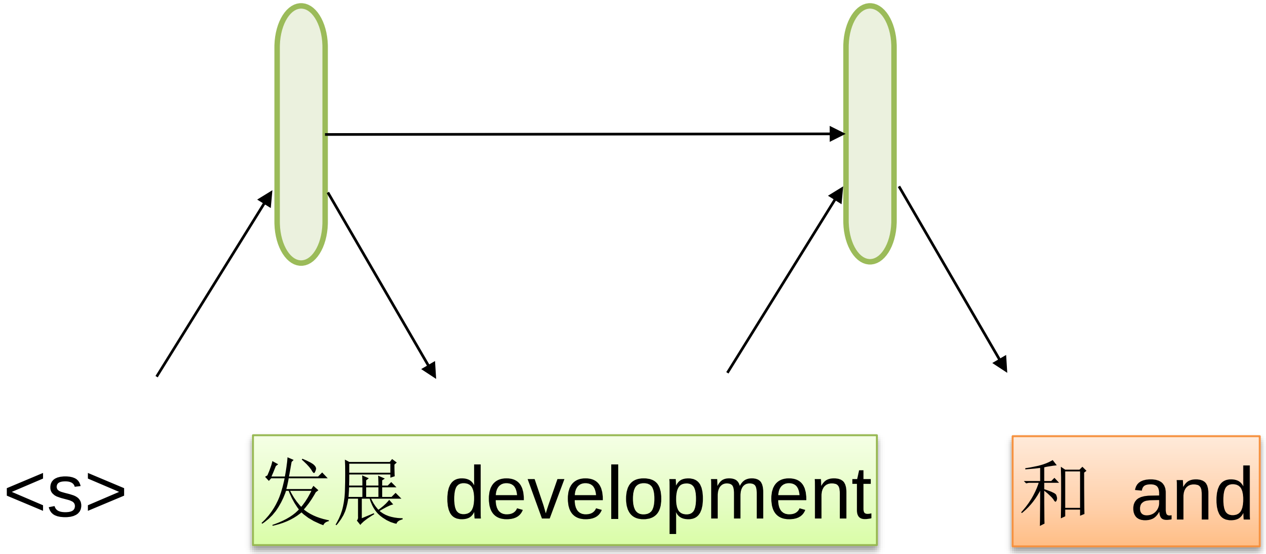
Model 1: Recurrent Atomic MTU



Model 1: Recurrent Atomic MTU



Model 1: Recurrent Atomic MTU

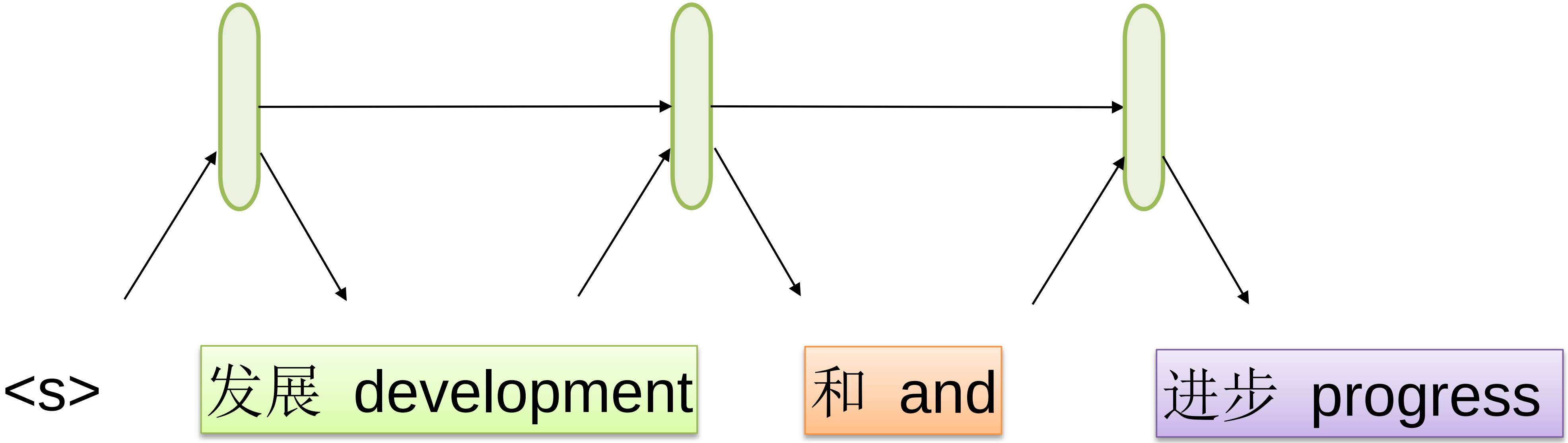


<s>

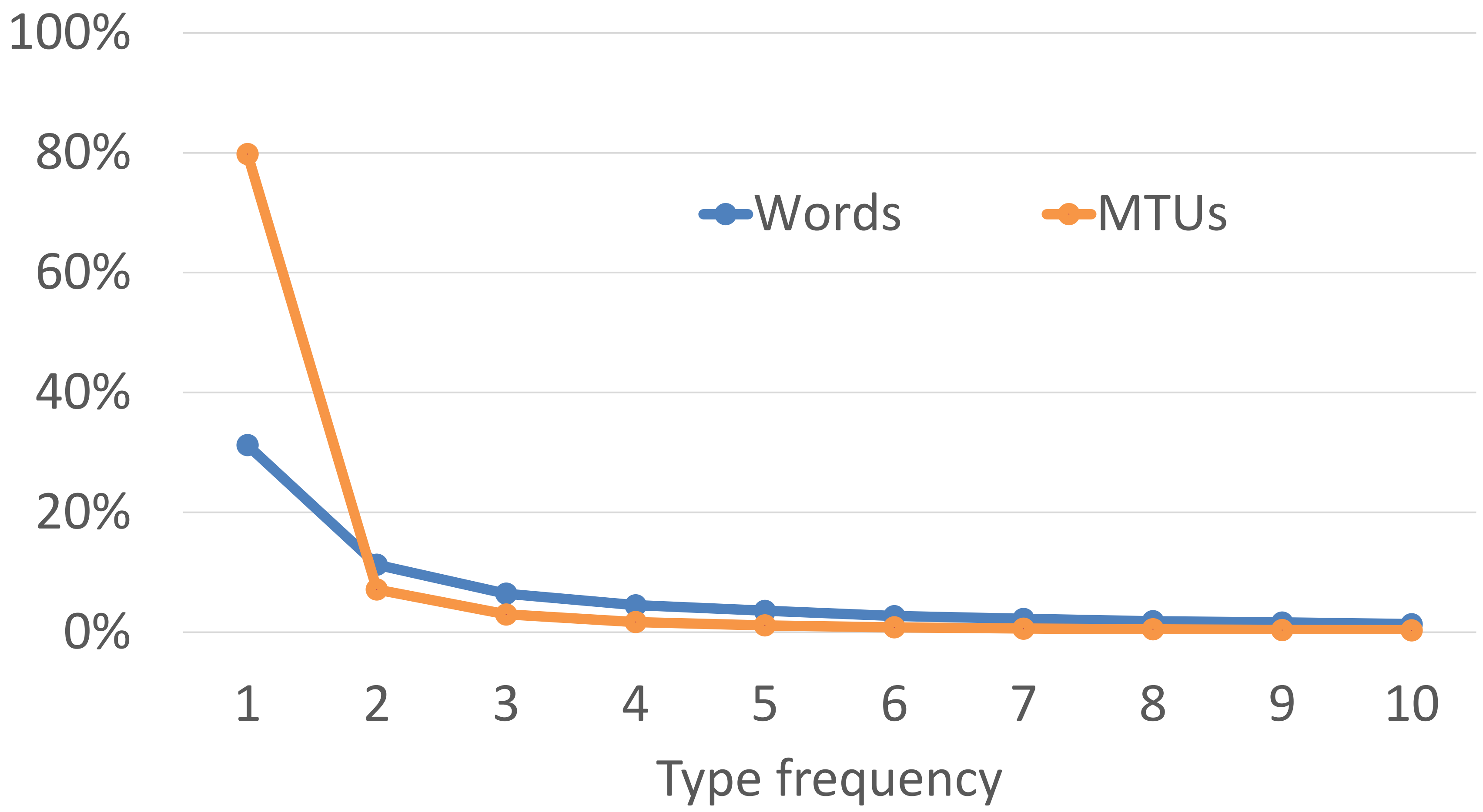
发展 development

和 and

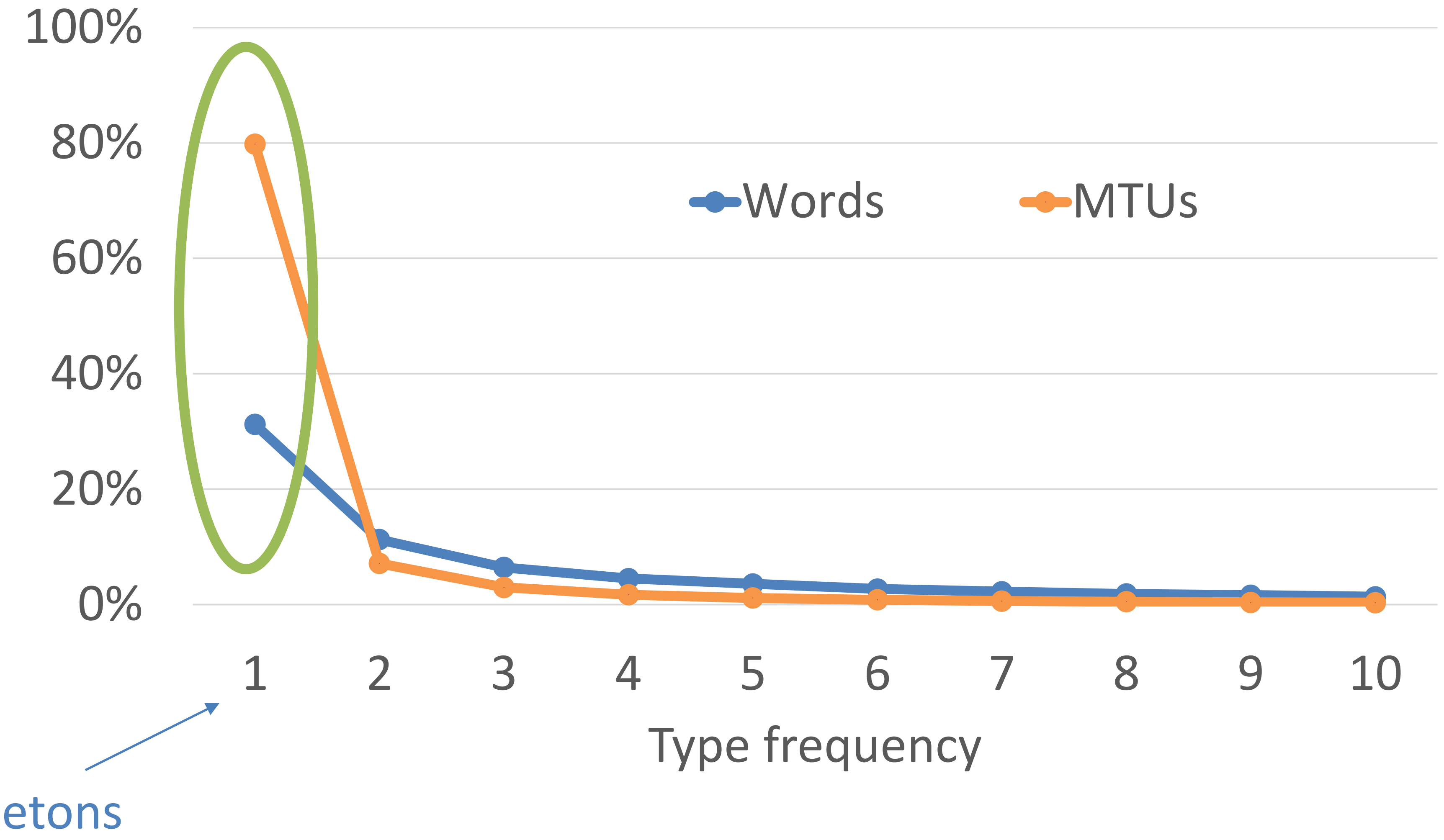
Model 1: Recurrent Atomic MTU



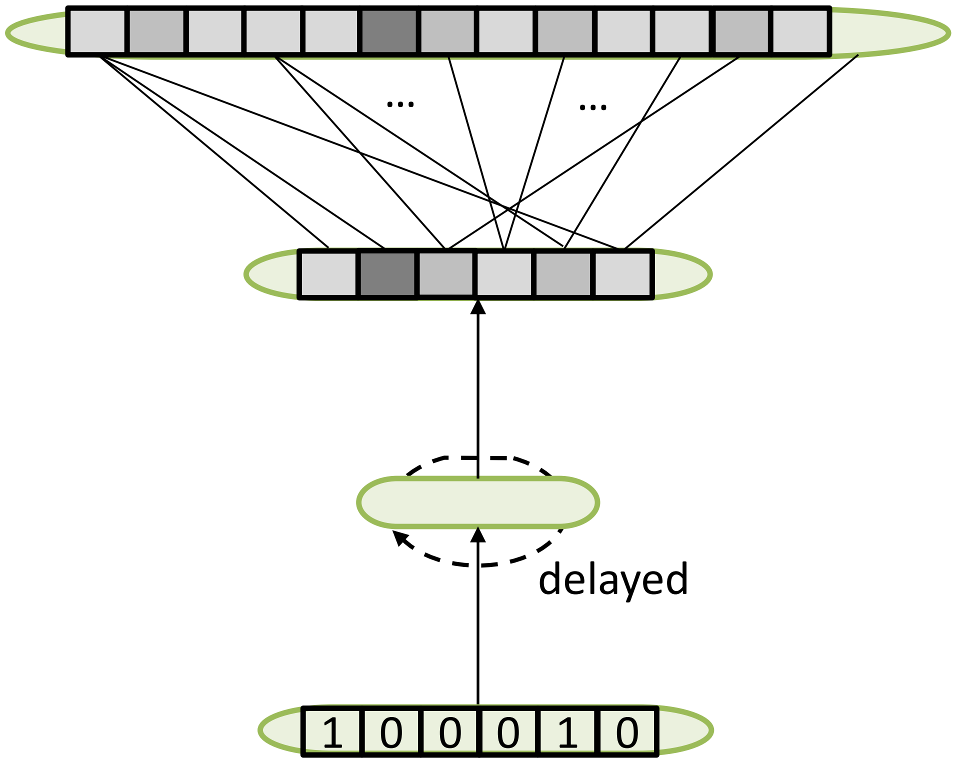
Data sparsity



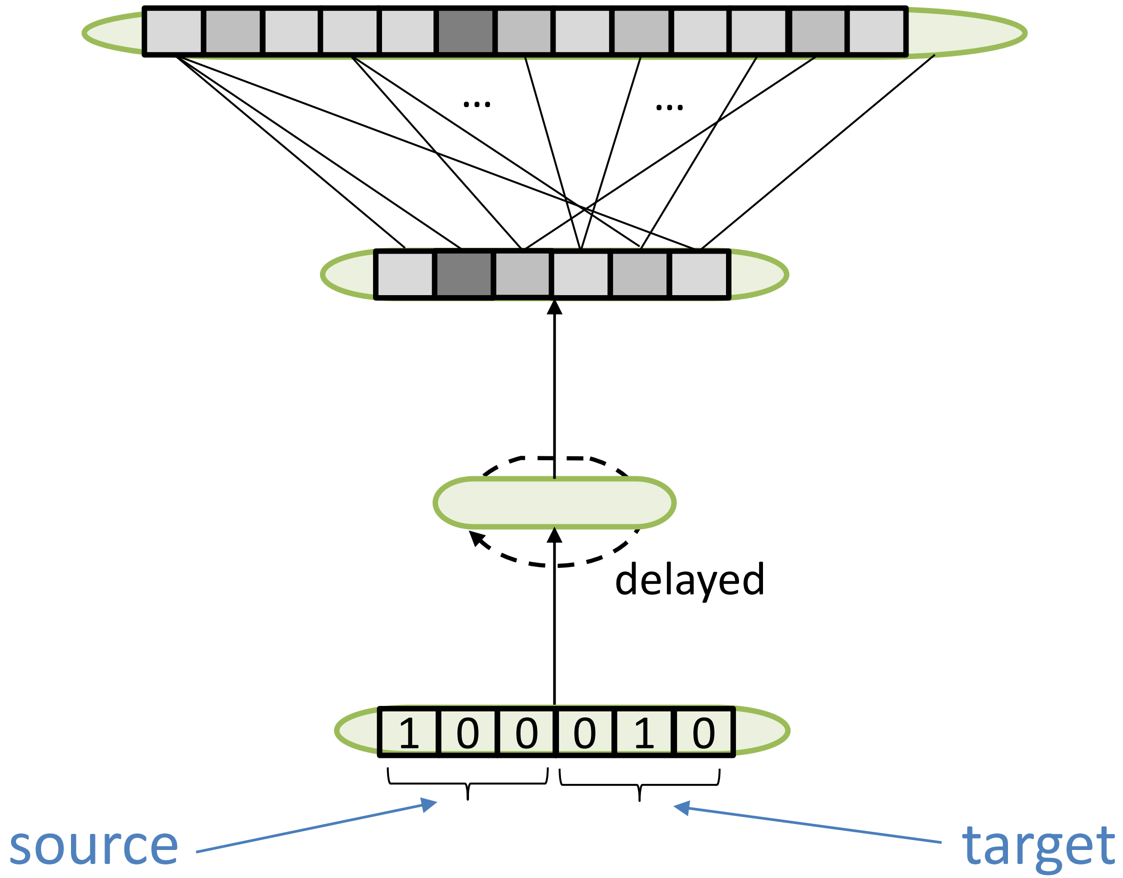
Data sparsity



Model 2: Bag of Words MTU

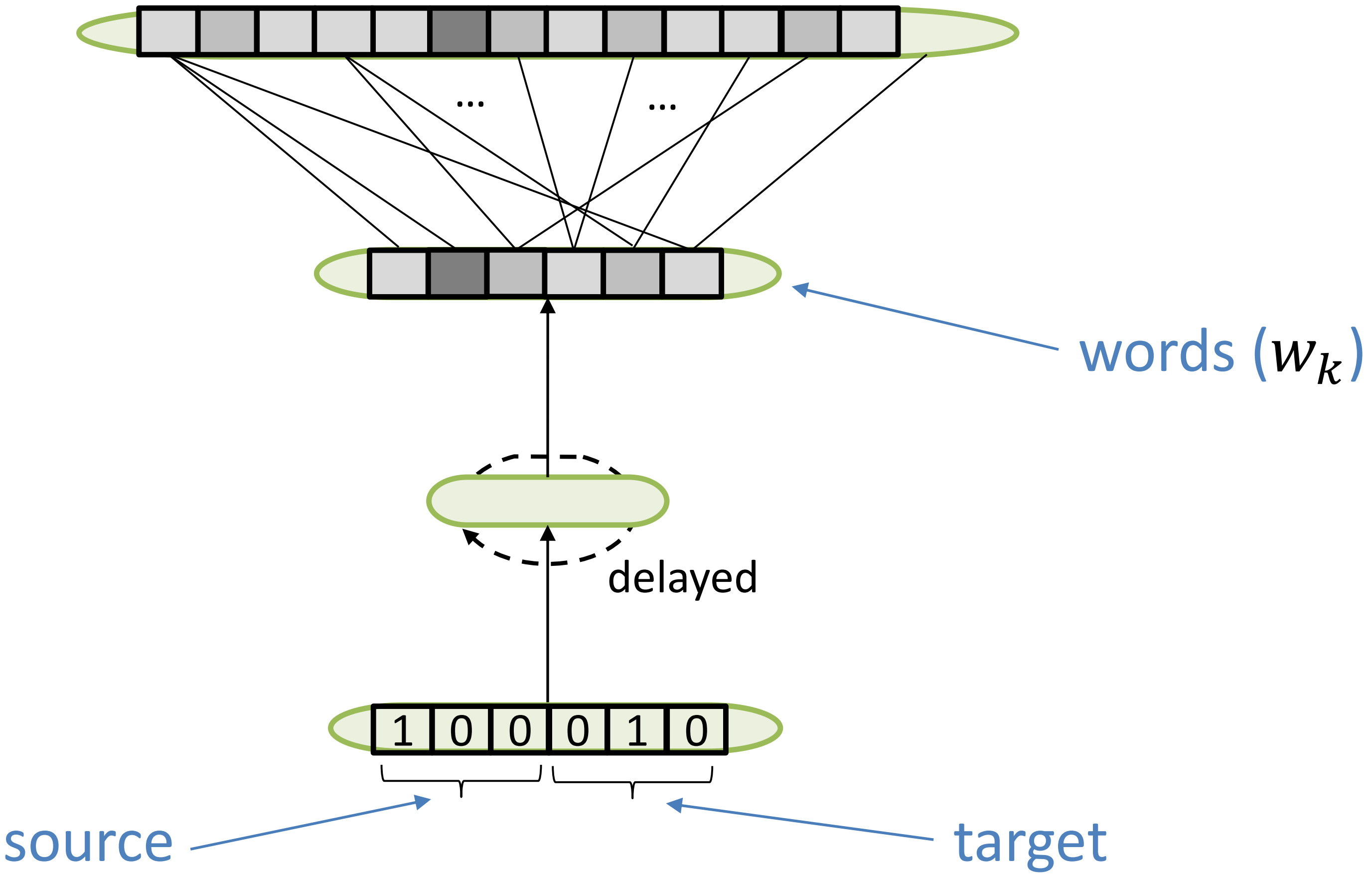


Model 2: Bag of Words MTU



Input: previous MTU as
bag of words

Model 2: Bag of Words MTU

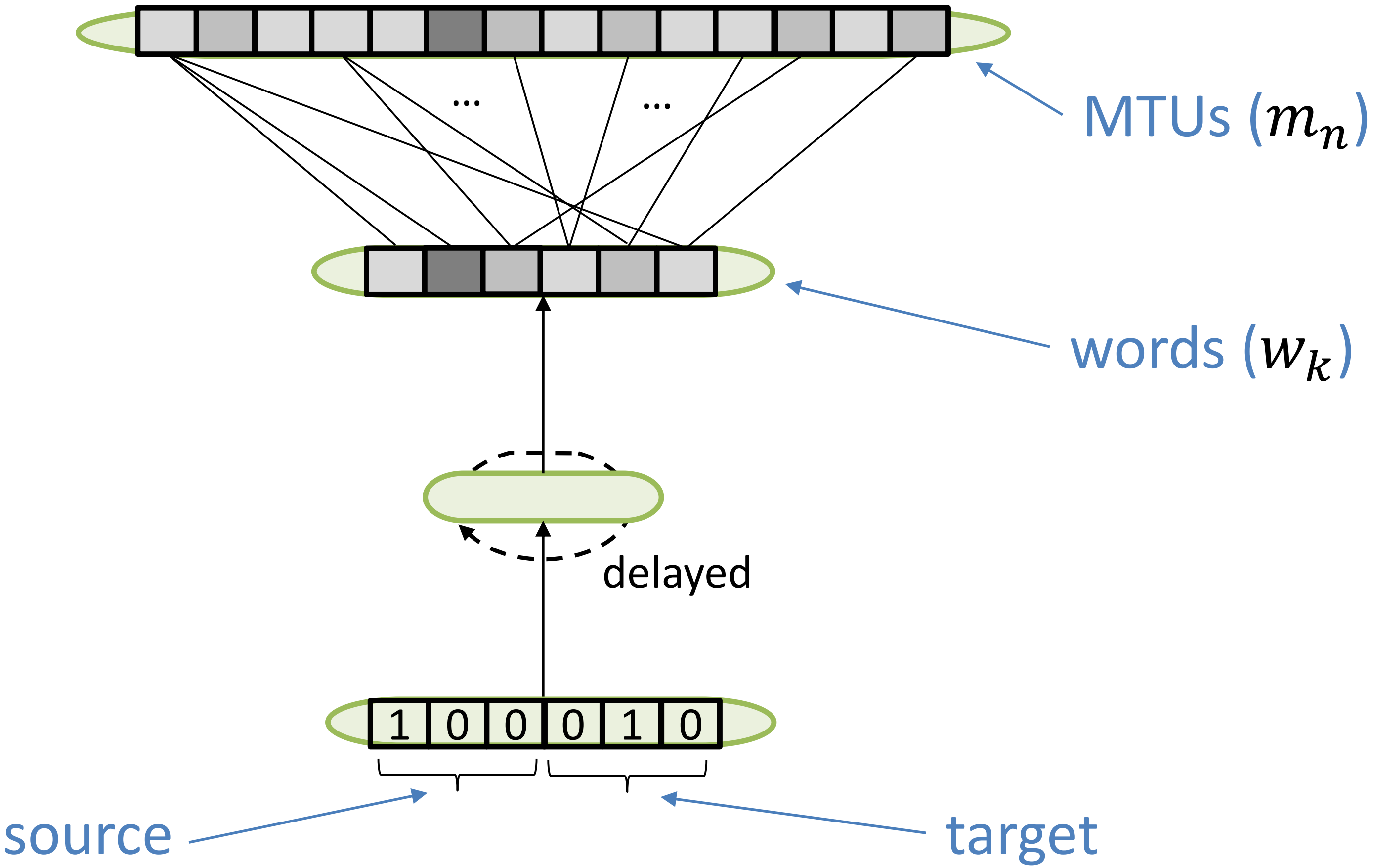


Input: previous MTU as
bag of words

Model 2: Bag of Words MTU

Add MTU output layer

Input: previous MTU as bag of words

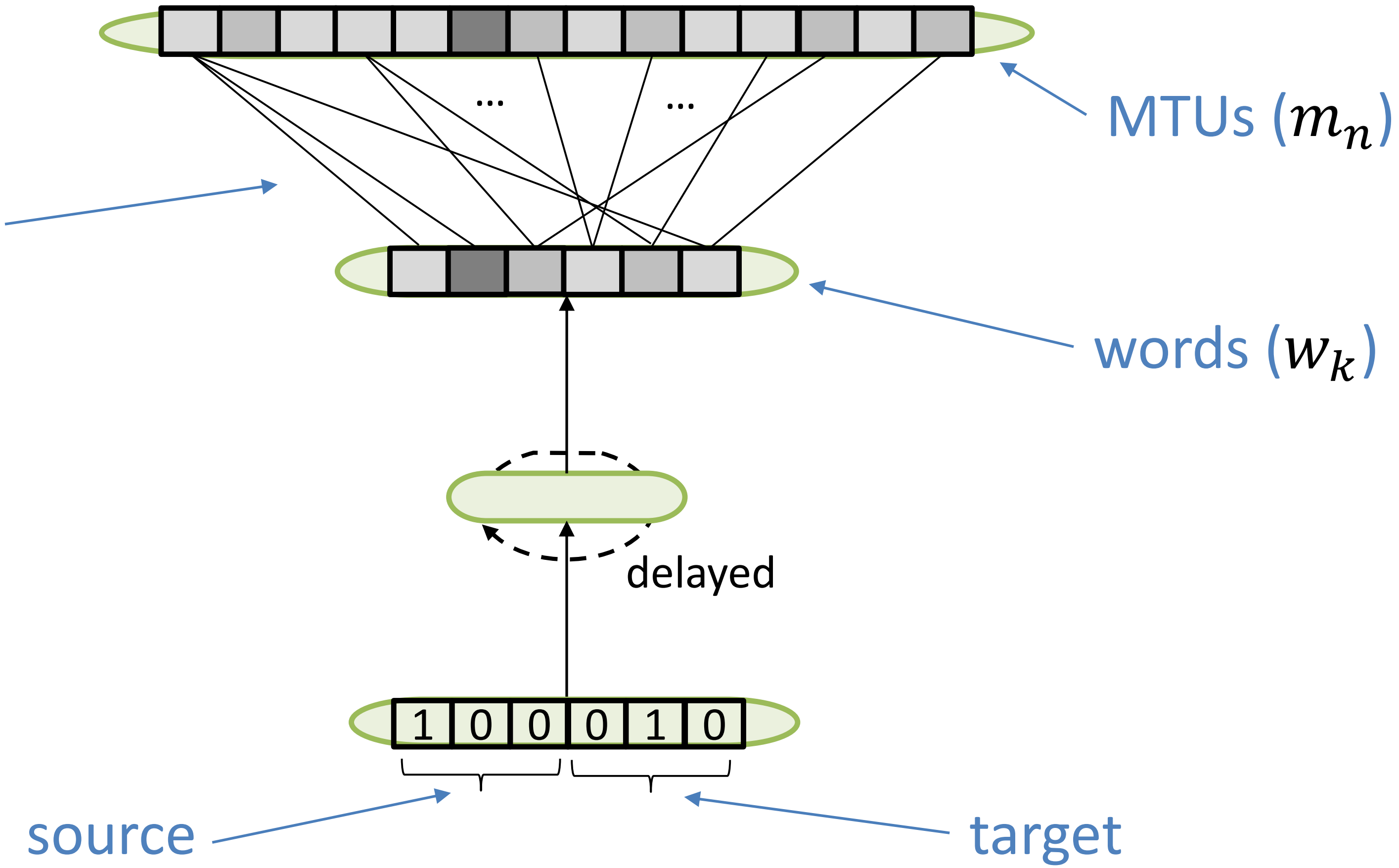


Model 2: Bag of Words MTU

Sparse Mapping MTUs - words

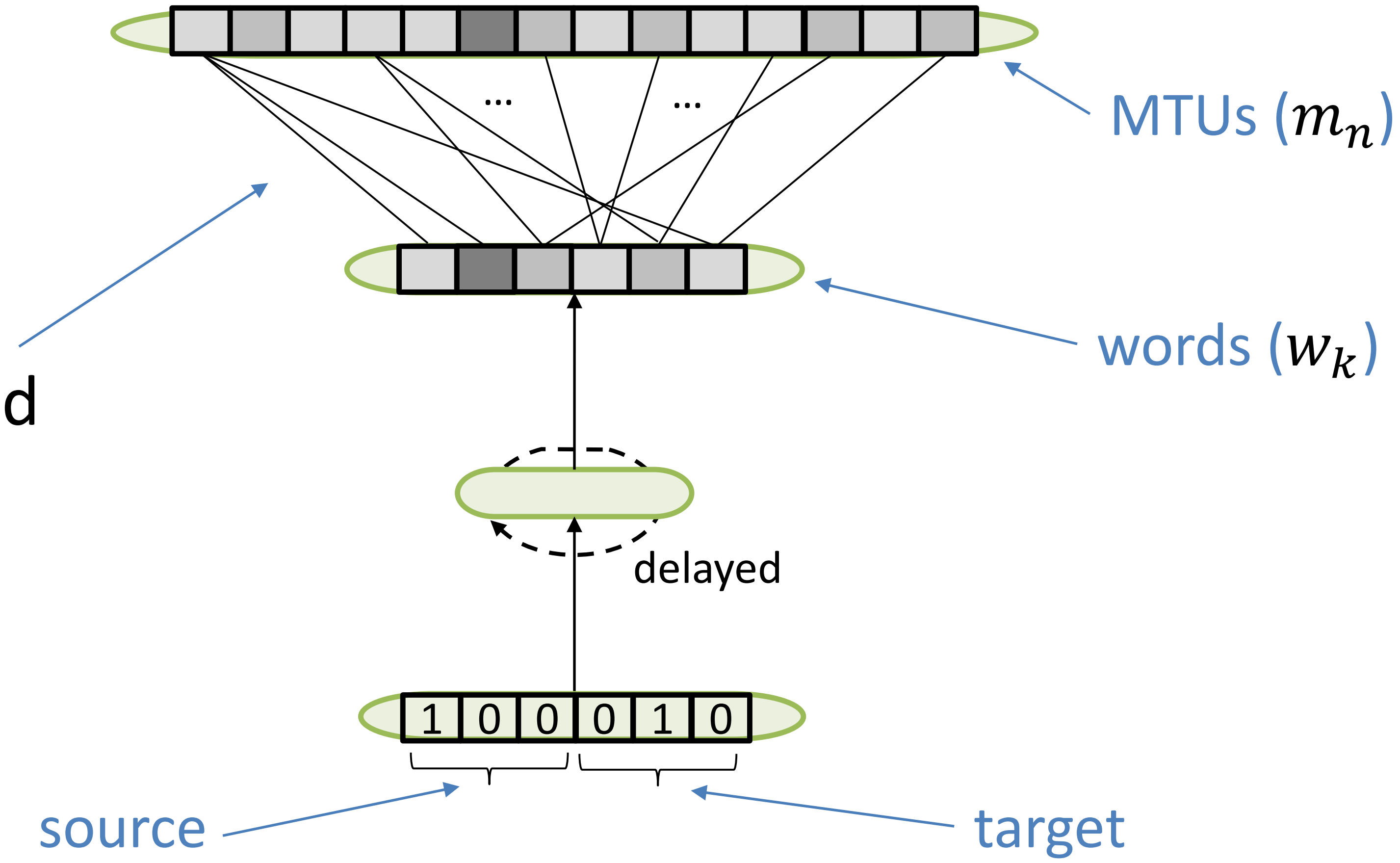
Add MTU output layer

Input: previous MTU as
bag of words

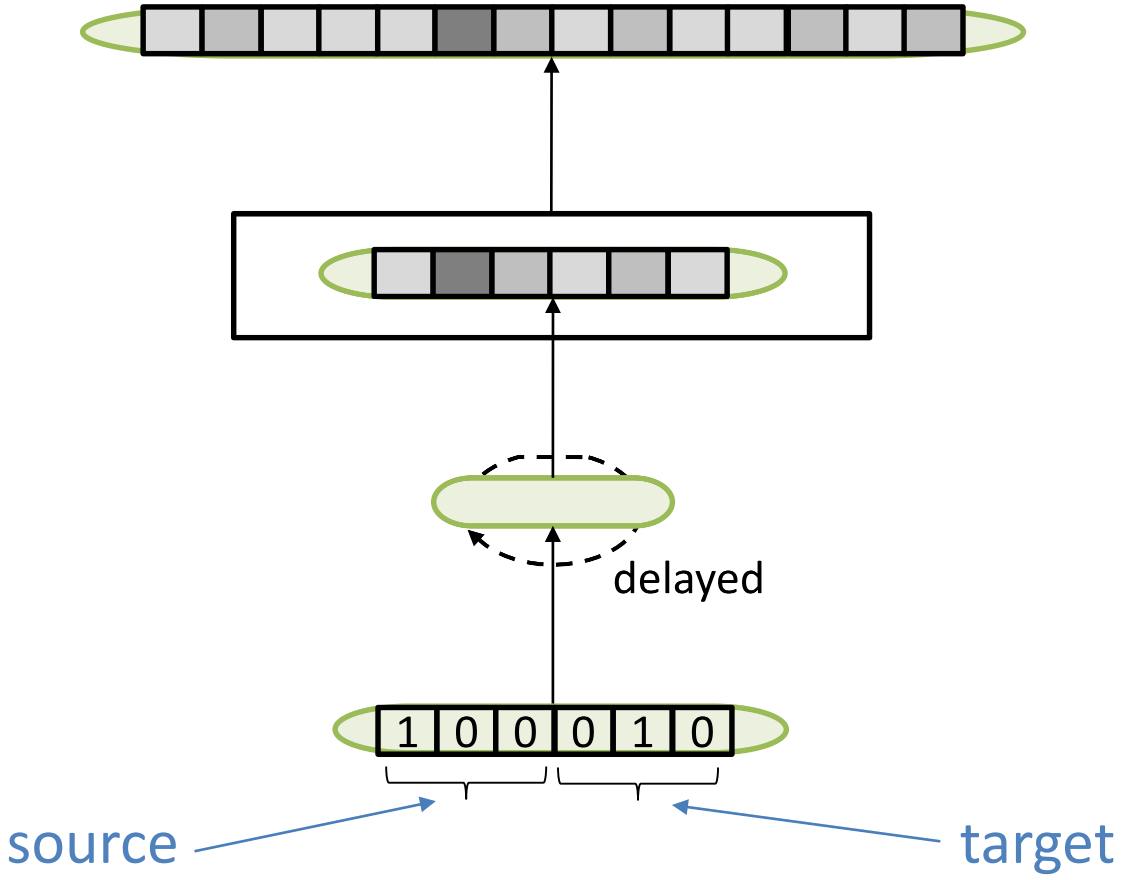


Model 2: Bag of Words MTU

Sparse mapping very imbalanced

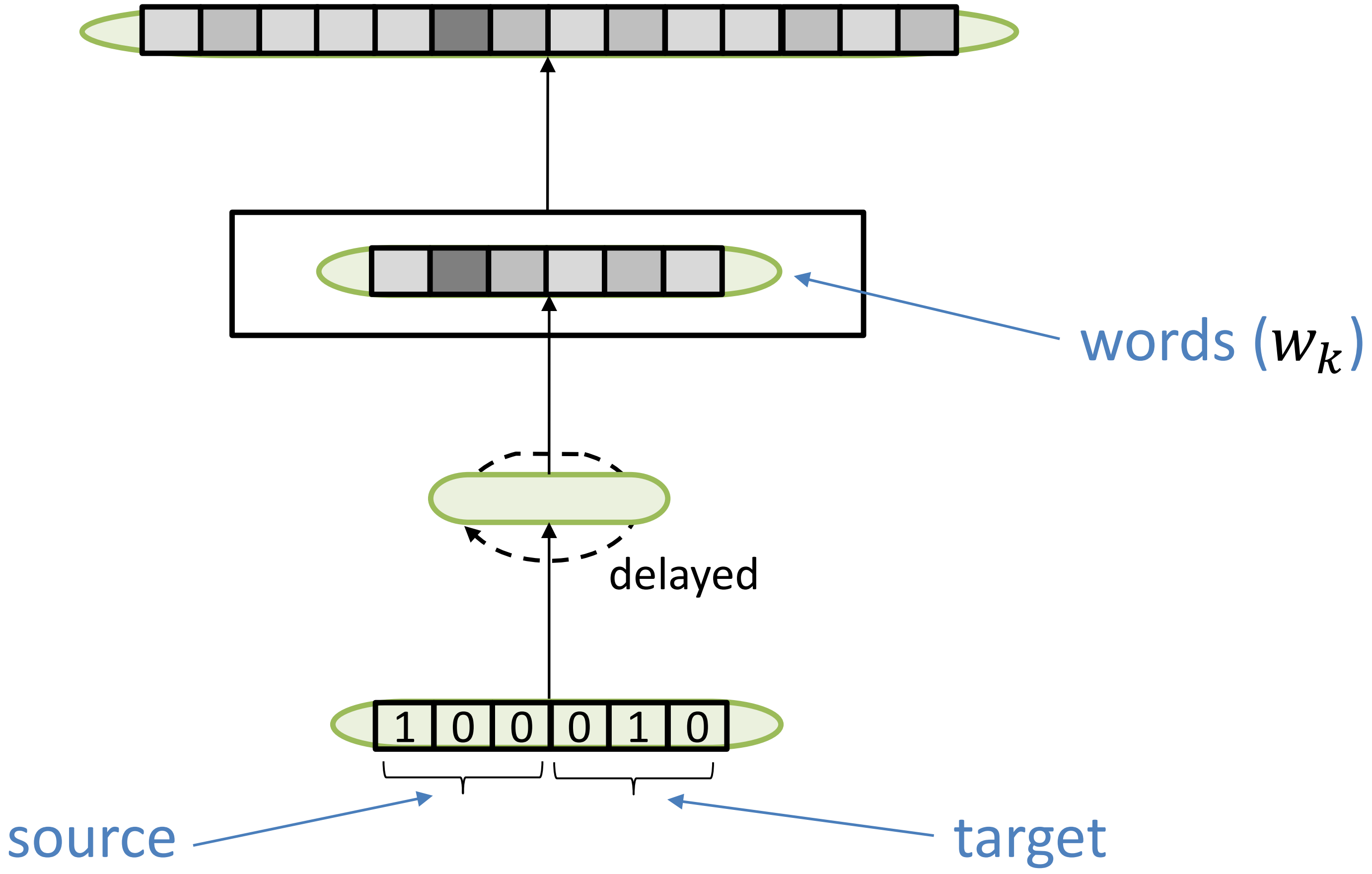


Model 2: Simplified Bag of Words MTU



Input: previous MTU as
bag of words

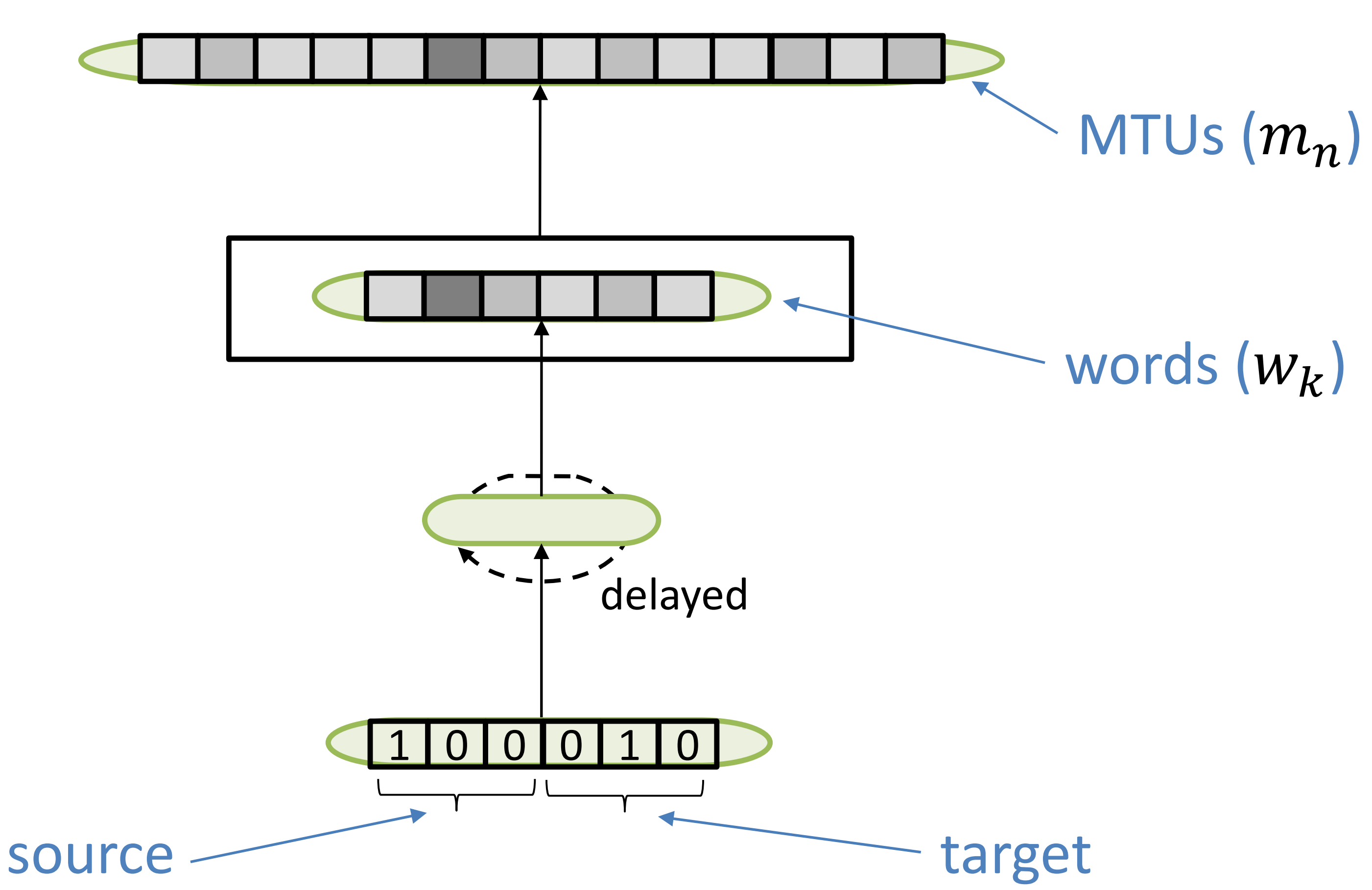
Model 2: Simplified Bag of Words MTU



Input: previous MTU as
bag of words

Model 2: Simplified Bag of Words MTU

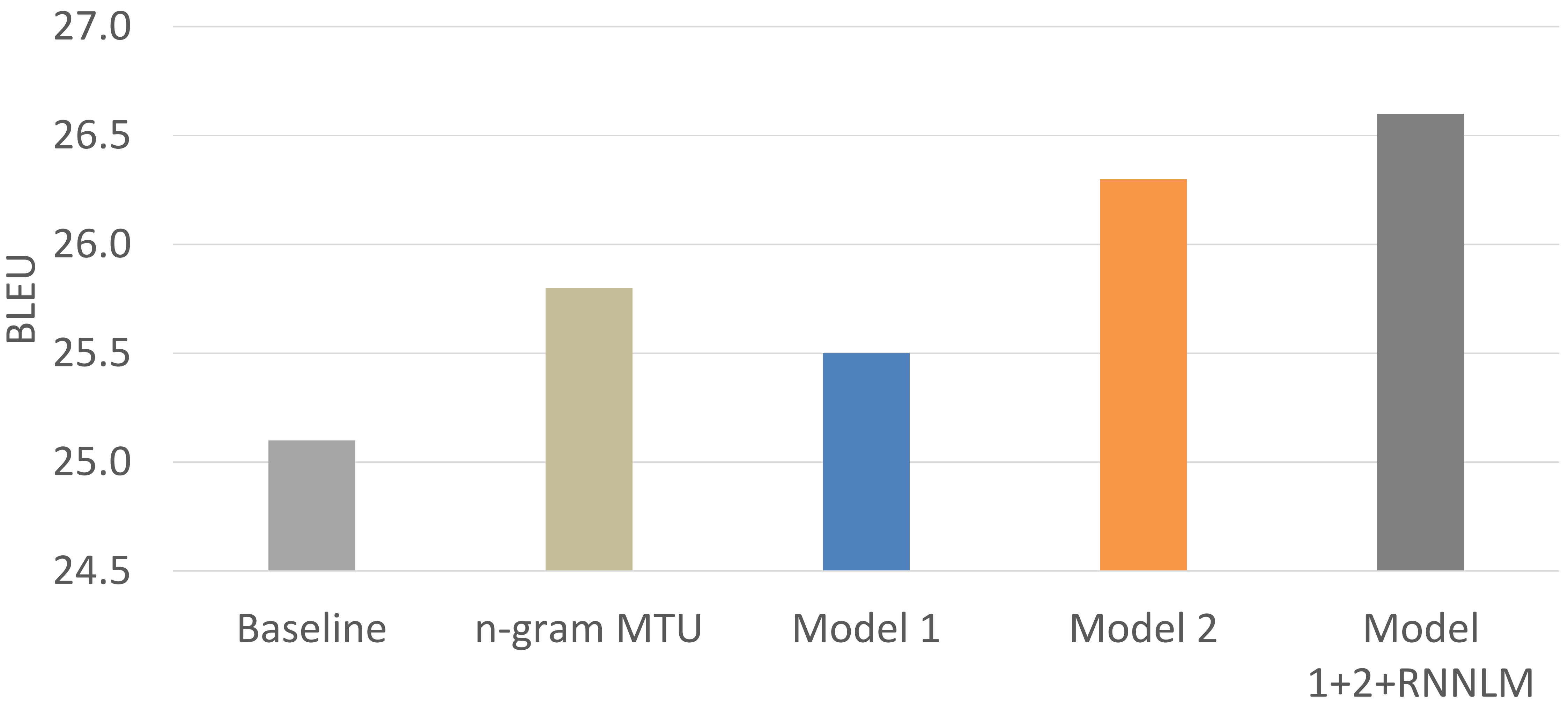
$$p(m_n|h) = \prod_{w \in m_n} p(w|h)$$



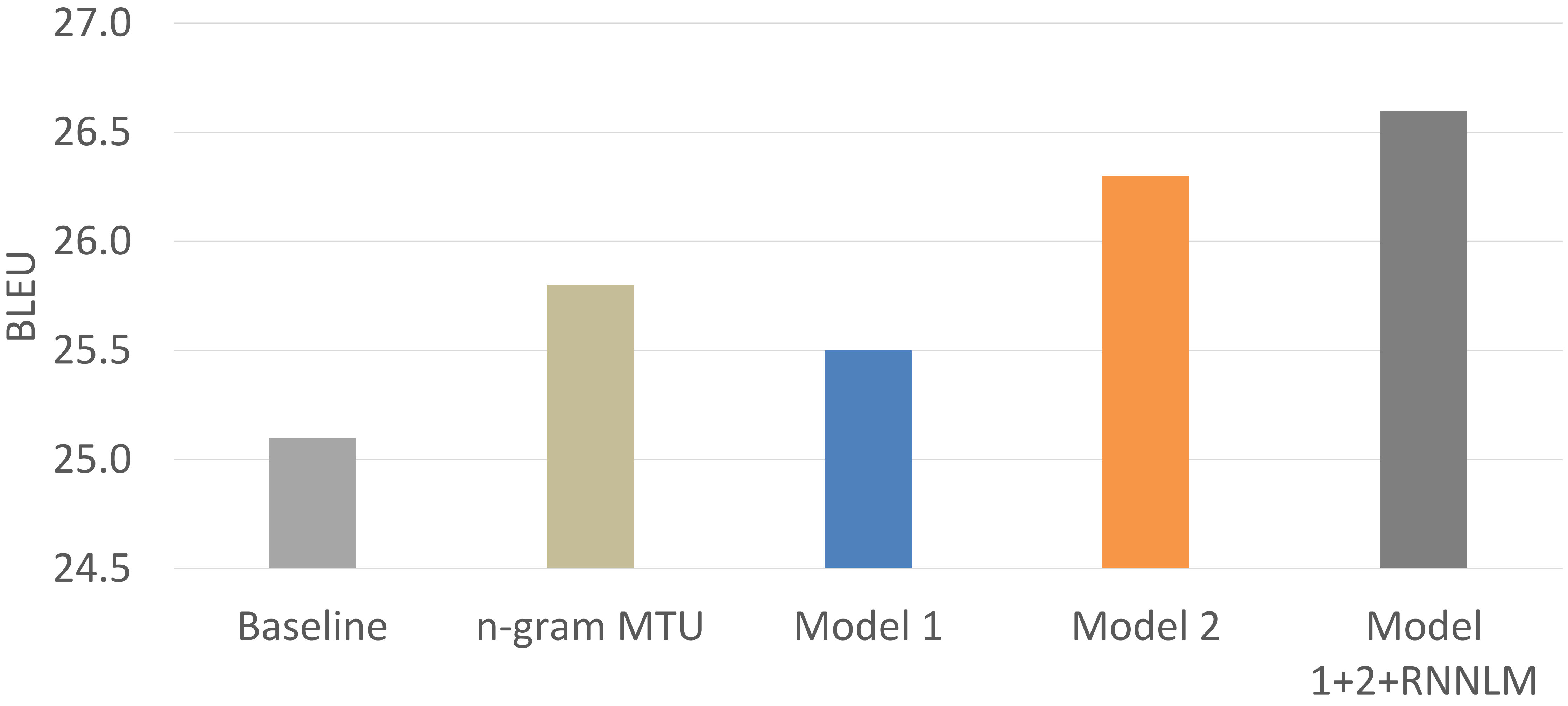
Input: previous MTU as bag of words

Results

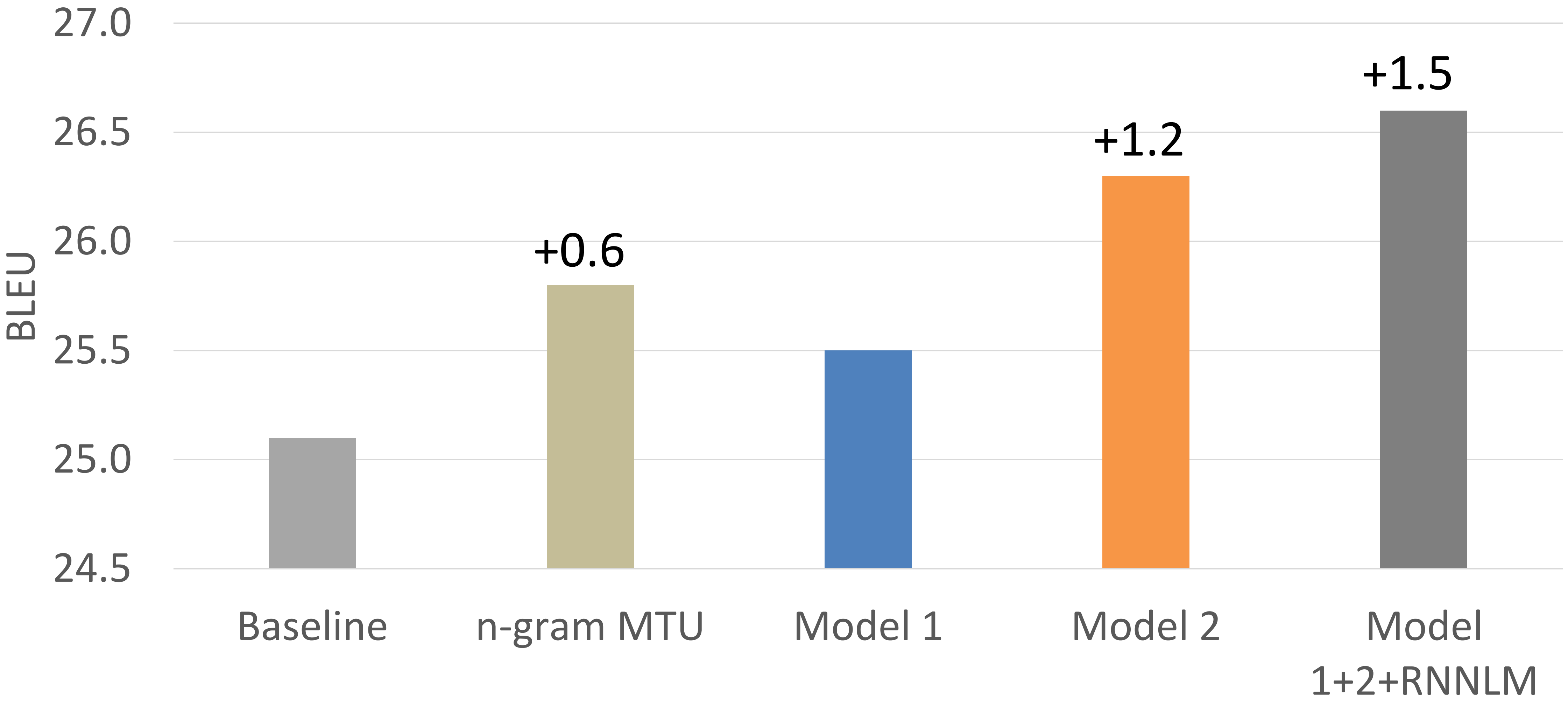
Results


27.0
26.5
26.0
25.5
25.0
24.5
BLEU
Baseline
n-gram MTU
Model 1
Model 2
Model 1+2+RNNLM

Results


BLEU
27.0
26.5
26.0
25.5
25.0
24.5
Baseline
n-gram MTU
Model 1
Model 2
Model 1+2+RNNLM

Results


27.0
26.5
26.0
25.5
25.0
24.5
BLEU
+0.6
+1.2
+1.5
Baseline
n-gram MTU
Model 1
Model 2
Model
1+2+RNNLM

Summary so far

- Recurrent net translation models improve phrase-based models (+1.4 BLEU)
- Word-window approach superior to simple sentence representations
- Recurrent MTU models need to be carefully factored
- Bag-of-words factorization adds up to +1.5 BLEU

Overview

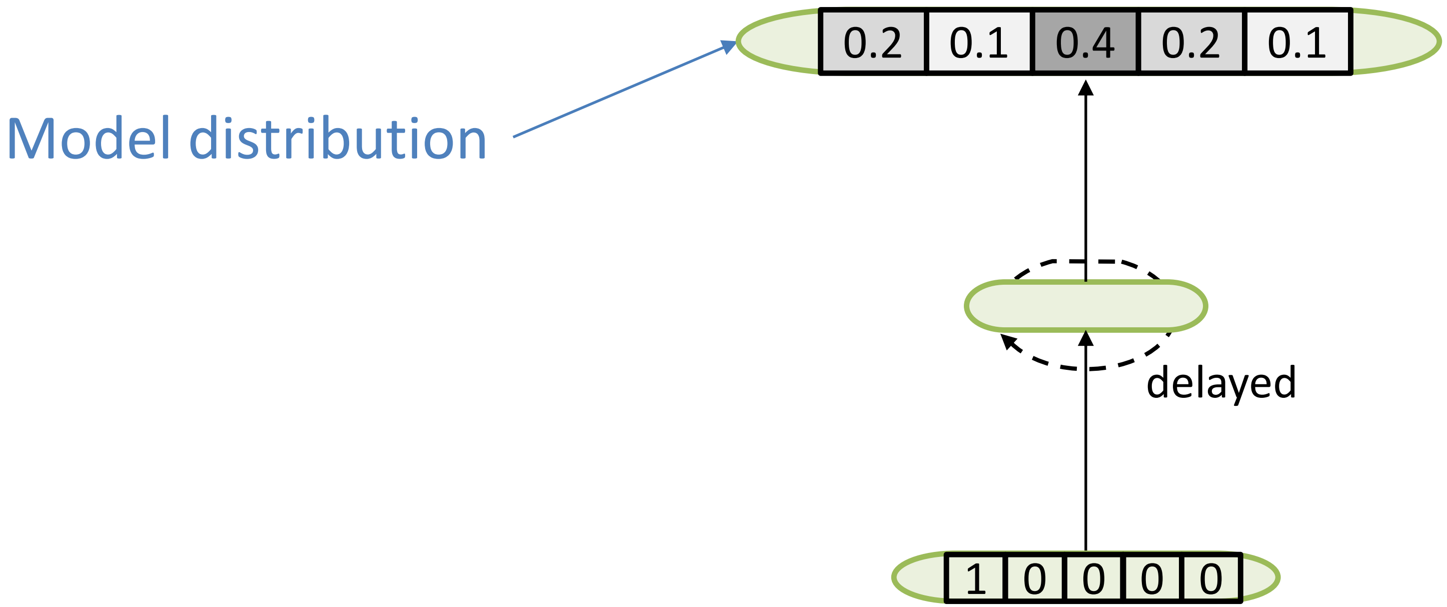
- Recurrent neural network joint models (Auli et al., EMNLP 2013)
  Combined language and translation modeling

- Minimum translation modeling with recurrent nets (Hu et al., EACL 2014)
  Sequence models over bilingual units

- Task-specific training of neural nets (Auli & Gao, ACL 2014)
  Expected BLEU training for neural network translation models

- Large-scale discriminative sparse ordering models (Auli et al., in submission)
  Training millions of linear ordering features with expected BLEU

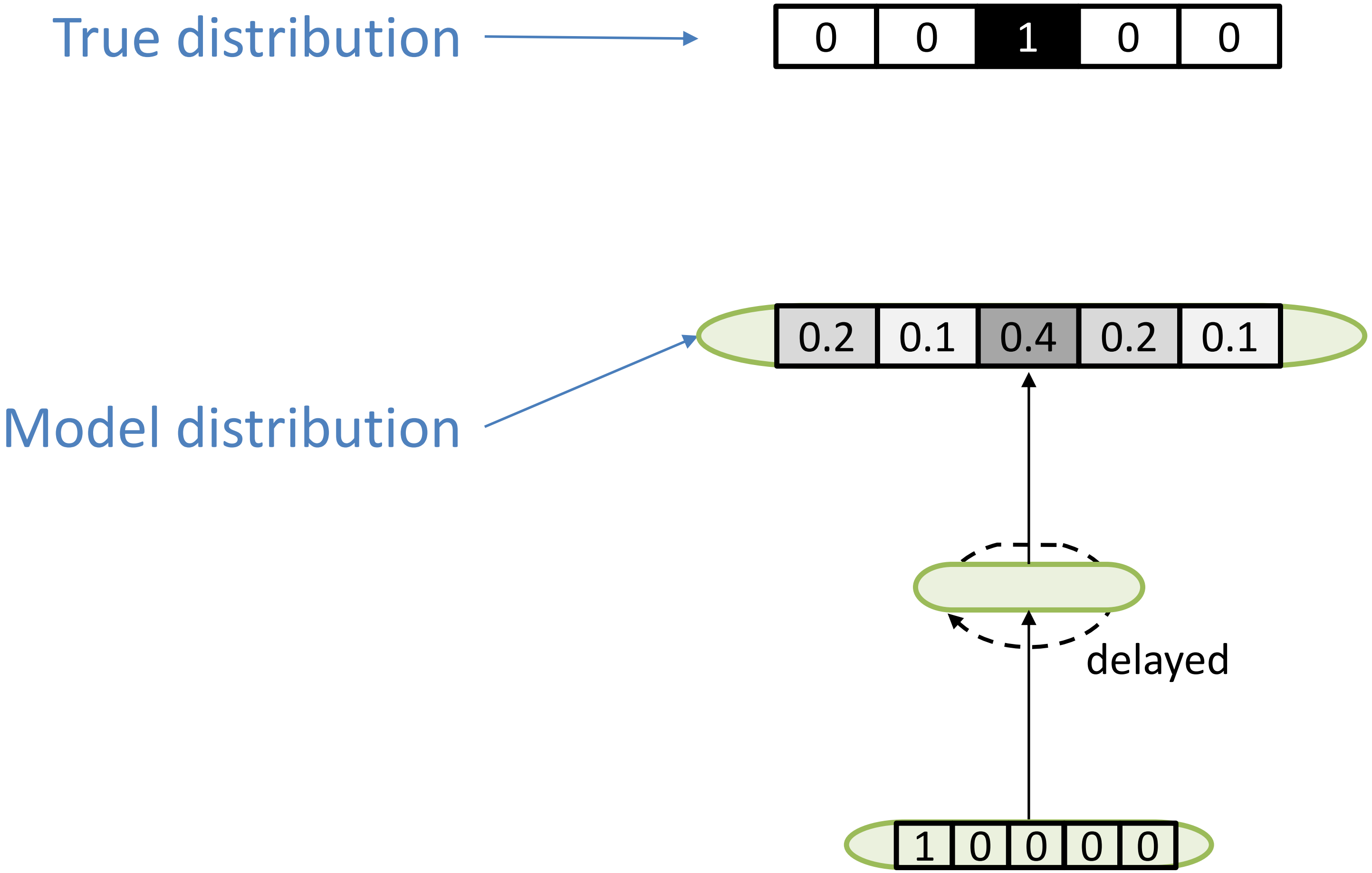
Back propagation with the Cross Entropy Error



$$\max_{\phi} \sum_{i} p(e_i; \phi)$$

Goal: Make reference most likely

Back propagation with the Cross Entropy Error



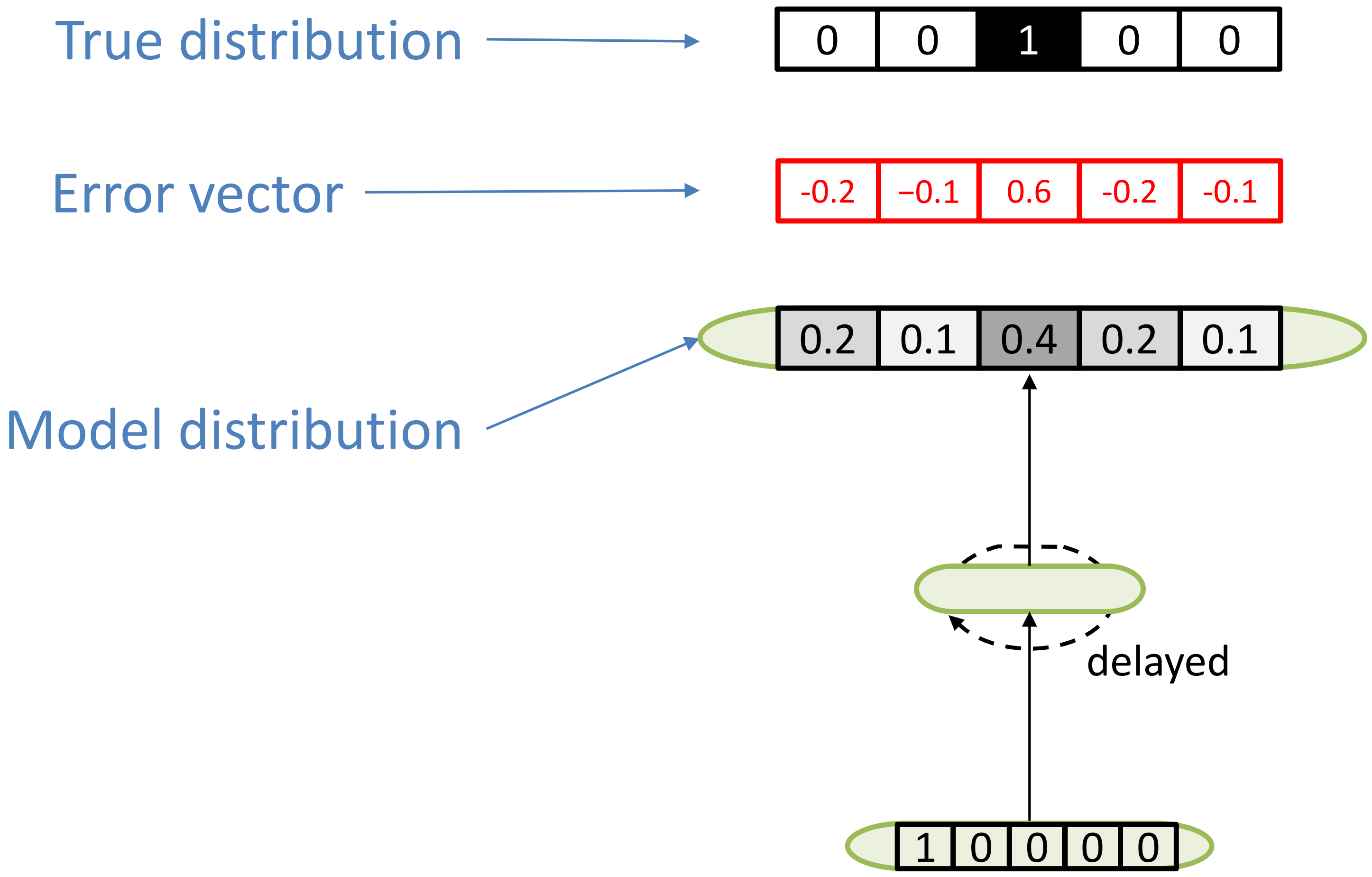
True distribution

Model distribution

$$\max_{\phi} \sum_{i} p(e_i; \phi)$$

Goal: Make reference most likely

Back propagation with the Cross Entropy Error



True distribution

0	0	1	0	0

Error vector

-0.2	–0.1	0.6	-0.2	-0.1

Model distribution

0.2	0.1	0.4	0.2	0.1

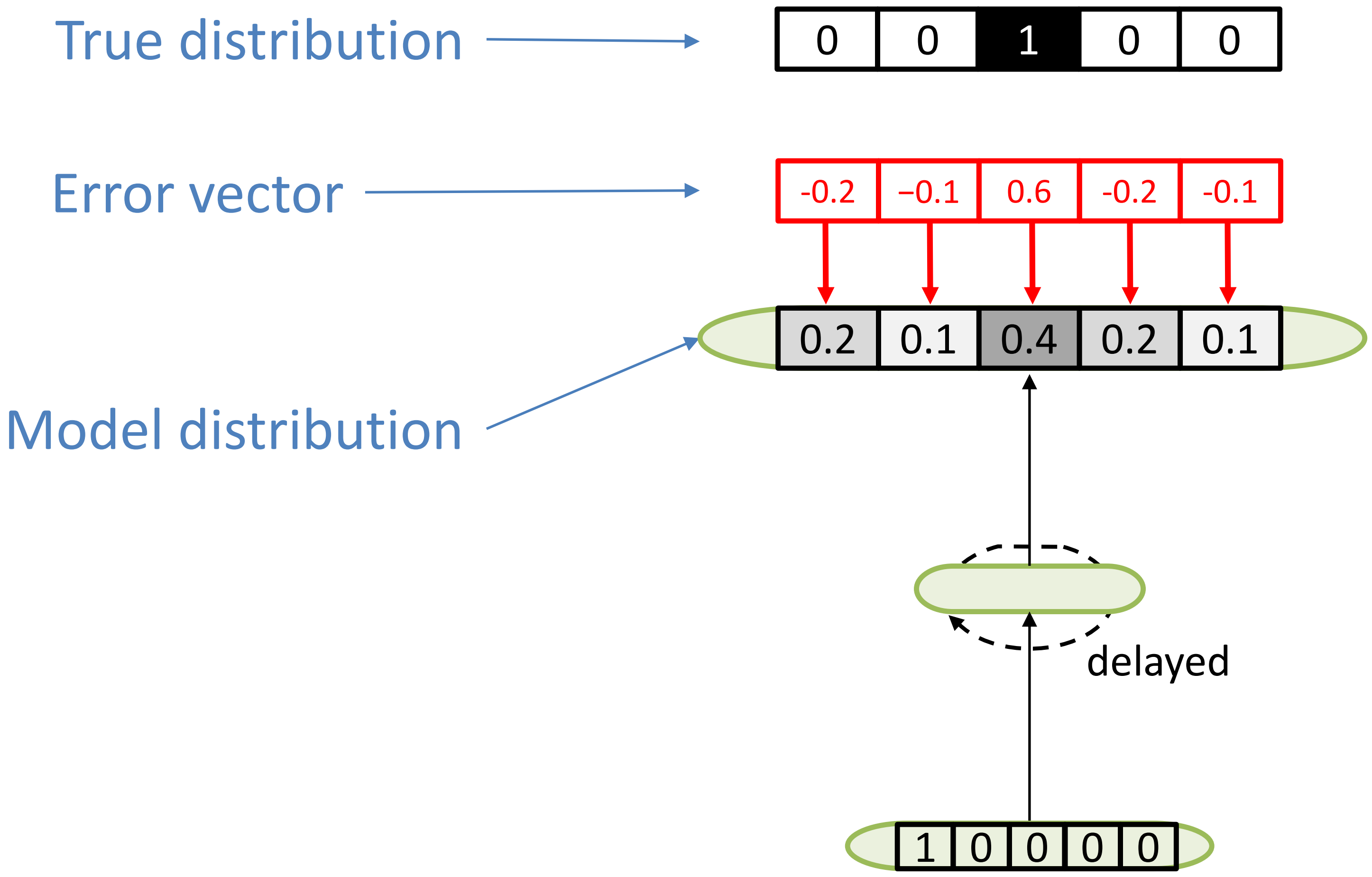
delayed

1	0	0	0	0

$$\max_{\phi} \sum_{i} p(e_i; \phi)$$

Goal: Make reference most likely

Back propagation with the Cross Entropy Error



True distribution

0	0	1	0	0

Error vector

-0.2	–0.1	0.6	-0.2	-0.1

Model distribution

0.2	0.1	0.4	0.2	0.1

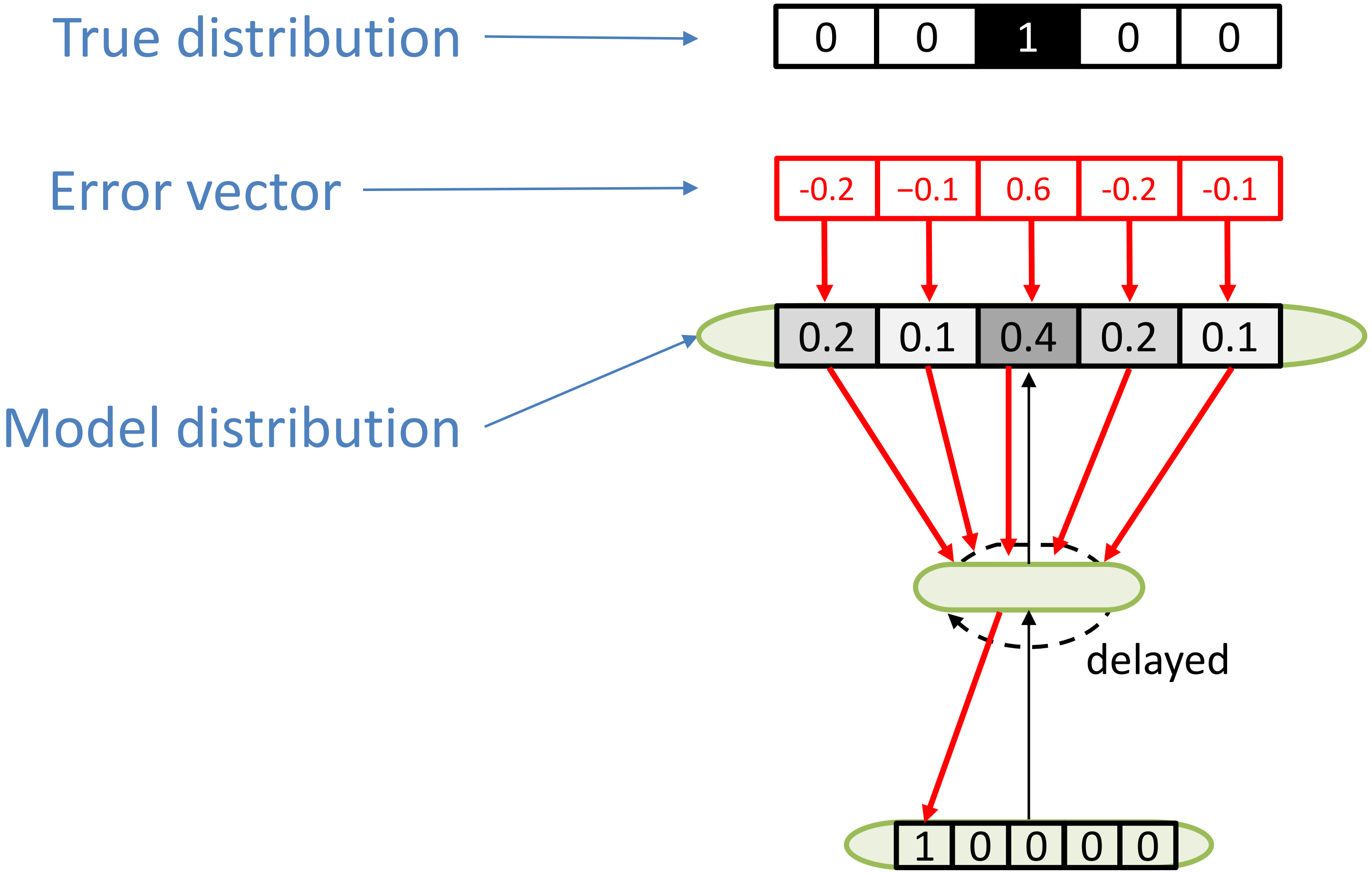
delayed

1	0	0	0	0

$$\max_{\phi} \sum_{i} p(e_i; \phi)$$

Goal: Make reference most likely

Back propagation with the Cross Entropy Error



True distribution

0	0	1	0	0

Error vector

-0.2	−0.1	0.6	-0.2	-0.1

Model distribution

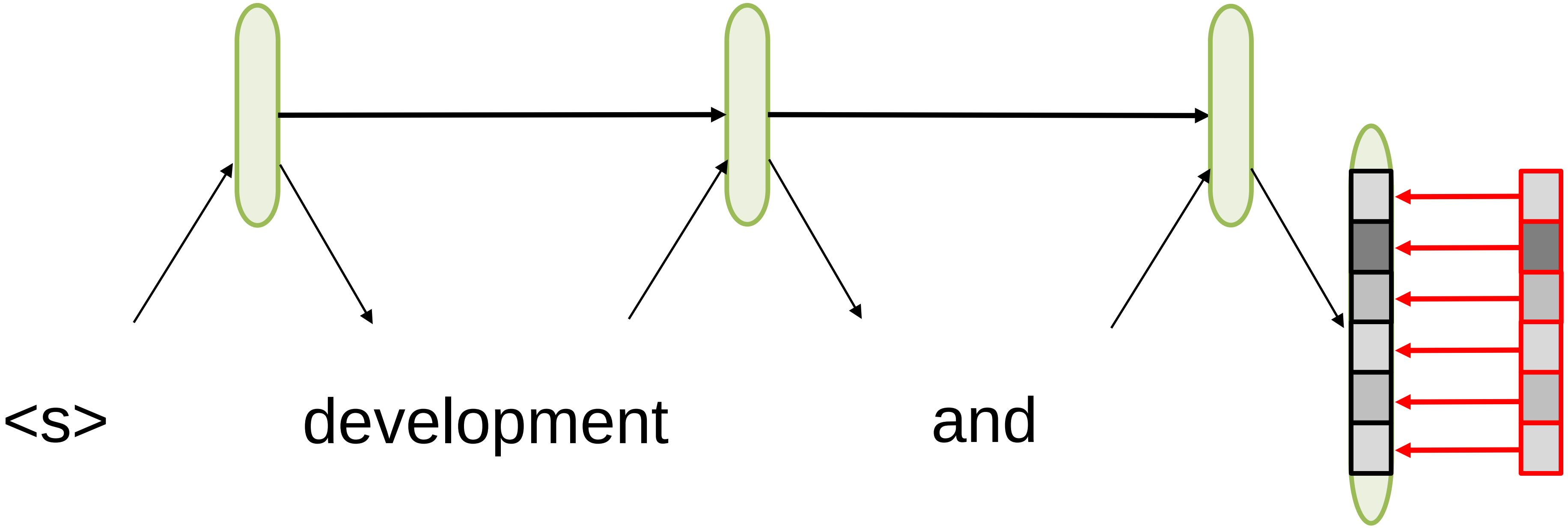
0.2	0.1	0.4	0.2	0.1

1	0	0	0	0

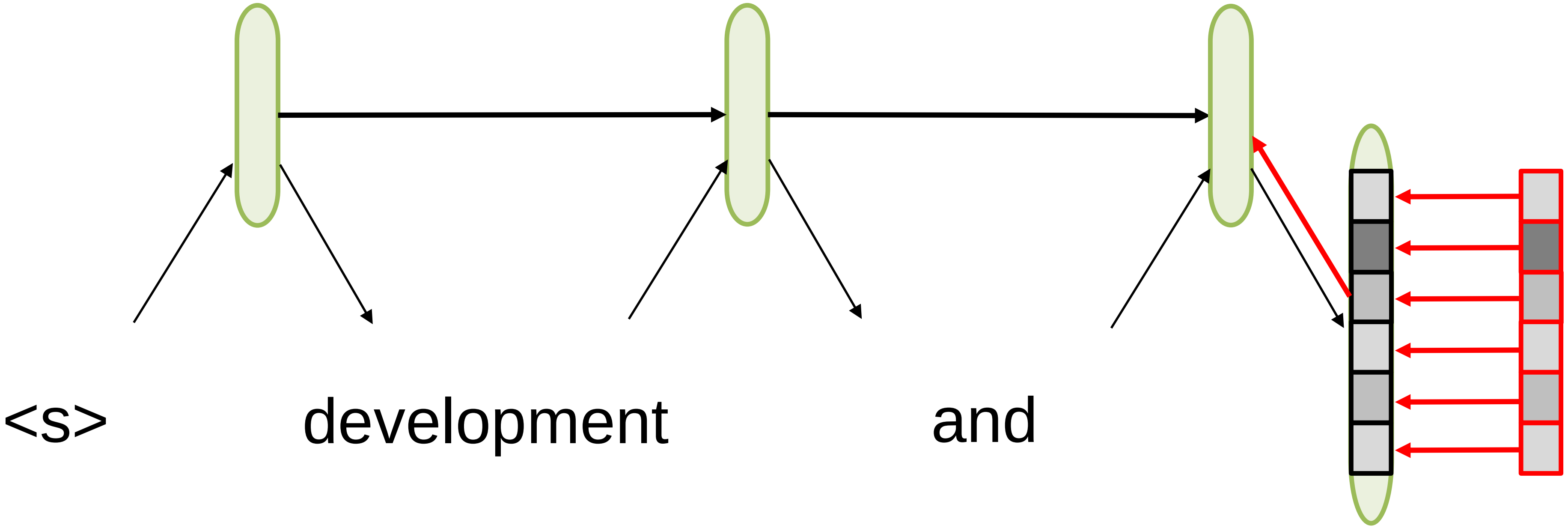
$$\max_{\phi} \sum_{i} p(e_i; \phi)$$

Goal: Make reference most likely

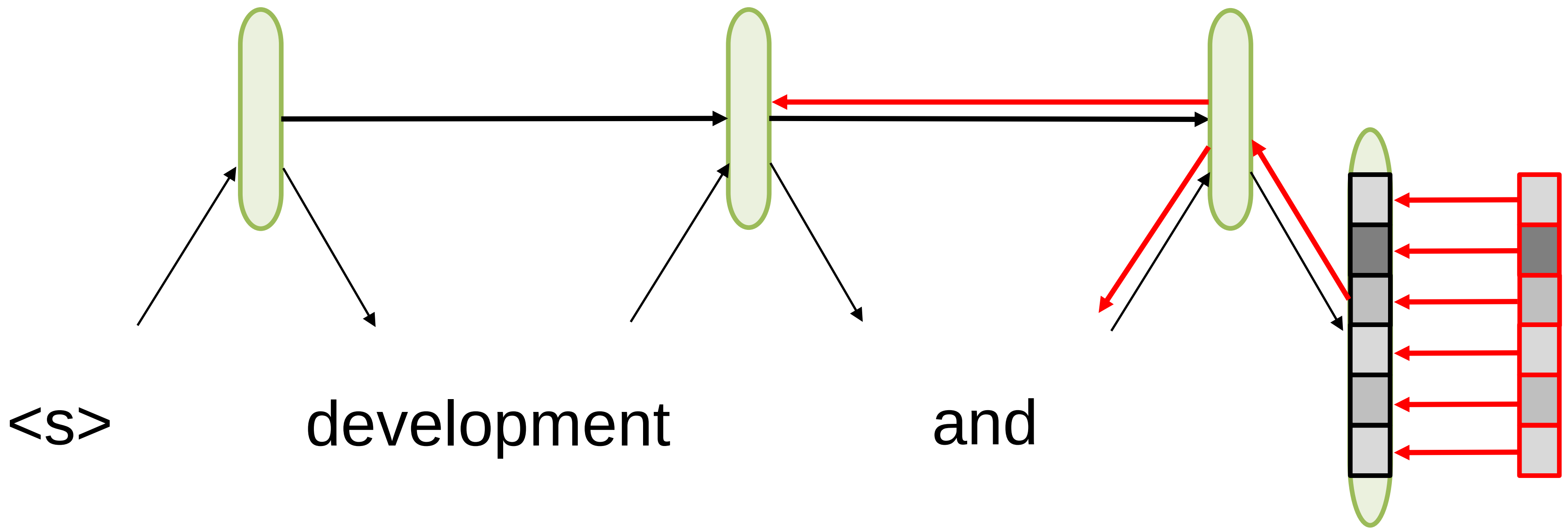
Back propagation through time



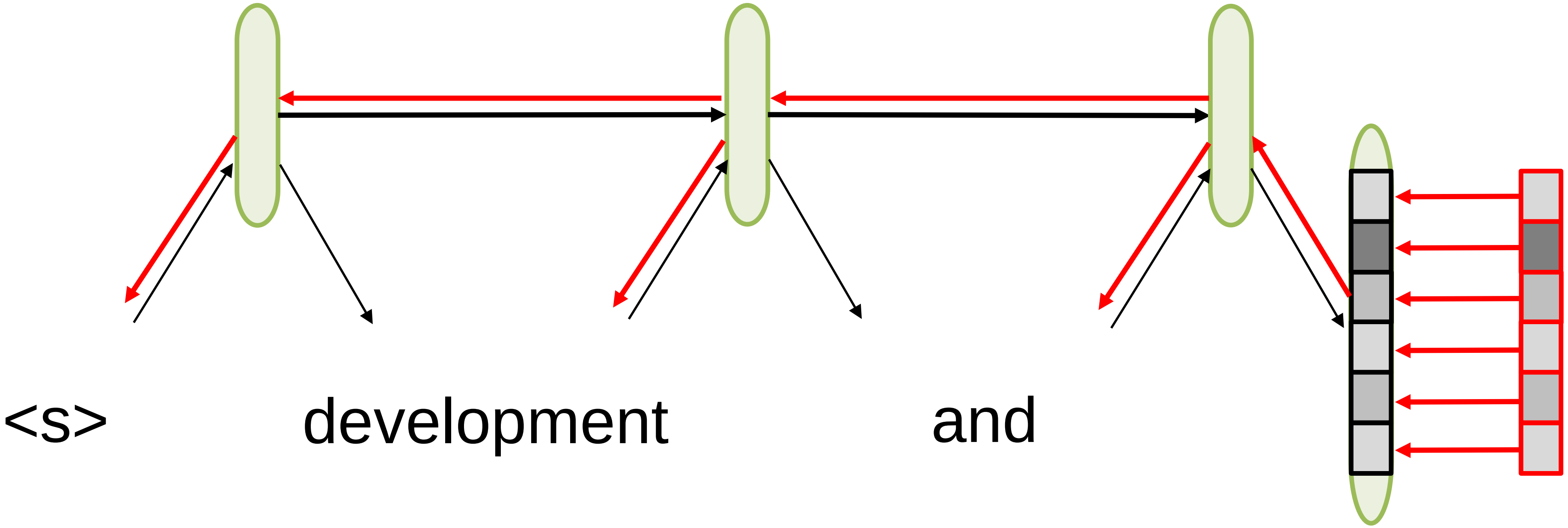
Back propagation through time



Back propagation through time



Back propagation through time



<s>

development

and

Task-specific Optimization

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

Task-specific Optimization

- Likelihood training very common
- Optimizing for evaluation metrics difficult, but empirically successful (Och 2003, Smith 2006, Chiang 2009, Gimpel 2010, Hopkins 2011)
- Objectives usually not convex

Task-specific Optimization

- Likelihood training very common
- Optimizing for evaluation metrics difficult, but empirically successful (Och 2003, Smith 2006, Chiang 2009, Gimpel 2010, Hopkins 2011)
- Objectives usually not convex
- But empirically effective

Task-specific Optimization

- Likelihood training very common
- Optimizing for evaluation metrics difficult, but empirically successful (Och 2003, Smith 2006, Chiang 2009, Gimpel 2010, Hopkins 2011)
- Objectives usually not convex
- But empirically effective
- **Next:** Task-specific training of neural nets for translation

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

Modified precision scores

Brevity penalty

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

Modified precision scores

Brevity penalty

Human: development and progress of the region

System: advance and progress of region

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

Modified precision scores

Brevity penalty

Human: development and progress of the region

System: advance and progress of region

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

Modified precision scores

Brevity penalty

Human: development and progress of the region

System: advance and progress of region

BLEU Metric

(Bilingual Evaluation Understudy; Papineni 2002)

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

Modified precision scores

Brevity penalty

Human: development and progress of the region

System: advance and progress of region

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

L: $\max_{\phi} \sum_{i} p(e_i | f_i; \phi)$

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

Desired translation output

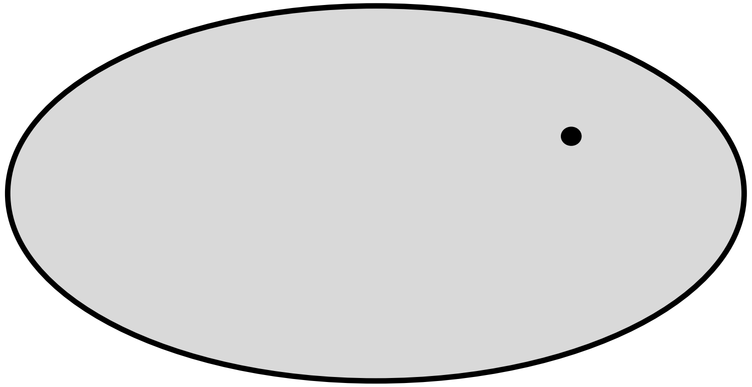
L: $$\max_{\phi} \sum_{i} p(e_i | f_i; \phi)$$

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

Desired translation output

L: $$\max_{\phi} \sum_{i} p(e_i|f_i;\phi)$$



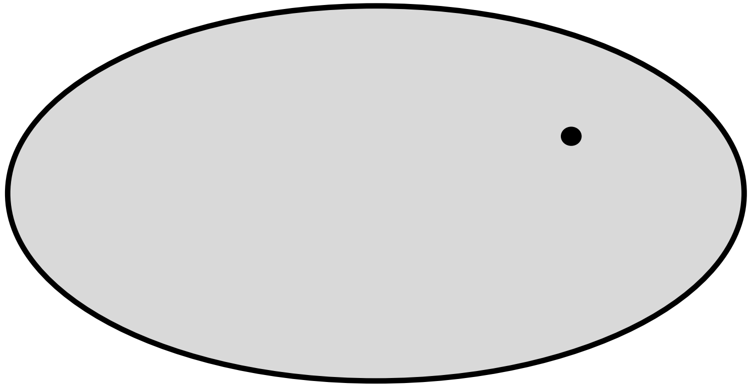
xBLEU: $$\max_{\phi} \sum_{i} \sum_{e \in E(f_i)} \mathrm{sBLEU}(e,e_i)\, p(e|f_i;\phi)$$

Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

Desired translation output

L: $$\max_{\phi} \sum_{i} p(e_i|f_i;\phi)$$



xBLEU: $$\max_{\phi} \sum_{i} \sum_{e\in E(f_i)} \mathrm{sBLEU}(e,e_i)\, p(e|f_i;\phi)$$

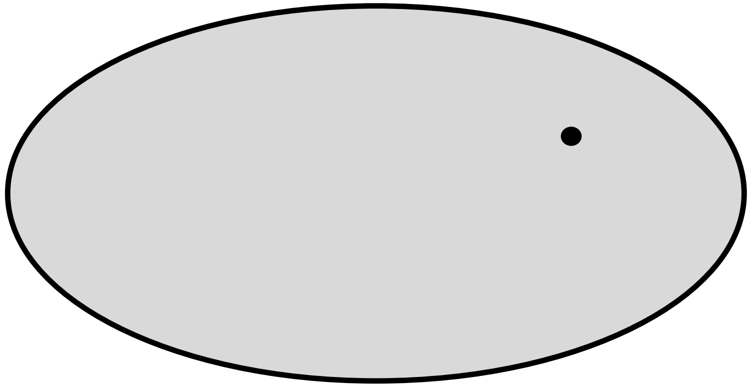
Generated outputs Gain function Human translation

Expected BLEU Training

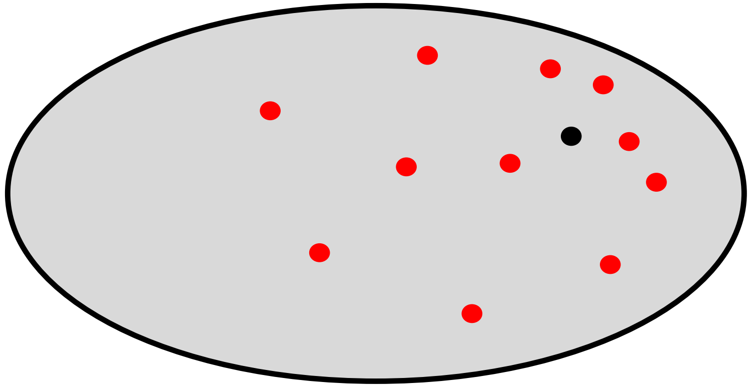
(Smith 2006, He 2012, Gao 2014)

Desired translation output

L: $$\max_{\phi} \sum_{i} p(e_i | f_i; \phi)$$



xBLEU: $$\max_{\phi} \sum_{i} \sum_{e \in E(f_i)} \mathrm{sBLEU}(e, e_i)\, p(e | f_i; \phi)$$

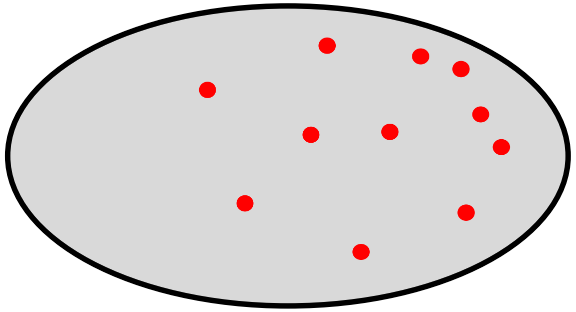


Generated outputs

Gain function

Human translation

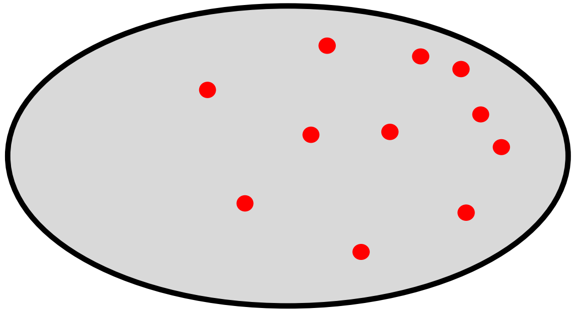
Expected BLEU Training



本 地区 的 发展 和 进步

Human: development and progress of the region

Expected BLEU Training

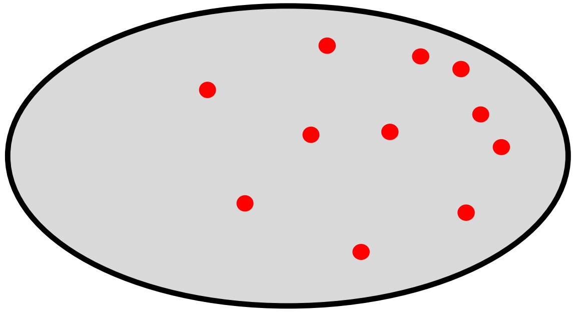


本 地区 的 发展 和 进步

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

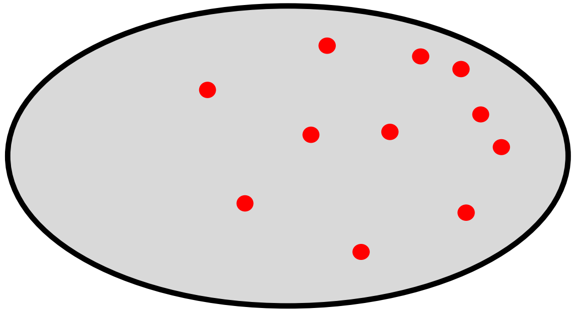


本 地区 的 发展 和 进步

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

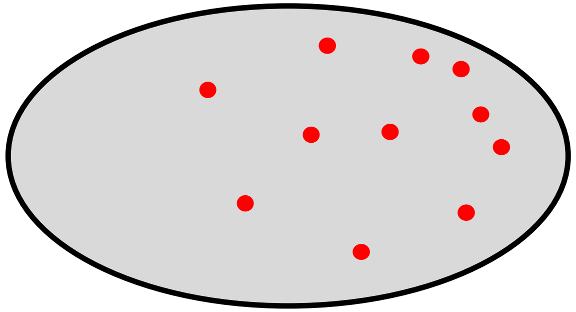


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$
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

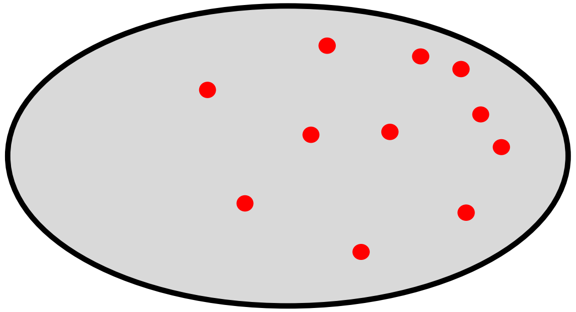


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$
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

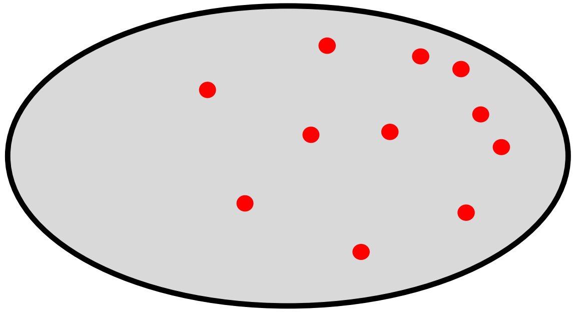


本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$
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



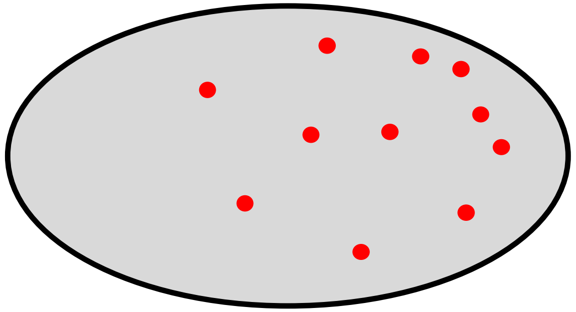
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$
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

$$\text{xBLEU} = \sum_{i} \sum_{e \in E(f_i)} \text{sBLEU}(e, e_i)\, p(e|f_i) \; = \; 0.5$$

Expected BLEU Training



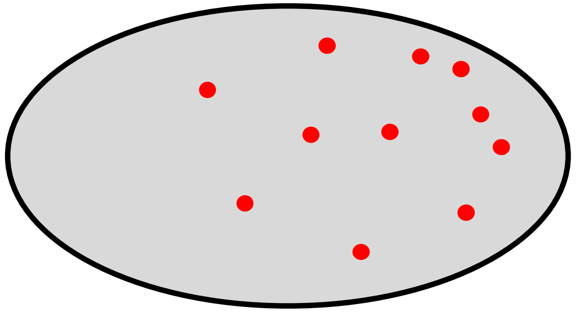
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$	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

$$\text{xBLEU} = \sum_{i} \sum_{e \in E(f_i)} \text{sBLEU}(e, e_i)\, p(e|f_i) = \boxed{0.5}$$

Expected BLEU Training



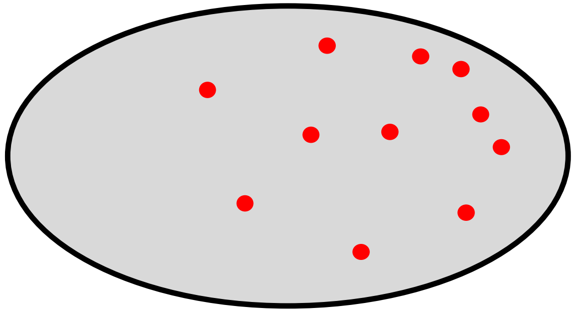
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$	δ_t	$p_{t+1}(e\|f_i)$
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	0.1

$$\text{xBLEU} = \sum_{i} \sum_{e \in E(f_i)} \text{sBLEU}(e, e_i)\, p(e|f_i) \;=\; 0.5$$

Expected BLEU Training



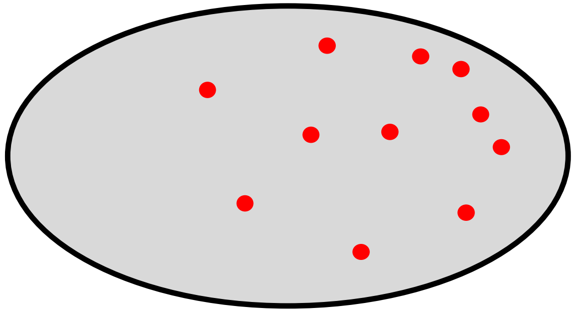
本 地区 的 发展 和 进步

Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$	δ_t	$p_{t+1}\ (e\|f_i)$
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	0.1

$$\text{xBLEU} = \sum_{i} \sum_{e \in E(f_i)} \text{sBLEU}(e, e_i)\, p(e|f_i) \ = \ 0.5$$

Expected BLEU Training



本 地区 的 发展 和 进步

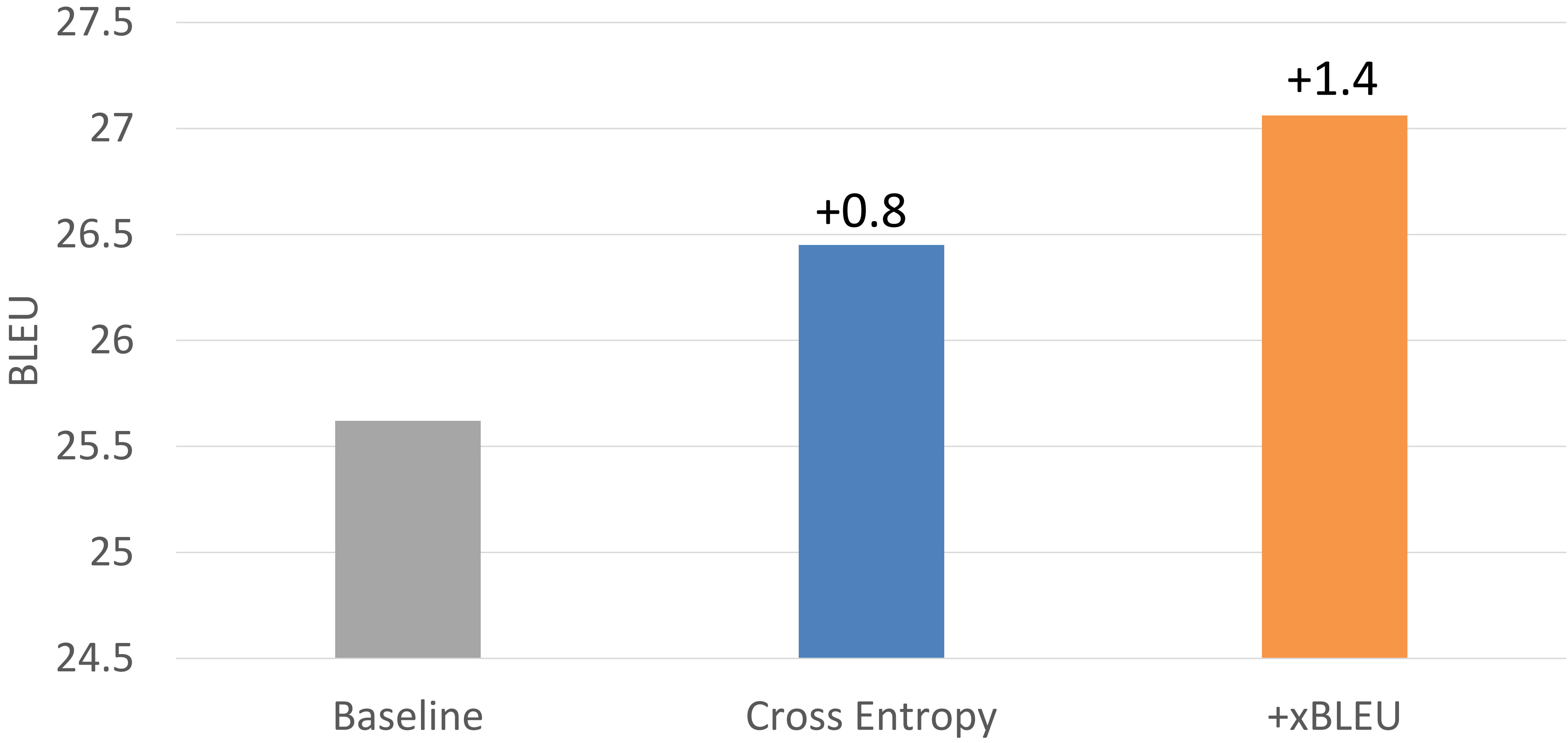
Human: development and progress of the region

	sBLEU	$p_t(e\|f_i)$	δ_t	$p_{t+1}(e\|f_i)$
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	0.1

$$\mathrm{xBLEU} = \sum_{i} \sum_{e \in E(f_i)} \mathrm{sBLEU}(e, e_i)\, p(e|f_i) = 0.5 \rightarrow \boxed{\mathbf{0.6}}$$

Results

Results


BLEU
27.5
27
26.5
26
25.5
25
24.5
+0.8
+1.4
Baseline
Cross Entropy
+xBLEU

Overview

- Recurrent neural network joint models (EMNLP 2013)
  Combined language and translation modeling

- Minimum translation modeling with recurrent nets (EACL 2014)
  Sequence models over bilingual units

- Training recurrent nets (ACL 2014)
  Expected BLEU training for neural network translation models

- Large-scale discriminative training for SMT (Auli et al., in submission)
  Training millions of linear ordering features with expected BLEU

Large-scale Discriminative Training for SMT

- Tuning: Minimum Error Rate Training: ~30 features (Och, 2003)

Large-scale Discriminative Training for SMT

- Tuning: Minimum Error Rate Training: ~30 features (Och, 2003)
- Perceptron (Liang, 2006), Max-violation perceptron (Yu et al., 2013)

Large-scale Discriminative Training for SMT

- Tuning: Minimum Error Rate Training: ~30 features (Och, 2003)
- Perceptron (Liang, 2006), Max-violation perceptron (Yu et al., 2013)
- Several others: PRO (Hopkins, 2011), MIRA (Chiang 2009 Watanabe 2007)

Large-scale Discriminative Training for SMT

- Tuning: Minimum Error Rate Training: ~30 features (Och, 2003)
- Perceptron (Liang, 2006), Max-violation perceptron (Yu et al., 2013)
- Several others: PRO (Hopkins, 2011), MIRA (Chiang 2009 Watanabe 2007)
- Recent success: MIRA-trained sparse ordering models (Cherry, 2013)

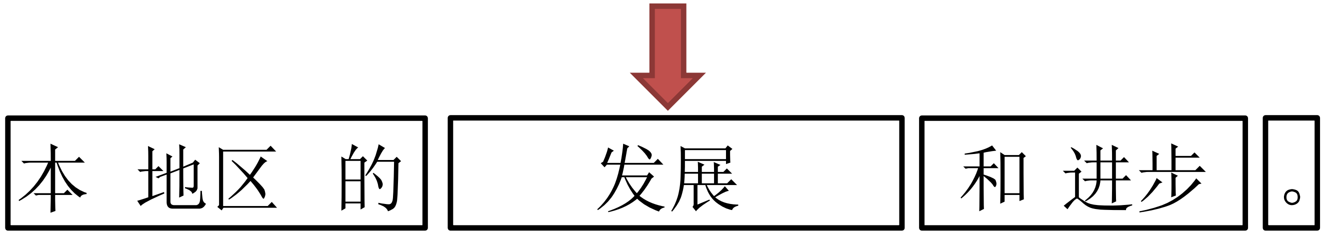
Large-scale Discriminative Training for SMT

- Tuning: Minimum Error Rate Training: ~30 features (Och, 2003)
- Perceptron (Liang, 2006), Max-violation perceptron (Yu et al., 2013)
- Several others: PRO (Hopkins, 2011), MIRA (Chiang 2009 Watanabe 2007)
- Recent success: MIRA-trained sparse ordering models (Cherry, 2013)
- **Next:** Training large-scale sparse ordering models with expected BLEU

Lexicalized Reordering

本 地区 的 | 发展 | 和 进步 | 。

Lexicalized Reordering



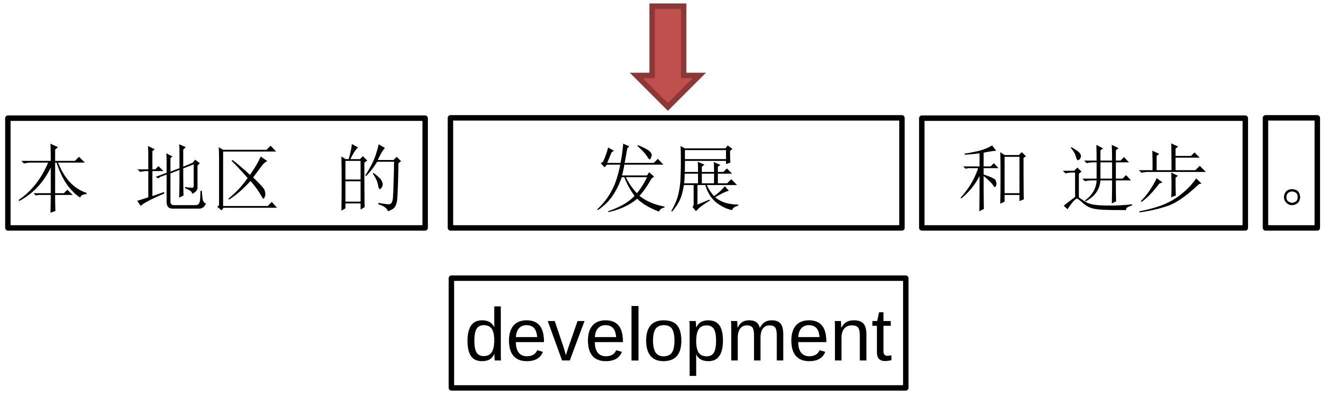
本 地区 的

发展

和 进步

。

Lexicalized Reordering



本 地区 的

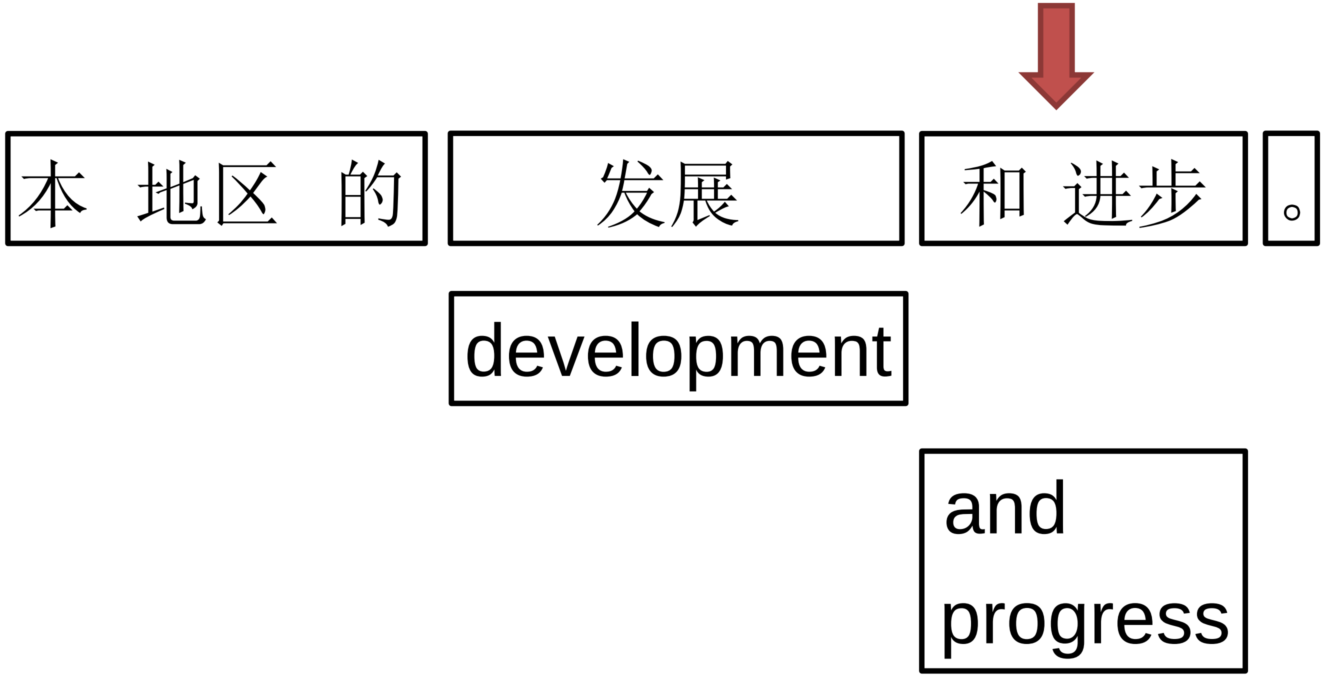
发展

和 进步

。

development

Lexicalized Reordering



本 地区 的

发展

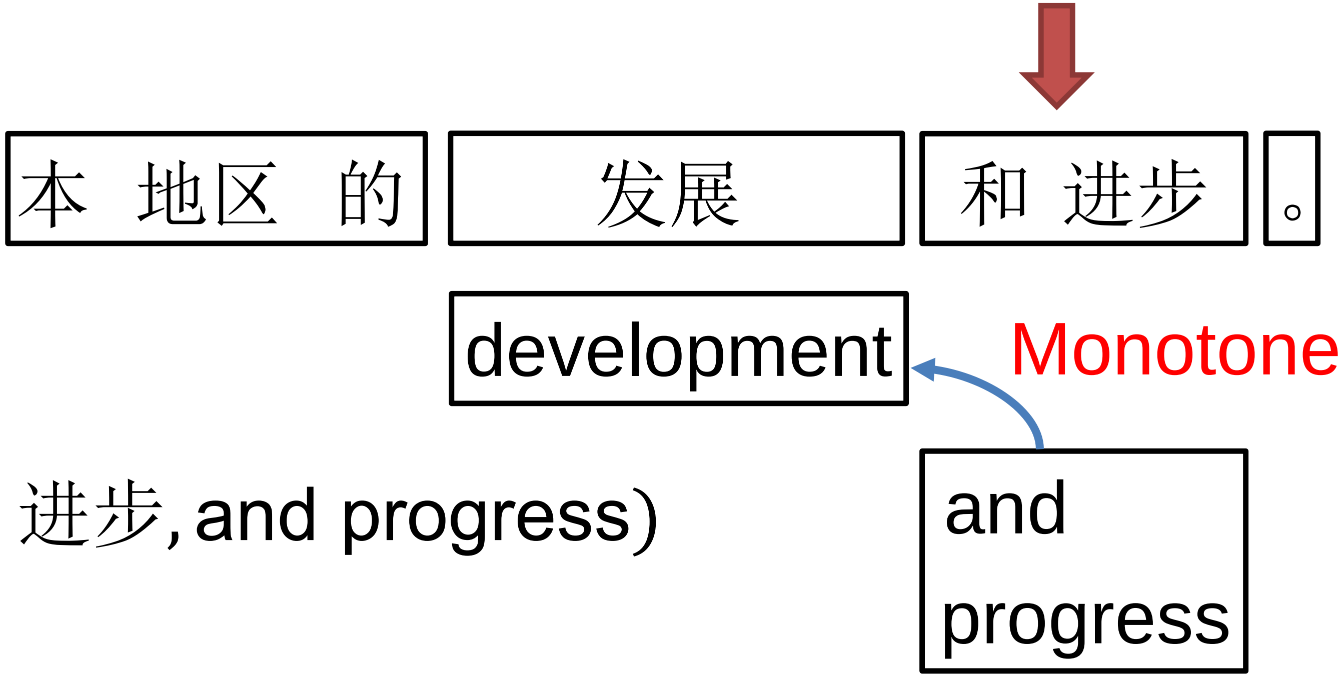
和 进步

。

development

and progress

Lexicalized Reordering



本 地区 的

发展

和 进步

。

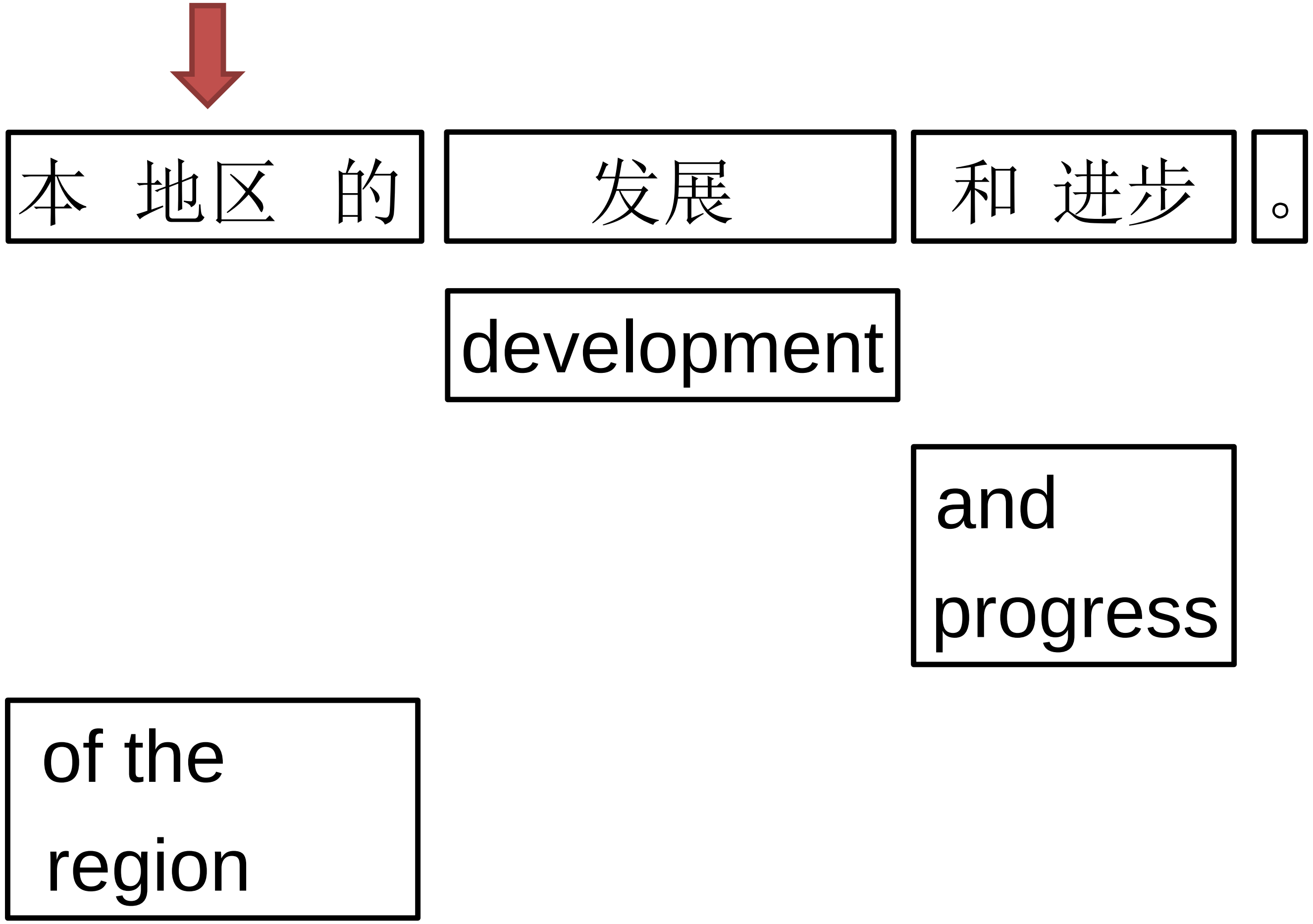
development

Monotone

and progress

p(Monotone|和 进步, and progress)

Lexicalized Reordering



本 地区 的

发展

和 进步

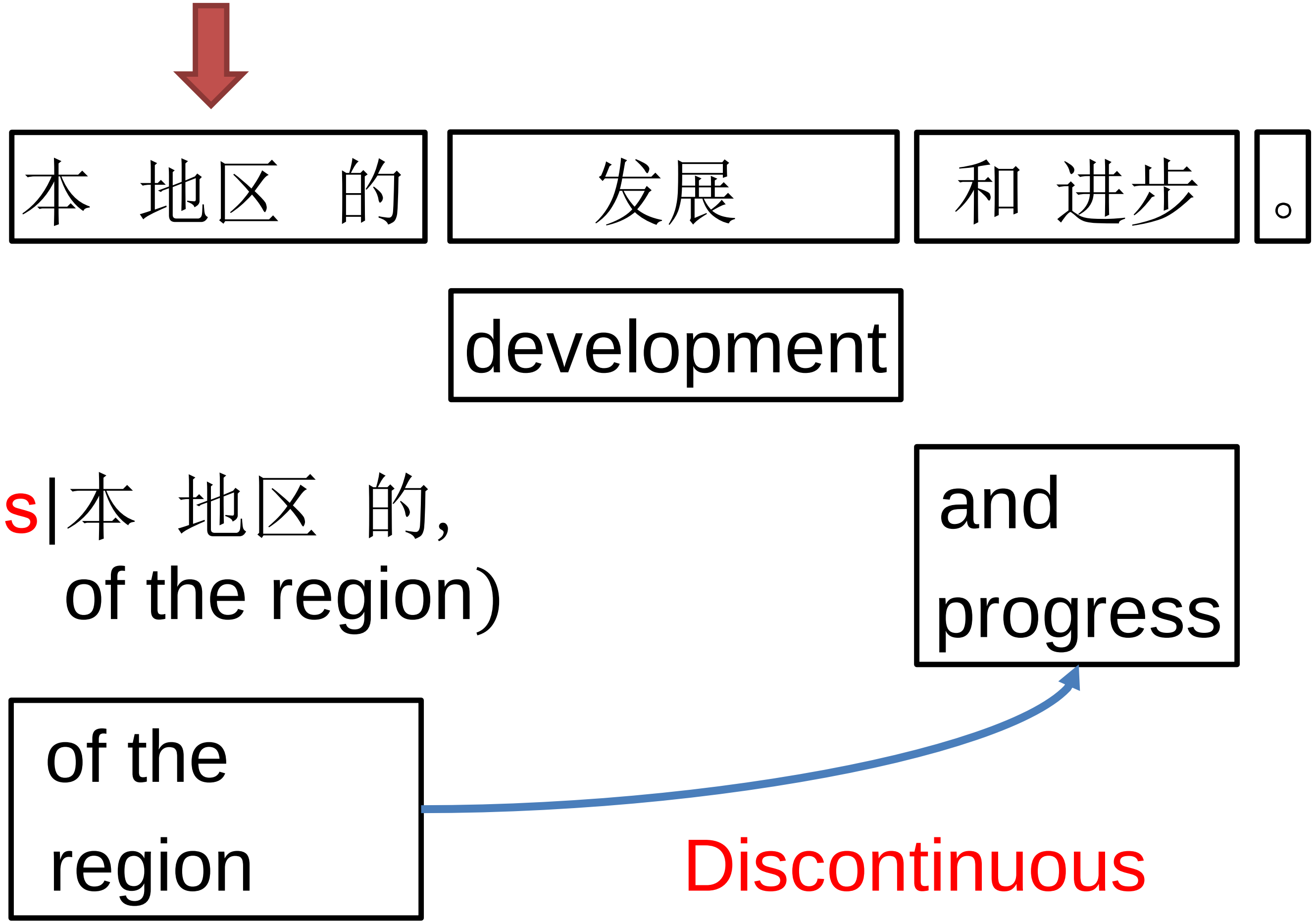
。

development

and progress

of the region

Lexicalized Reordering



本 地区 的

发展

和 进步

。

development

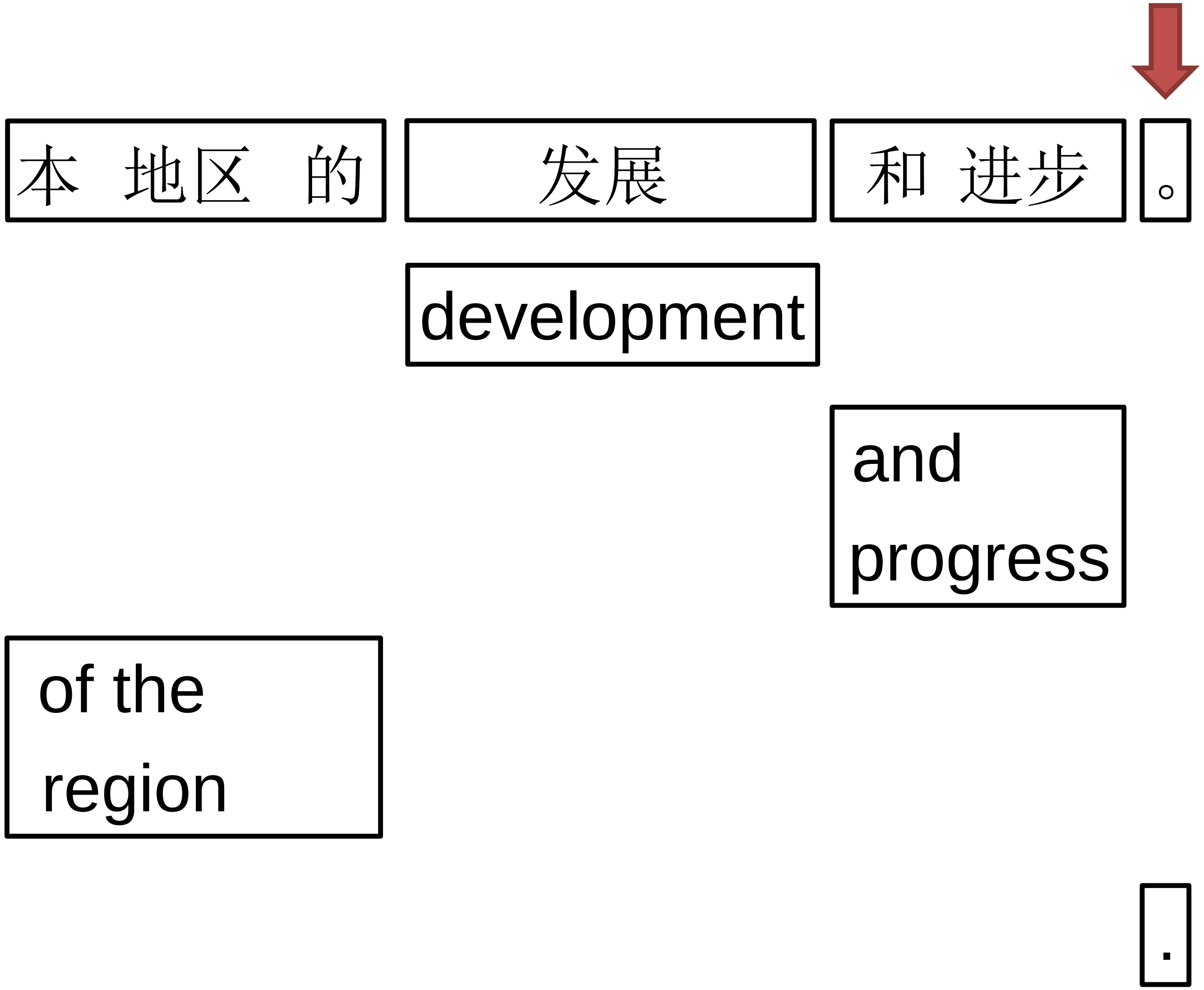
and progress

of the region

Discontinuous

p(Discontinuous|本 地区 的, of the region)

Lexicalized Reordering



本 地区 的

发展

和 进步

。

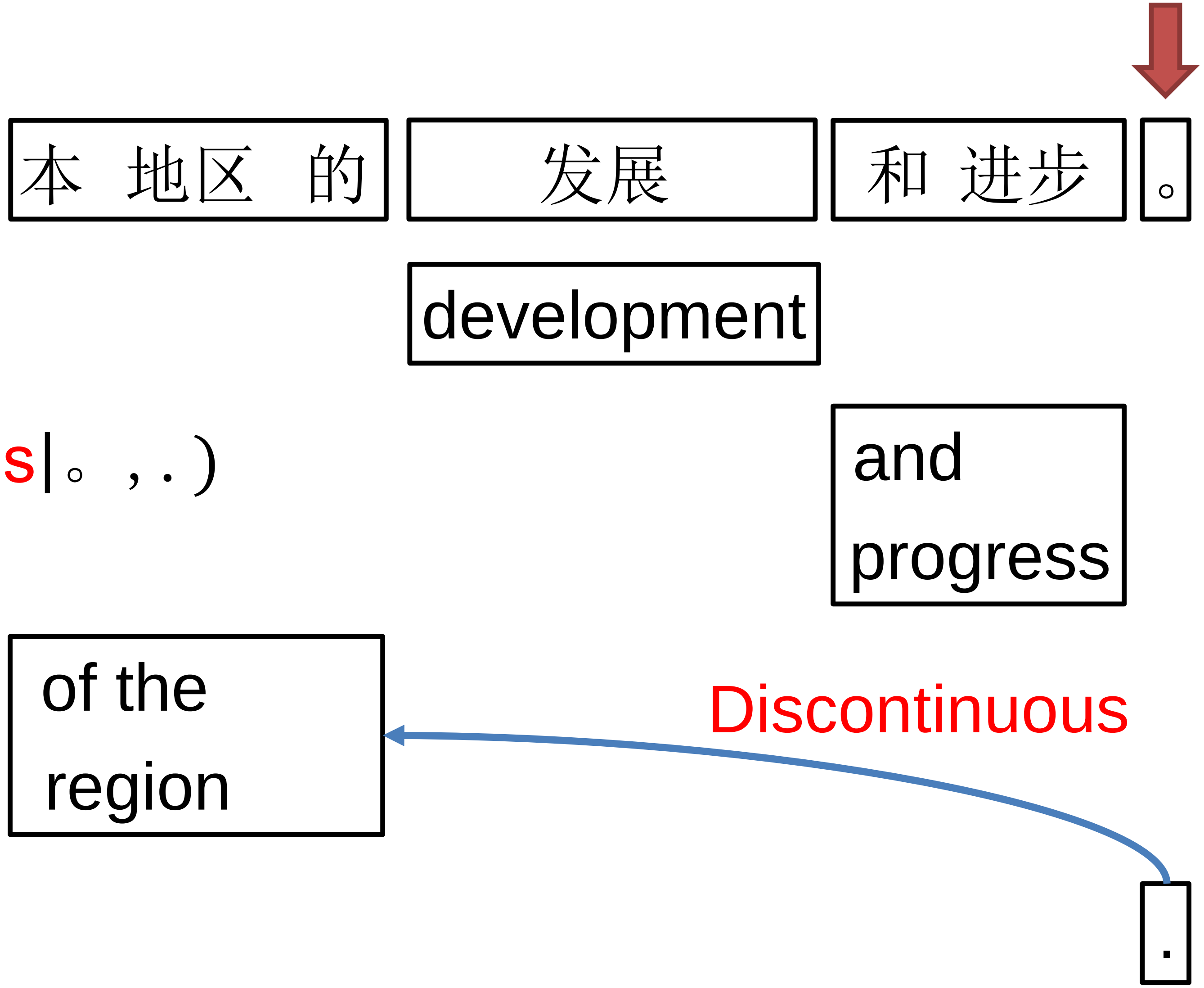
development

and progress

of the region

.

Lexicalized Reordering



本 地区 的

发展

和 进步

。

development

and progress

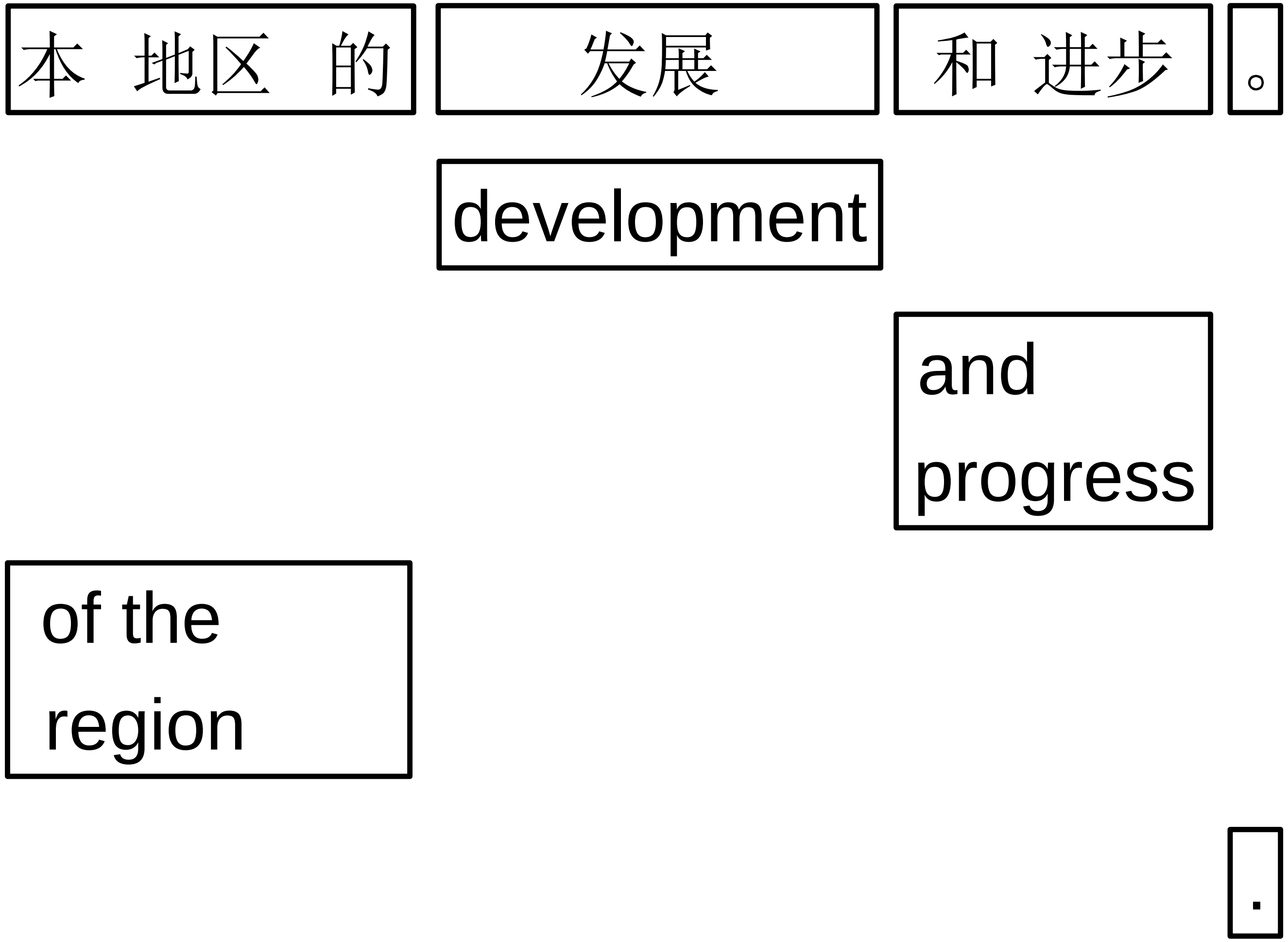
of the region

Discontinuous

.

p(Discontinuous|。, .)

Hierarchical Lexicalized Reordering



本 地区 的

发展

和 进步

。

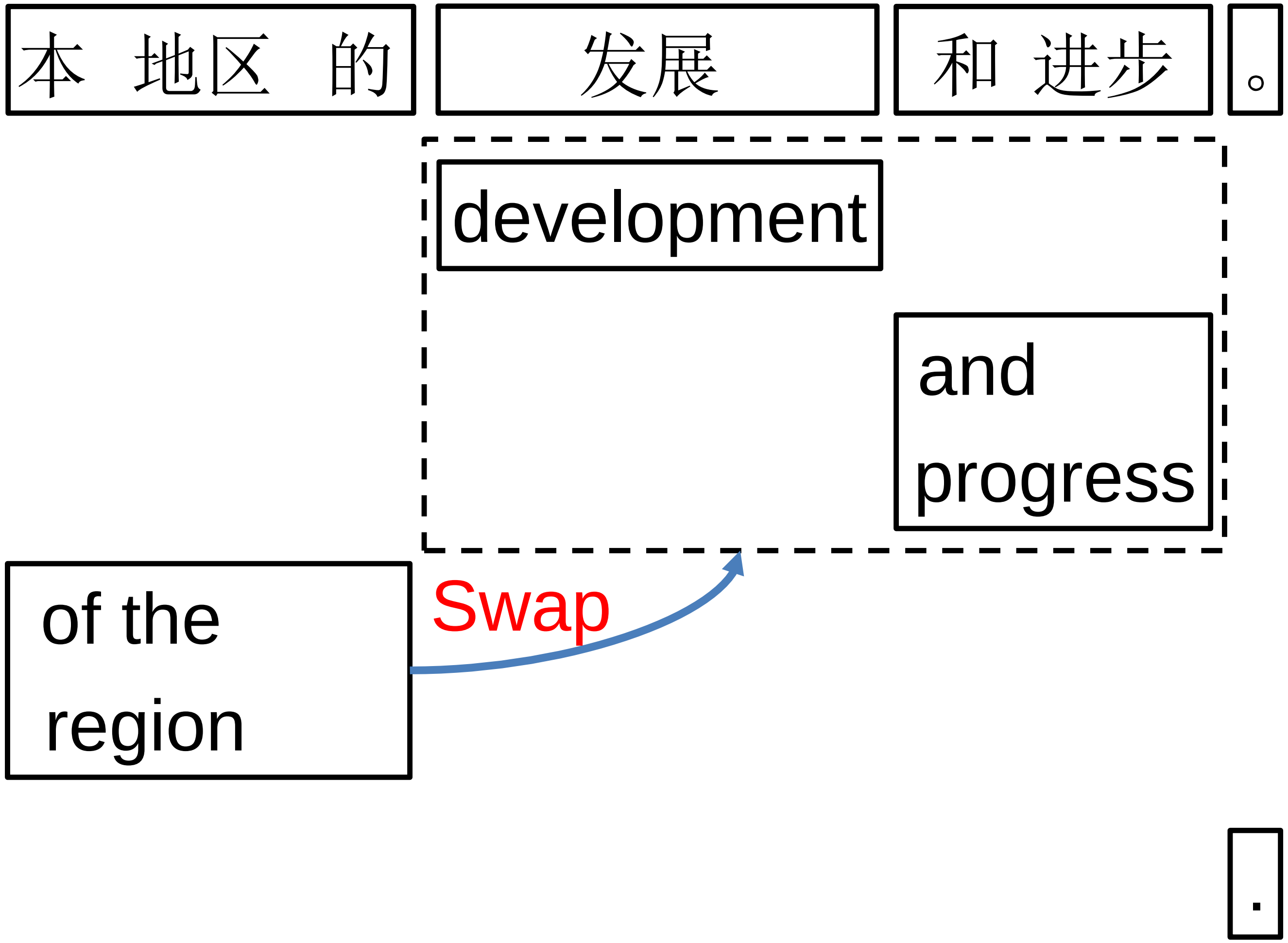
development

and progress

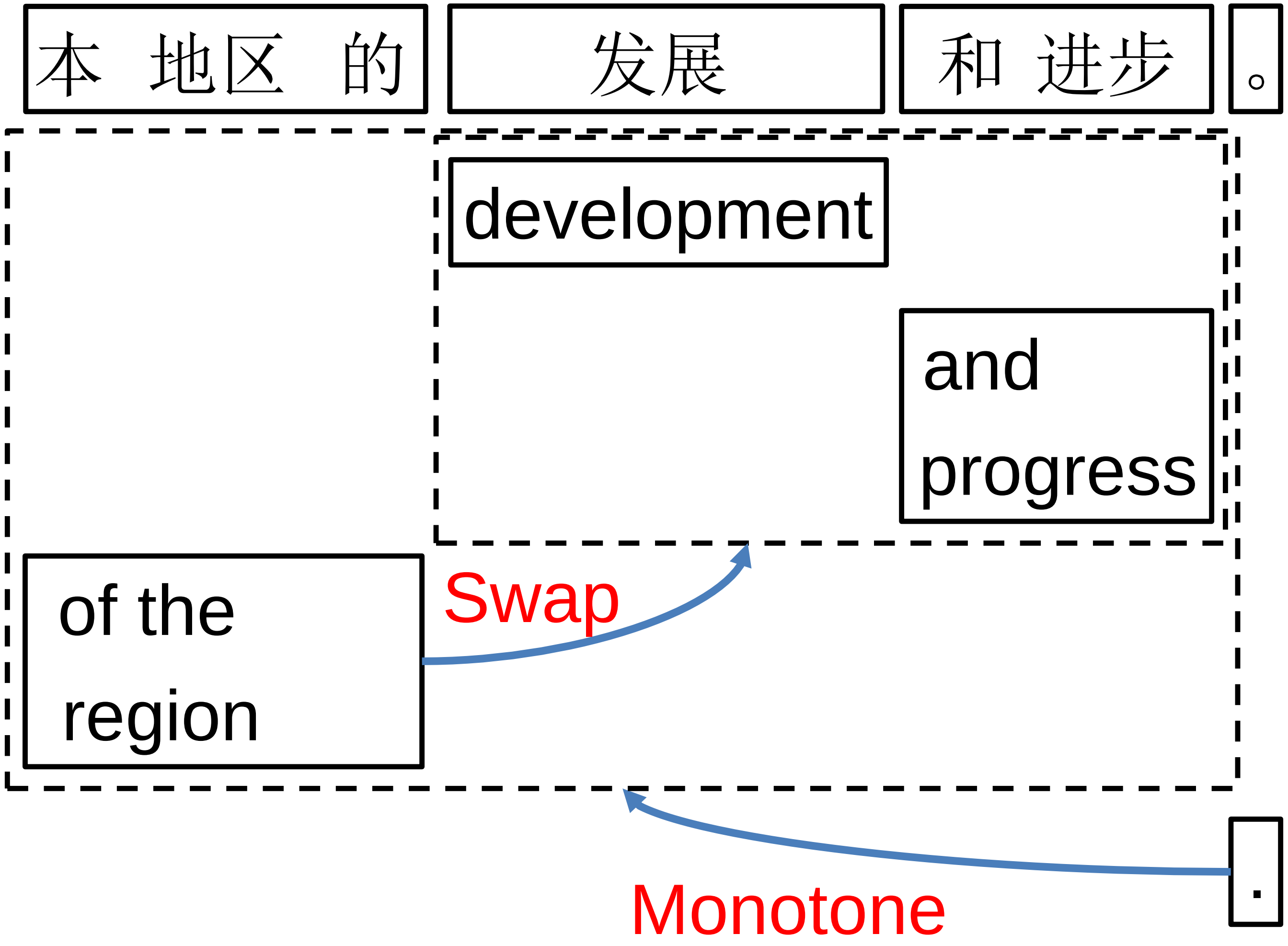
of the region

.

Hierarchical Lexicalized Reordering



Hierarchical Lexicalized Reordering



Hierarchical Lexicalized Reordering

p(Monotone|和 进步, and progress)

Hierarchical Lexicalized Reordering

$$\mathrm{p}(\mathrm{o}|pp) =$$

Hierarchical Lexicalized Reordering

$$\mathrm{p}(\mathrm{o}|pp) = \frac{\mathrm{count}(\mathrm{o}, pp)}{\mathrm{count}(pp)}$$

Hierarchical Lexicalized Reordering

$$\mathrm{p}(\mathrm{o}|pp) = \frac{\mathrm{count}(\mathrm{o}, pp)}{\mathrm{count}(pp)}$$

- Typically 100Ms of parameters

Hierarchical Lexicalized Reordering

$$\mathrm{p}(\mathrm{o}|pp) = \frac{\mathrm{count}(\mathrm{o}, pp)}{\mathrm{count}(pp)}$$

- Typically 100Ms of parameters
- Very sparse estimates

Hierarchical Lexicalized Reordering

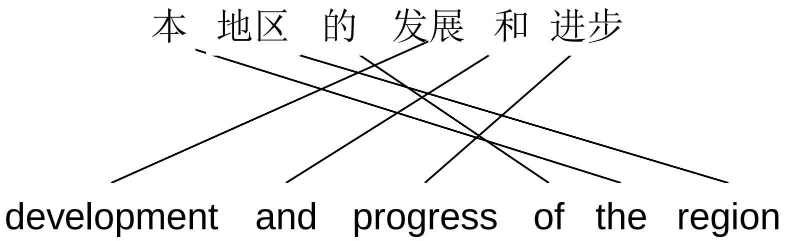
$$\mathrm{p}(\mathrm{o}|pp) = \frac{\mathrm{count}(\mathrm{o}, pp)}{\mathrm{count}(pp)}$$

- Typically 100Ms of parameters
- Very sparse estimates
- Objective: Likelihood

Hierarchical Lexicalized Reordering

$$\mathrm{p}(\mathrm{o}|pp) = \frac{\mathrm{count}(\mathrm{o}, pp)}{\mathrm{count}(pp)}$$

- Typically 100Ms of parameters
- Very sparse estimates
- Objective: Likelihood
- Training data: word-aligned bi-texts



MaxEnt Reordering (Xiong 2006, Nguyen 2009)

$\mathrm{p}(\mathrm{o}|pp) =$ indicator features!

e.g. Monotone_progress,
Monotone_和

MaxEnt Reordering (Xiong 2006, Nguyen 2009)

$$\mathrm{p}(\mathrm{o}|pp) = \frac{\exp\{\theta^T h(\mathrm{o}, \mathrm{pp})\}}{\sum_o \exp\{\theta^T h(o, \mathrm{pp})\}}$$

e.g. Monotone_progress,
Monotone_和

MaxEnt Reordering (Xiong 2006, Nguyen 2009)

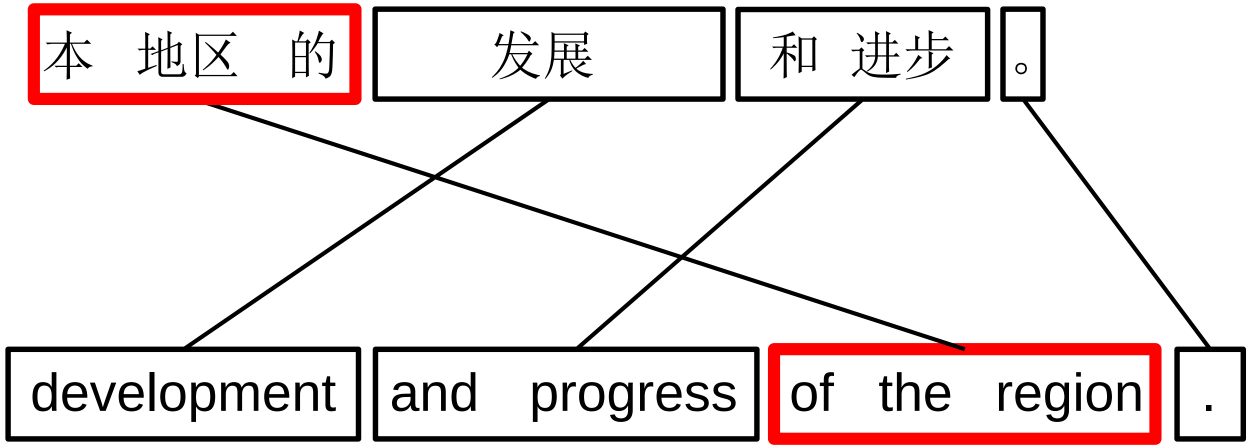
$$\mathrm{p}(\mathrm{o}|pp) = \frac{\exp\{\theta^T h(\mathrm{o}, \mathrm{pp})\}}{\sum_o \exp\{\theta^T h(o, \mathrm{pp})\}}$$

e.g. Monotone_progress,
Monotone_和

- Typically Ms of parameters
- *Better estimates*
- Objective: Likelihood
- Training data: word-aligned bi-texts

Sparse Reordering (Cherry 2013)

- Simple unigram features
- Most frequent 80 words, 20 or 50 class Brown Clusters
  e.g., Monotone_the, Monotone_C20, Monotone_C50
- About 3.5K features



- Discontinuous_src_本
- Discontinuous_tgt_of
- Discontinuous_src_C20
- …

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

- *Better estimates*

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

h_1: p(e|f)
h_2: p(f|e)
h_3: $p_{LM}(e)$

- *Better estimates*

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

- *Better estimates*

h_1: p(e|f)
h_2: p(f|e)
h_3: $p_{LM}(e)$
h_4: c(Monotone_progress)
h_5: c(Monotone_和)
…

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

h_1: p(e|f)
h_2: p(f|e)
h_3: $p_{LM}(e)$
h_4: c(Monotone_progress)
h_5: c(Monotone_和)
…

- *Better estimates*
- Objective: **BLEU**

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

h_1: p(e|f)
h_2: p(f|e)
h_3: $p_{LM}(e)$
h_4: c(Monotone_progress)
h_5: c(Monotone_和)
…

- *Better estimates*
- Objective: **BLEU**
- Training data: **machine translation output**

Sparse Reordering (Cherry 2013)

Idea: Add ordering features to top-level features and tune with MIRA

$$\hat{e} = \mathrm{argmax}_e \theta^T h(f, e)$$

h_1: p(e|f)
h_2: p(f|e)
h_3: $p_{LM}(e)$
h_4: c(Monotone_progress)
h_5: c(Monotone_和)
…

- *Better estimates*
- Objective: **BLEU**
- Training data: **machine translation output**
- Much better than MaxEnt

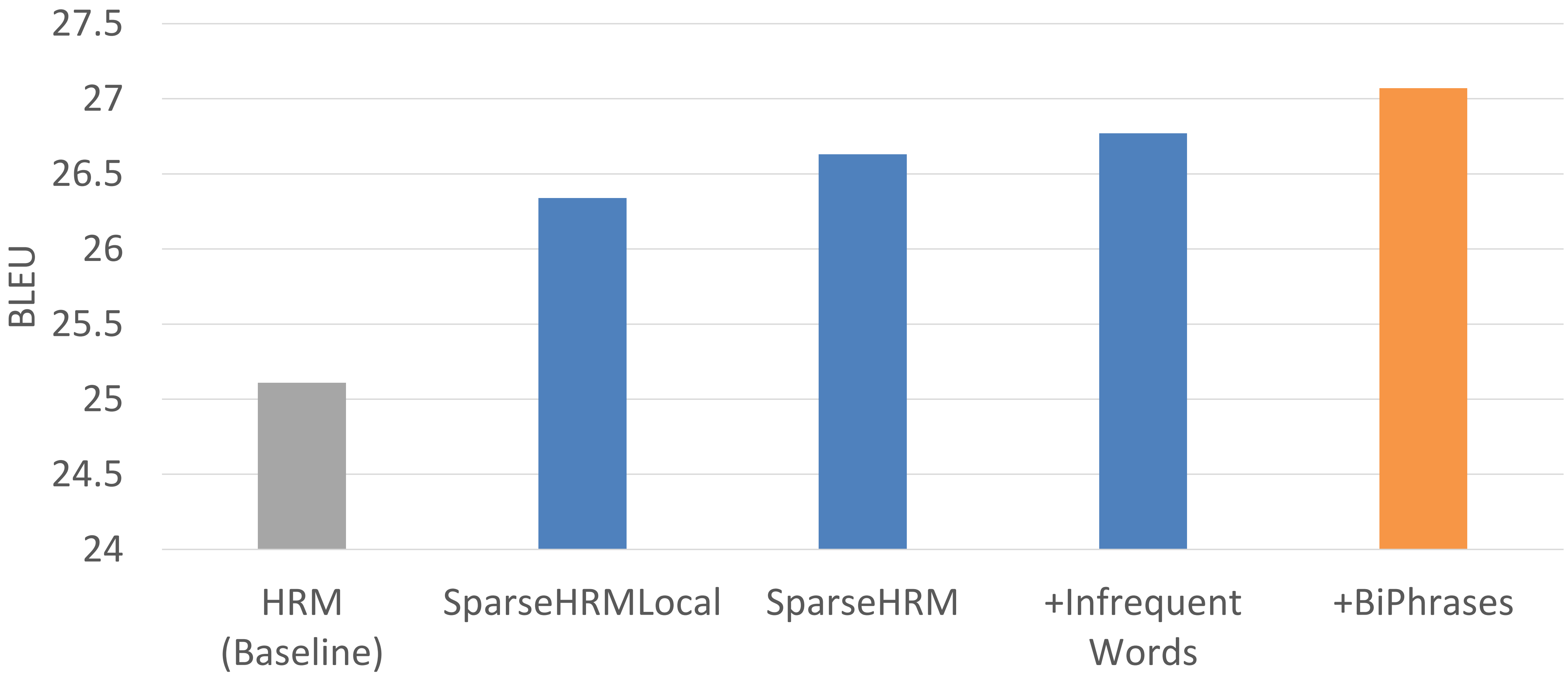
Sparse Reordering (Cherry 2013)

- Lexicalized models trained on Ms of sentences with 100Ms of parameters
- Cherry (2013): Ordering model with 3.5K features learned on 2K sentences
- Can we learn a general purpose ordering model this way?
- MIRA/PRO don’t scale to truly large settings (Yu 2013, Eidelman 2013)
- **Next:** Large-scale discriminative models with Ms of features trained on 100Ks of sentences using expected BLEU

Scaling the feature set

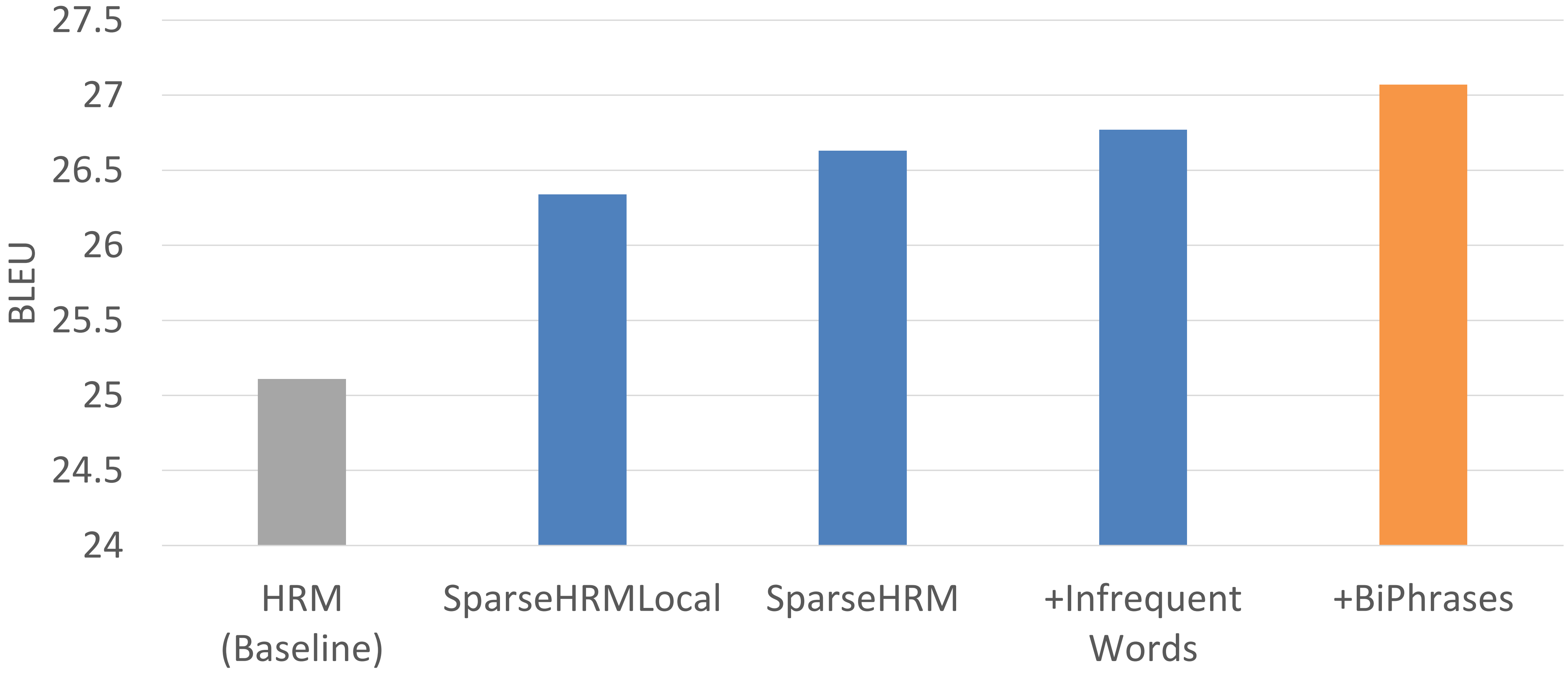
Features	4.5K	9K	900K	3M

Scaling the feature set



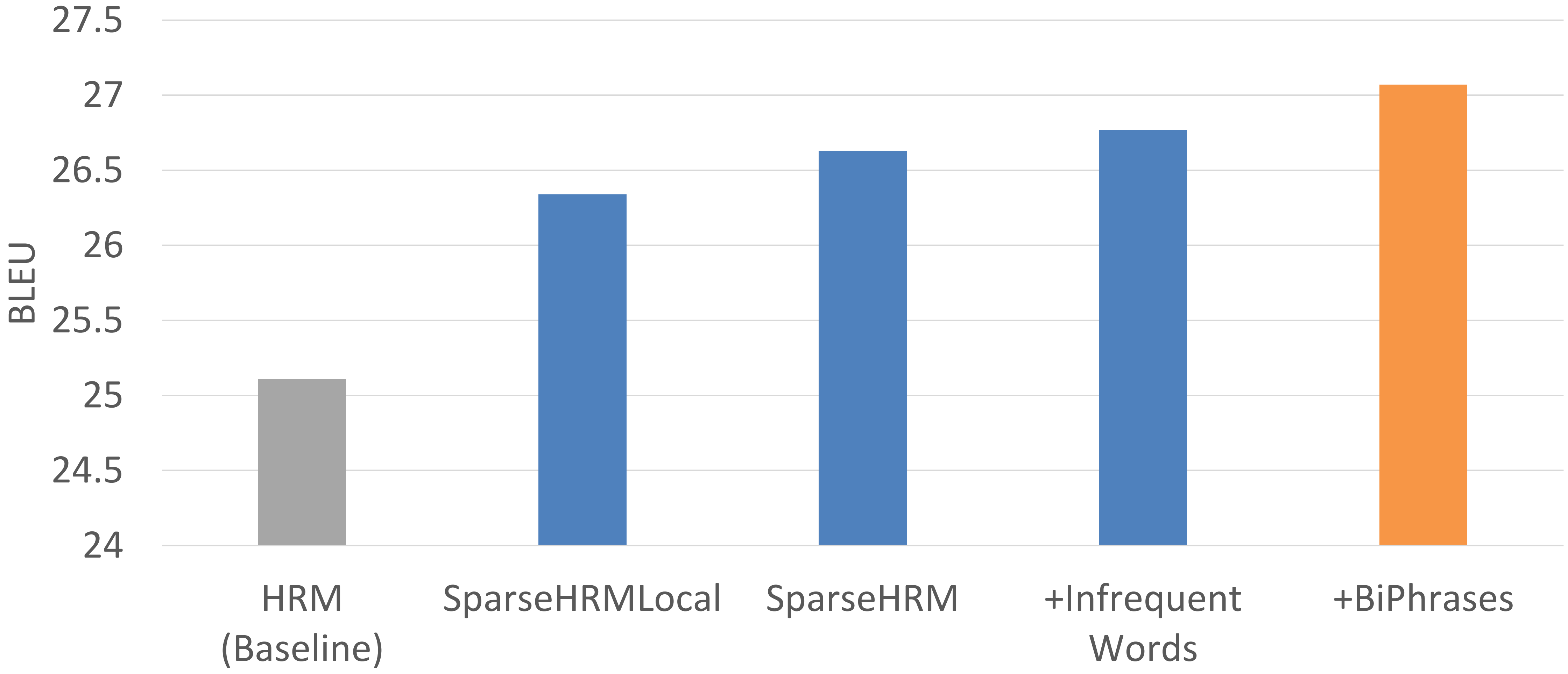
Features 4.5K 9K 900K 3M

Scaling the feature set



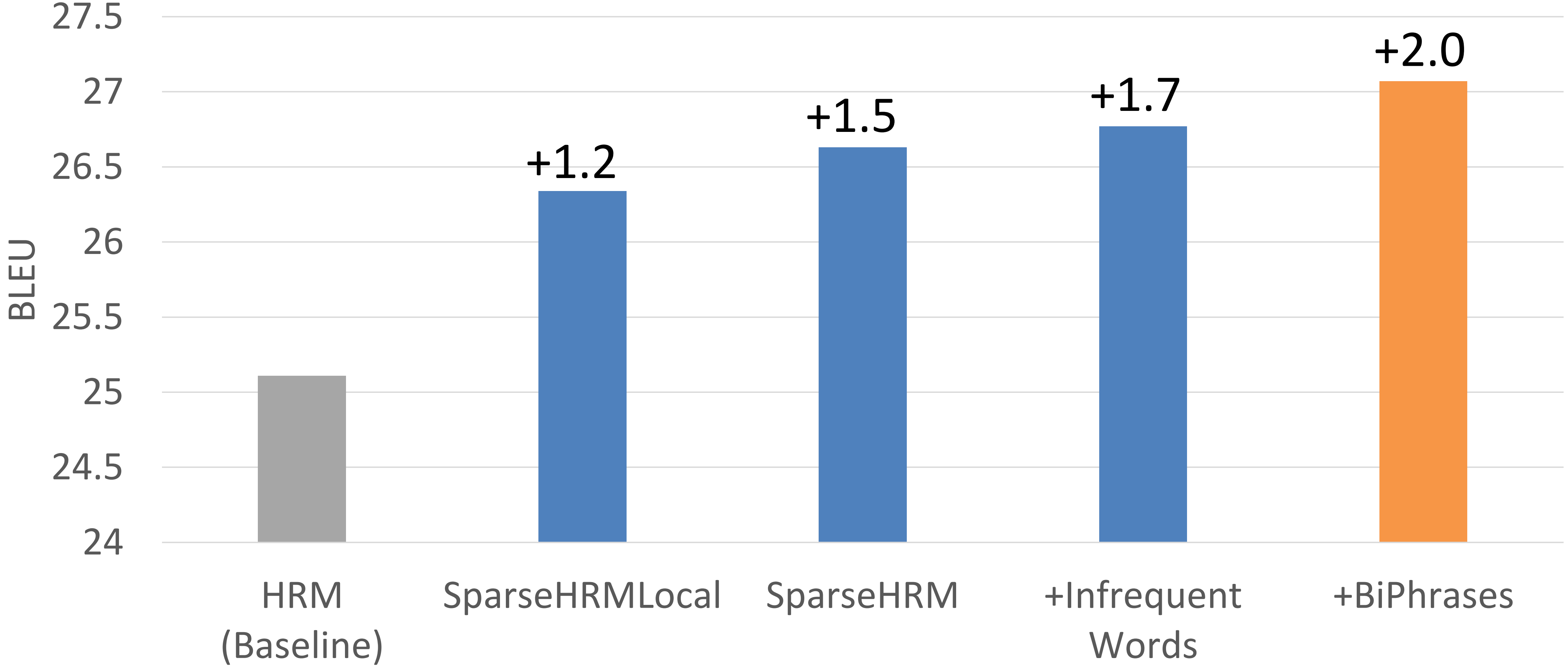
Features	4.5K	9K	900K	3M

Scaling the feature set



Features 4.5K 9K 900K 3M

Scaling the feature set



Features	4.5K	9K	900K	3M

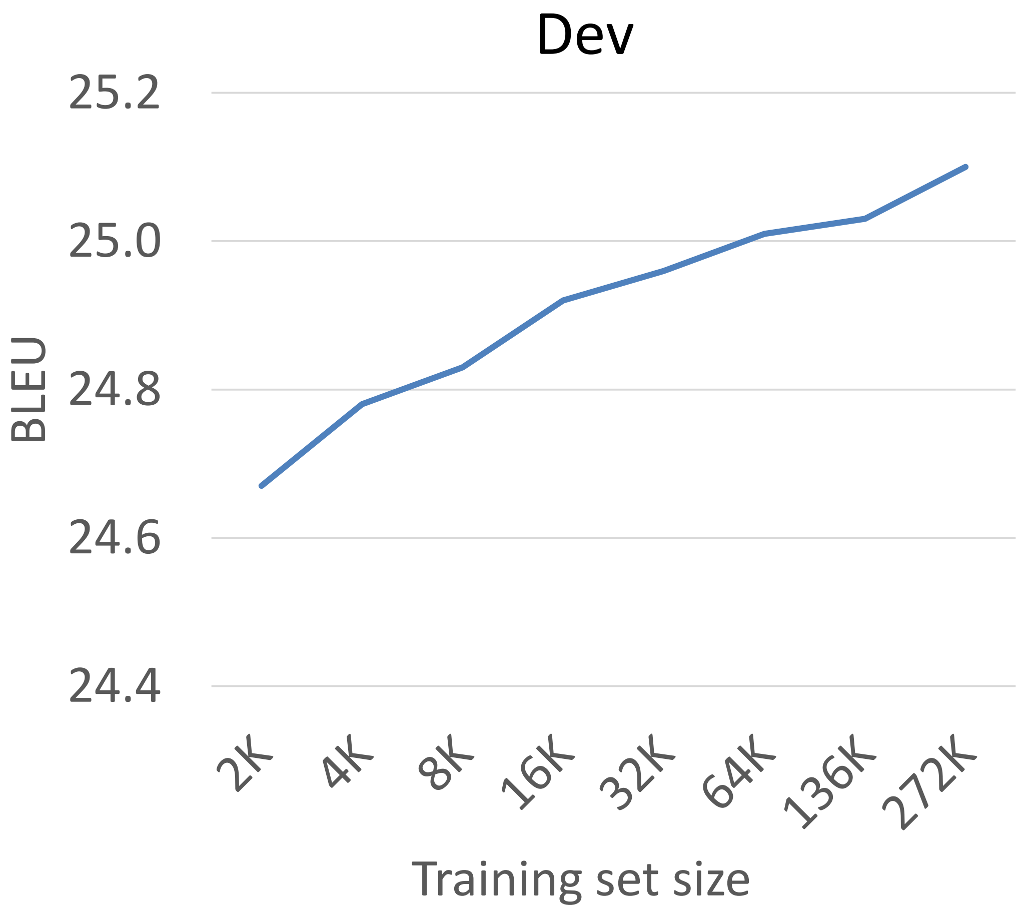
Scaling the training data

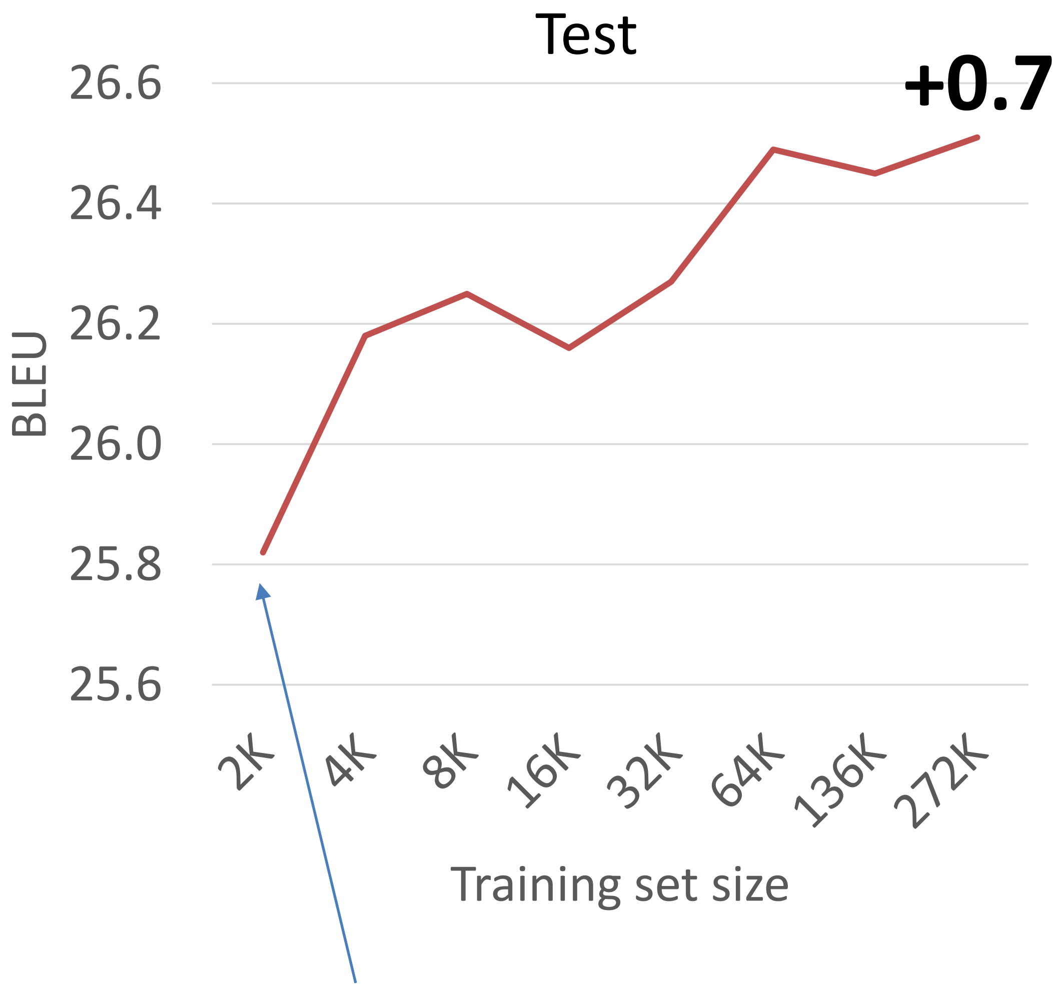
Dev

Test

N-best rescore with SparseHRMLocal (4.5K features)

Scaling the training data





N-best rescore with SparseHRMLocal (4.5K features)

Setup of Cherry (2013)

Why is this better than Lexicalized/Maxent models?

- Objective: Likelihood → BLEU
- Train data: bilingual corpus → machine translation output

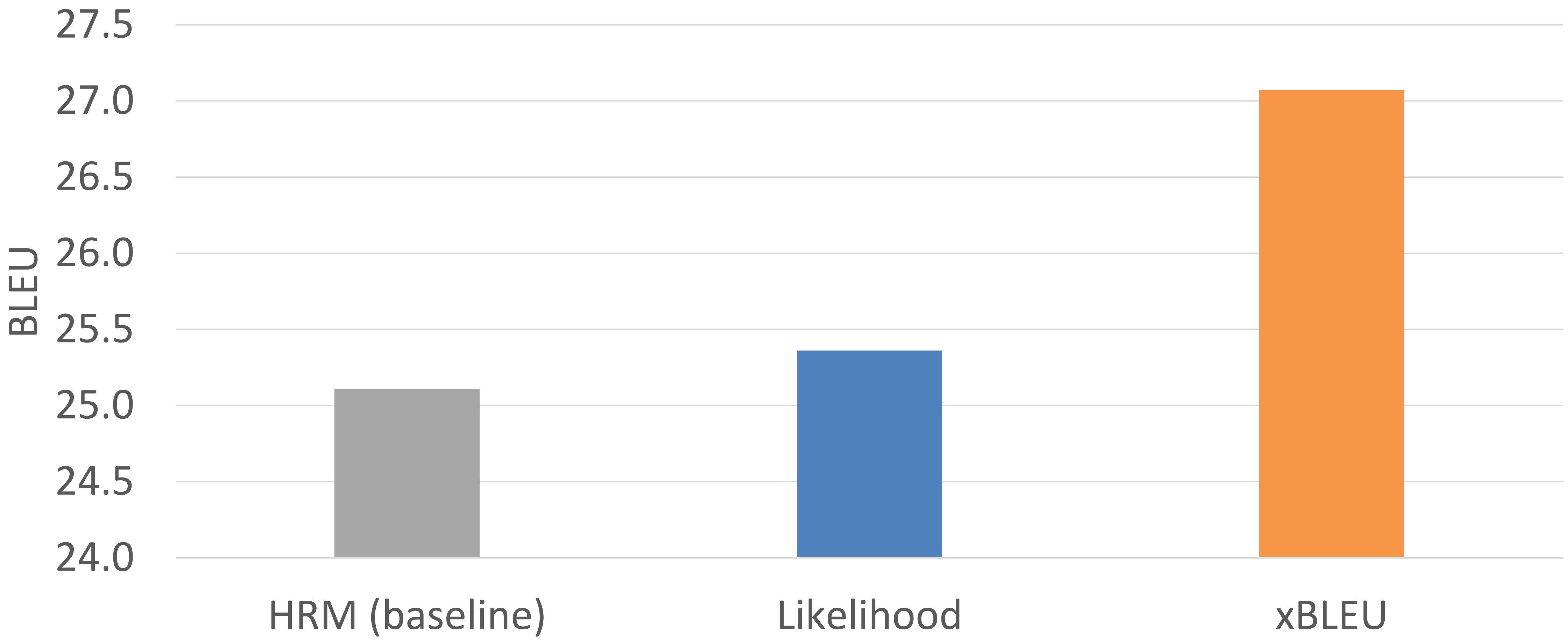
- Which one responsible for better performance?

- Experiment: Likelihood/xBLEU train on MT output

Likelihood vs. xBLEU

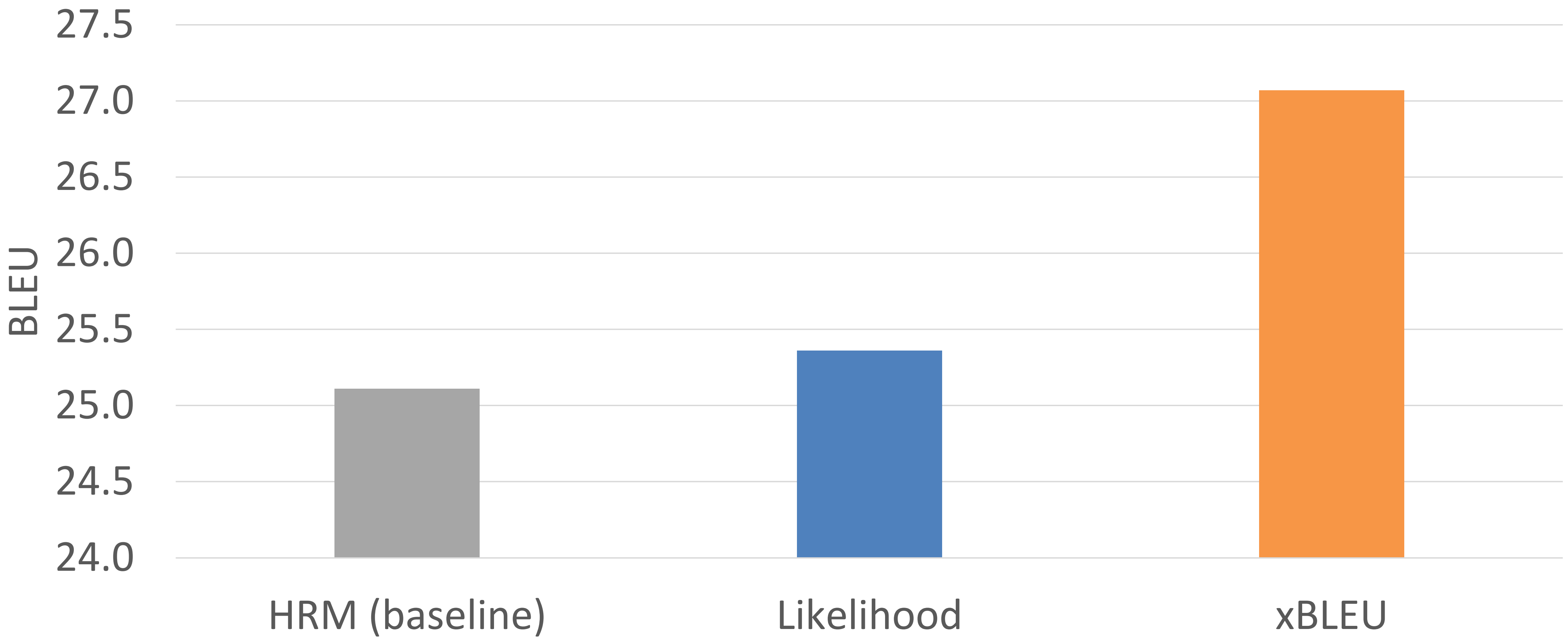
Based on BiPhrases (3M features)

Likelihood vs. xBLEU



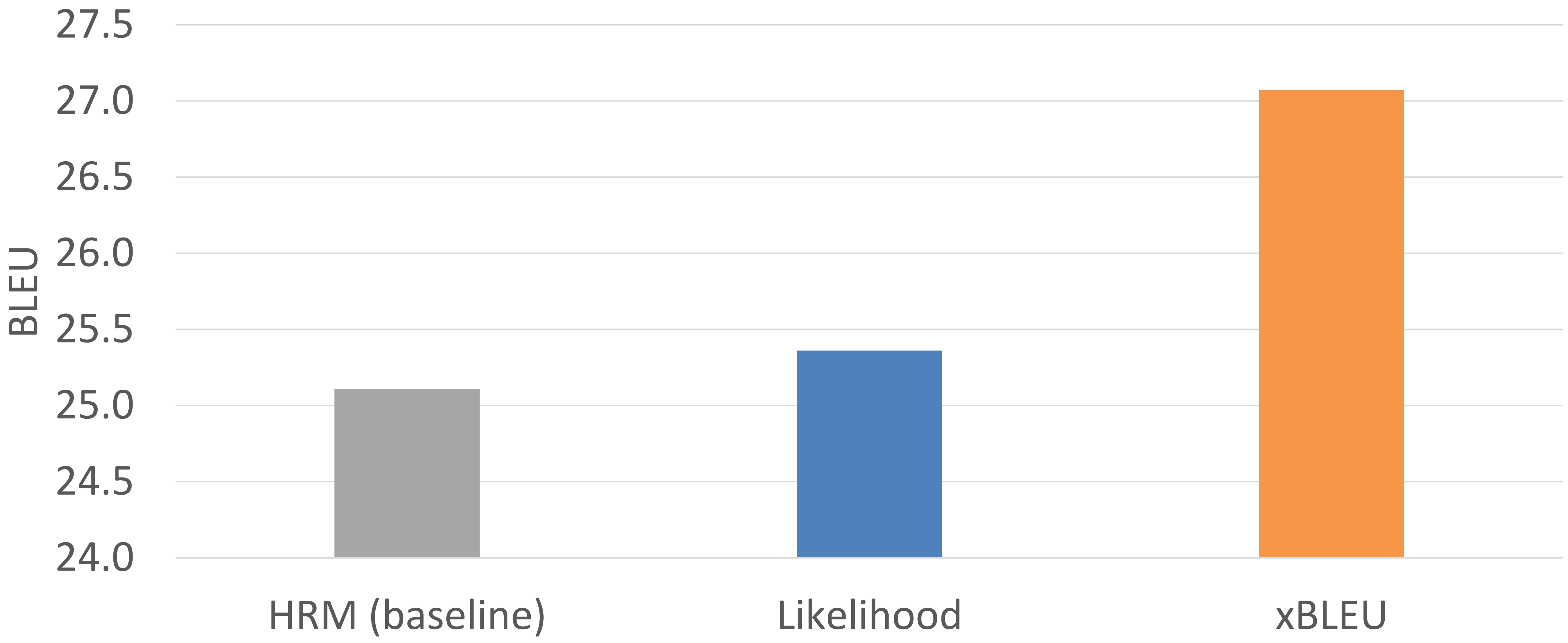
Based on BiPhrases (3M features)

Likelihood vs. xBLEU



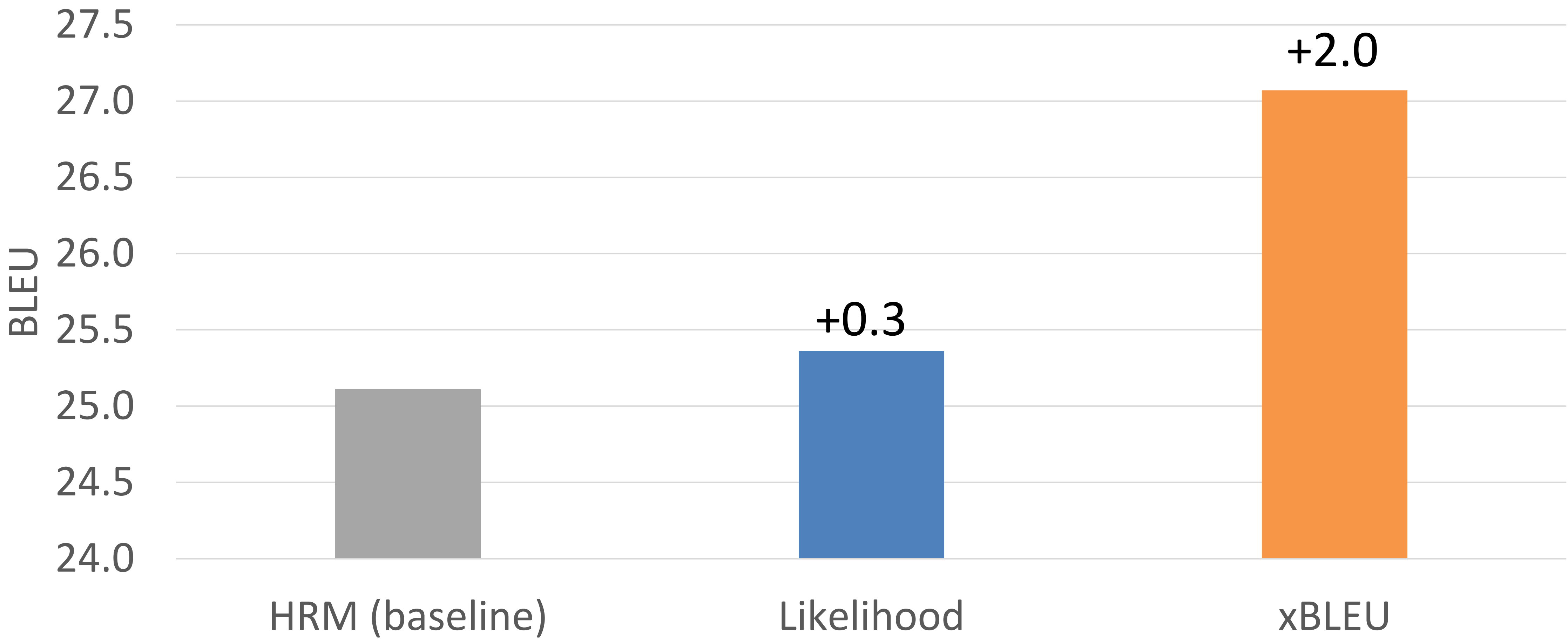
Based on BiPhrases (3M features)

Likelihood vs. xBLEU



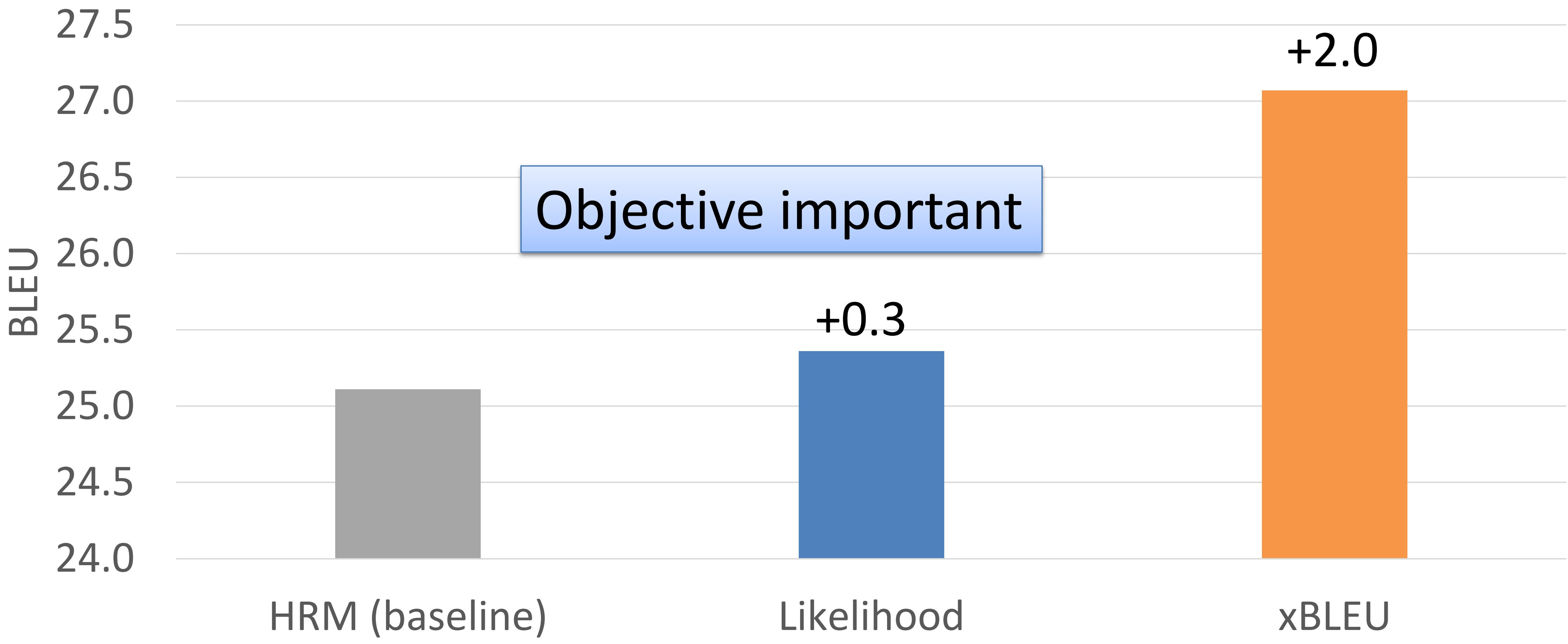
Based on BiPhrases (3M features)

Likelihood vs. xBLEU



Based on BiPhrases (3M features)

Likelihood vs. xBLEU



Objective important

Based on BiPhrases (3M features)

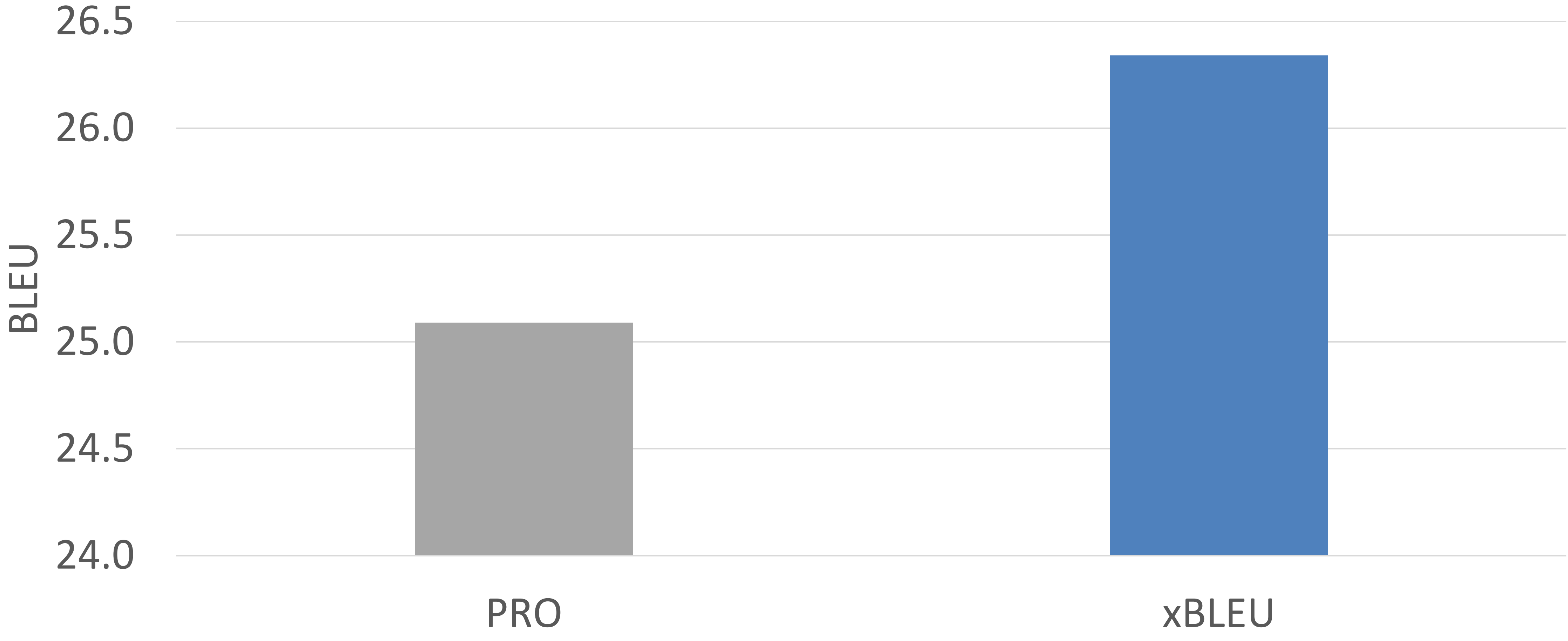
xBLEU vs. PRO

Based on SparseHRMLocal (4.5K features)

Train Data: 2.5K 140K

xBLEU vs. PRO

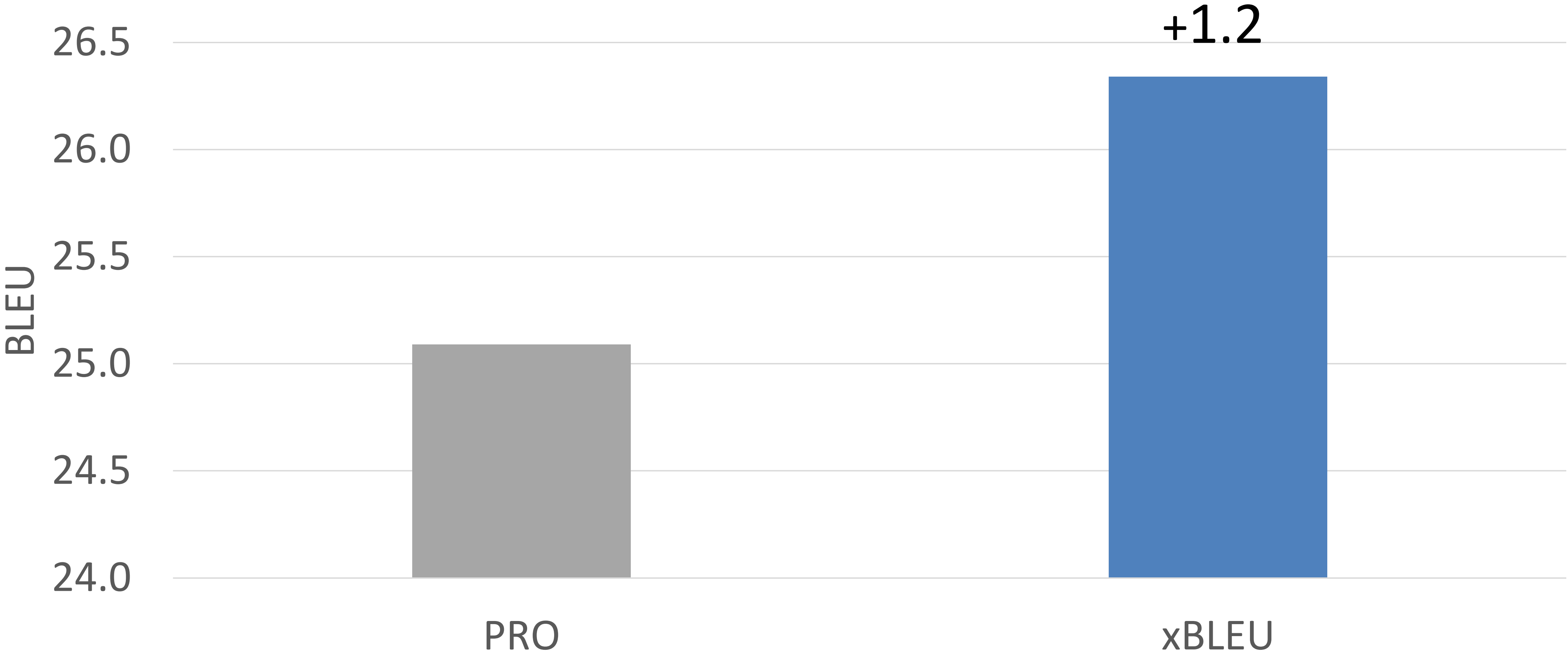
Based on SparseHRMLocal (4.5K features)



Train Data: 2.5K 140K

xBLEU vs. PRO

Based on SparseHRMLocal (4.5K features)



Train Data: 2.5K 140K

Comparison to Max-Violation Perceptron

	Max-Violation Perceptron	xBLEU

Comparison to Max-Violation Perceptron

	Max-Violation Perceptron	**xBLEU**
Loss	No partial credit (0/1)	**partial credit**

Comparison to Max-Violation Perceptron

	Max-Violation Perceptron	**xBLEU**
Loss	No partial credit (0/1)	**partial credit**
Train data	Mostly short sentences (reference must be reachable)	Uses **all data**

Comparison to Max-Violation Perceptron

	Max-Violation Perceptron	**xBLEU**
Loss	No partial credit (0/1)	**partial credit**
Train data	Mostly short sentences (reference must be reachable)	Uses **all data**
Updates	Based 1-best and reference	Based on **all outputs** in gen-set

Summary

- Directly optimizing sub-models towards BLEU improves translation accuracy
- xBLEU allows estimation of millions of features
- More training data helps
- Objective crucial to good performance

Conclusion

- Recurrent nets are very well suited to model translation
- They complement and improve simpler models
- xBLEU training effective for both neural nets and linear models
- xBLEU scales to millions of features on hundreds of thousands of sentences

Future Directions

What can we do with the presented methods?

- LSTM nets for translation
- Recurrent nets for other NLP tasks, e.g., CCG parsing
- xBLEU training: Large-scale discriminative training of all models

Thank you!