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)
Machine translation

The development and progress of the region.
Machine translation

Translation modeling

发展和进步。  

development and progress of the region.
Machine translation

本地地区的发展和进步。

Development and progress of the region.

Translation modeling

Language modeling
Machine translation

Translation modeling

Language modeling

Optimization

development and progress of the region.

本地地区的 发展和进步。
Machine translation

Translation modeling
Language modeling
Optimization
Reordering

本地区的发展和进步。

development and progress of the region.
Machine translation

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
Discrete phrase-based translation
Discrete phrase-based translation

development and progress of the region.
Discrete phrase-based translation

Koehn et al. (2003)
Discrete phrase-based translation

Koehn et al. (2003)
Discrete phrase-based translation

Koehn et al. (2003)

Phrase Length

Train

Test

0% 1 2 3 4 5 6 7

Development and progress of the region.

本地区的发展和进步。
Discrete phrase-based translation

Koehn et al. (2003)

The region of development and progress.

Phrase Length

0% 23% 47% 70%

Phrase Length

Train Test

1 2 3 4 5 6 7
Discrete phrase-based translation

Development and progress of the region.

Koehn et al. (2003)

Average in train: 2.3 words
Average in test: 1.5 words
Discrete n-gram language modeling

\[ p(\text{progress in the region}) = \]
Discrete n-gram language modeling

\[ p(\text{progress in the region}) = \]

Train data:

... development and progress of the region. in ...

...
Discrete n-gram language modeling

\[ p(\text{progress in the region}) = p(\text{progress}) \ p(\text{in}) \ p(\text{the}) \ p(\text{region | the}) \]

Train data:

... development and progress of the region. in ...

...
Discrete n-gram language modeling

\[ p(\text{progress in the region}) = p(\text{progress}) \cdot p(\text{in}) \cdot p(\text{the}) \cdot p(\text{region} \mid \text{the}) \]

Train data:

... development and progress of the region. in ...

... 

Average: 2.7 words

Kneser & Ney (1996)

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

development and progress of the region.
This talk

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

\[ h_t = \sigma(Ux_t) \]
\[ \sigma(z) = \frac{1}{1 + \exp\{-z\}} \]
Feed-forward network

e.g. bigram LM

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

\[ h_t = \sigma(U x_t) \]

\[ y_t = s(V h_t) \]

\[ \sigma(z) = \frac{1}{1 + \exp\{-z\}} \]
Feed-forward network

e.g. bigram LM

\[ p(\text{progress} \mid \text{and}) \]

\[ y_t = s(V h_t) \]
\[ s(z) = \frac{\exp\{z\}}{\sum_{z'} \exp\{z'\}} \]
\[ h_t = \sigma(U x_t) \]
\[ \sigma(z) = \frac{1}{1 + \exp\{-z\}} \]

and
Feed-forward network

e.g. bigram LM

Classification

Feature learning

\[ p(\text{progress} \mid \text{and}) \]

\[ y_t = s(V h_t) \]

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

\[ h_t = \sigma(U x_t) \]

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

and
Feed-forward network

e.g. bigram LM
Classification
Feature learning

\[ s(z) = \frac{\exp\{z\}}{\sum_{z'} \exp\{z'\}} \]
\[ \sigma(z) = \frac{1}{1 + \exp\{-z\}} \]

Still based on limited context!
Recurrent network

e.g. bigram LM

\[ p(\text{progress} | \text{and}) \]

\[ h_t = \sigma(Ux_t) \]
Recurrent network

e.g. bigram LM

$p(\text{progress} \mid \text{and})$

$W$

$V$

$U$

$0 \ 0 \ 0 \ 1 \ 0$

$ht = \sigma(Ux_t)$

Dependence on previous time step
Recurrent network

e.g. bigram LM

\[ h_t = \sigma(Ux_t + Wh_{t-1}) \]
Recurrent network

e.g. bigram LM

\[ h_t = \sigma(Ux_t + Wh_{t-1}) \]
Recurrent network
Recurrent network

<s> p(development | <s>)
Recurrent network

<s> development
Recurrent network

<s>    development
Recurrent network

<s> development  p(\text{and} \mid 
\text{development},<s>)
Recurrent network

<s> development and
Recurrent network

<s> development and
Recurrent network

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

<s> development and progress
Recurrent network

<\textit{s}\textit{> development and progress

History of inputs up to current time-step
State of the art in language modeling \cite{mikolov2011}

More accurate than feed-forward nets \cite{sundermeyer2013}
Combined language and translation model

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
Combined language and translation model

Auli et al., EMNLP 2013

Entire source sentence representation
Combined language and translation model

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

Entire source sentence representation

Auli et al., EMNLP 2013
Combined language and translation model

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

Entire source sentence representation

Auli et al., EMNLP 2013
Combined language and translation model

\[ h_t = \sigma(U x_t + W h_{t-1} + F s) \]

Entire source sentence representation

Auli et al., EMNLP 2013
Combined language and translation model

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

Entire source sentence representation

Auli et al., EMNLP 2013
Combined language and translation model

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

Auli et al., EMNLP 2013

Entire source sentence representation
Combined language and translation model

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

Entire source sentence representation

Auli et al., EMNLP 2013
Combined language and translation model

本地的发展和进步

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

\[
h_t = \sigma(U x_t + W h_{t-1} + F s) (\ldots, \ldots, \ldots)
\]

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> 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

<\text{s}>  development
Combined language and translation model

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

Source word-window

development

本地 区 的 发展 和 进步
Combined language and translation model

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

Source word-window

本地区的发展和进步

<s> development and
Combined language and translation model

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

Source word-window
Combined language and translation model

本地的 发展 和 进步

\[ h_t = \sigma(Ux_t + Wh_{t-1} + Fs) \]
Combined language and translation model

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

Source word-window

<\text{s}> development and progress
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

<\text{s}> development and progress

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

<table>
<thead>
<tr>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.0</td>
</tr>
<tr>
<td>23.5</td>
</tr>
<tr>
<td>23.0</td>
</tr>
<tr>
<td>22.5</td>
</tr>
<tr>
<td>22.0</td>
</tr>
</tbody>
</table>

Phrase-based Koehn et al. (2003)
RNNLM Mikolov (2011)
RNNTM full-sent
RNNTM word-window

WMT 2012 French-English, 100M words, phrase-based baseline, n-best rescoring
Does the neural model learn to translate?

- Discrete translation model

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Phrase-based</th>
<th>RNNLM</th>
<th>RNNTM full-sent</th>
<th>RNNTM word-window</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.0</td>
<td>Koehn et al. (2003)</td>
<td>Mikolov (2011)</td>
<td>+0.8</td>
<td>+1.5</td>
</tr>
</tbody>
</table>

WMT 2012 French-English, 100M words, phrase-based baseline, n-best rescoring
<table>
<thead>
<tr>
<th>Improving a phrase-based baseline</th>
<th>+Discrete translation model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koehn et al. (2003)</td>
<td>24.5</td>
</tr>
<tr>
<td>Mikolov (2011)</td>
<td>25.0</td>
</tr>
<tr>
<td>RNLM</td>
<td>25.5</td>
</tr>
<tr>
<td>RNNTM word-window</td>
<td>26.0</td>
</tr>
<tr>
<td></td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>27.0</td>
</tr>
</tbody>
</table>

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

+Discrete translation model

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Phrase-based</th>
<th>RNNLM</th>
<th>RNNTM word-window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Koehn et al. (2003)</td>
<td>Mikolov (2011)</td>
<td>+1.4</td>
</tr>
</tbody>
</table>

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

BLEU

Phrases-based

n-best rescore

lattice rescore

decoder integration

24.9

25.7

26.4

26.9

+2.0
Neural nets vs discrete models

Average: 2.7 words
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

\[ \hat{e} = \text{argmax}_e \sum_i \lambda_i h_i(f, e) \quad h_1(f, e) = \log p_{lm}(e) \quad h_2(f, e) = \log p_{rnn}(e) \]
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

\[ \hat{e} = \text{argmax}_e \sum_i \lambda_i h_i(f, e) \quad h_1(f, e) = \log p_{lm}(e) \quad h_2(f, e) = \log p_{rnn}(e) \]

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

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

\[ \hat{e} = \arg\max_e \sum_i \lambda_i h_i(f, e) \quad h_1(f, e) = \log p_{lm}(e) \quad h_2(f, e) = \log p_{rnn}(e) \]

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

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

• Standard optimizer (MERT) to find weights for each n-gram order
Neural nets vs discrete models

Optimizer: MERT

Average: 2.7 words
Neural nets vs discrete models

Optimizer: MERT

Average: 2.7 words
Neural nets vs discrete models

“more important”

0.05

0.04

0.03

0.01

0

“less important”

1 2 3 4 5

n-gram order

Optimizer: MERT

Average: 2.7 words
Neural nets vs discrete models

"more important"

Feature weight

"less important"

n-gram order

Optimizer: MERT

Average: 2.7 words
Neural nets vs discrete models

"more important"  "less important"

Feature weight

0 0.01 0.02 0.03 0.04 0.05

1 2 3 4 5

n-gram order

n-gram RNN

Optimzer: MERT
Neural nets vs discrete models

"more important" weights:

n-gram model relies on unigram counts

Optimizer: MERT
Neural nets vs discrete models

A neural network model has higher-order statistics compared to a n-gram model. The n-gram model relies on unigram counts, while the n-gram order varies from 1 to 5. The diagram shows that the n-gram model has a higher feature weight for the first order, indicating it is more important. The average n-gram order is 2.7 words. Optimizer: 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

M1  M2  M3  M4  M5  M6
本 地区 的 发展 和 进步

development and progress of the region
Recurrent Minimum Translation Unit Models

The region of progress and development of the region
Recurrent Minimum Translation Unit Models

n-gram models over MTUs:
\[ p(M1) \ p(M2 | M1) \ p(M3 | M1,M2) \ldots \]

Banchs et al. (2005), Quirk & Menezes (2006)
Recurrent Minimum Translation Unit Models

<s>
Recurrent Minimum Translation Unit Models

<\text{s}>

\text{发展}

\text{development}
Recurrent Minimum Translation Unit Models
Recurrent Minimum Translation Unit Models

<s>

发展
development

和
and

Hu et al., EACL 2014
Recurrent Minimum Translation Unit Models

<\text{s}>

发展

development

和

and
Recurrent Minimum Translation Unit Models

Hu et al., EACL 2014

<s>

发展

development

和

and

进步

progress
Recurrent Minimum Translation Unit Models

Singletons

发展 development

和 and

进步 progress

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

<\text{s}>

development

and
Recurrent Minimum Translation Unit Models

Reduce sparsity by bag of words representation
Recurrent Minimum Translation Unit Models
<table>
<thead>
<tr>
<th>BLEU</th>
<th>Phrase-based</th>
<th>n-gram MTU</th>
<th>RNN MTU</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.0</td>
<td>Koehn et al. (2003)</td>
<td>Banchs et al. (2005), Quirk &amp; Menezes (2006)</td>
<td></td>
</tr>
<tr>
<td>26.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Type</td>
<td>BLEU Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phrase-based</td>
<td>25.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koehn et al. (2003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n-gram MTU</td>
<td>25.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banchs et al. (2005), Quirk &amp; Menezes (2006)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNN MTU</td>
<td>26.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+1.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This talk

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

True distribution (t)

Model distribution (y)

Delayed
Back propagation with cross entropy error

Model distribution (y)

True distribution (t)

\[ J = - \sum_i t^i \log y^i \]

Goal: Make correct outputs most likely
Back propagation with cross entropy error

True distribution \( (t) \):

\[
\begin{array}{cccccc}
0 & 0 & 1 & 0 & 0 & 0
\end{array}
\]

Error vector \( (\delta) \):

\[
\begin{array}{cccccc}
0.2 & 0.1 & -0.6 & 0.2 & 0.1 & 0
\end{array}
\]

Model distribution \( (y) \):

\[
\begin{array}{cccccc}
0.2 & 0.1 & 0.4 & 0.2 & 0.1 & 0
\end{array}
\]

\[\delta^i = y^i - t^i\]

\[J = - \sum_i t^i \log y^i\]

Goal: Make correct outputs most likely
Back propagation with cross entropy error

True distribution \((t)\)

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
</table>

Error vector \((\delta)\)

<table>
<thead>
<tr>
<th>0.2</th>
<th>0.1</th>
<th>-0.6</th>
<th>0.2</th>
<th>0.1</th>
</tr>
</thead>
</table>

Model distribution \((y)\)

\[
\delta^i = y^i - t^i
\]

\[
J = - \sum_i t^i \log y^i
\]

Goal: Make correct outputs most likely
Back propagation with cross entropy error

True distribution (t)

Error vector ($\delta$)

Model distribution (y)

\[ \delta^i = y^i - t^i \]

\[ J = - \sum_i t^i \log y^i \]

Goal: Make correct outputs most likely
Back propagation through time
Back propagation through time
Back propagation through time
Back propagation through time

<s> development and
Optimization

• Likelihood training very common

Auli & Gao, ACL 2014
Optimization

- Likelihood training very common
- **Next:** Task-specific training of neural nets for translation

Auli & Gao, ACL 2014
**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)

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

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 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 and progress of the region

System: advance 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

\[
L: \quad \max_{\theta} p(\tilde{e} | f ; \theta)
\]
Expected BLEU Training

(Smith 2006, He 2012, Gao 2014)

Desired translation output

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

xBLEU:

\[ \max_{\theta} \sum_{e \in E(f)} s\text{BLEU}(e, \tilde{e}) p(e | f; \theta) \]

Generated outputs

e.g. n-best, lattice

Gain function

Model probability
Expected BLEU Training

Human: development and progress of the region
Expected BLEU Training

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

<table>
<thead>
<tr>
<th>expression</th>
<th>sBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>advance and progress of the region</td>
<td>0.8</td>
</tr>
<tr>
<td>development and progress of this province</td>
<td>0.5</td>
</tr>
<tr>
<td>progress of this region</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Expected BLEU Training

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

| Human:                                                                 | sBLEU | $p_t(e|f;\theta)$ |
|-----------------------------------------------------------------------|-------|-------------------|
| development 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

<table>
<thead>
<tr>
<th>Human:</th>
<th>development and progress of the region</th>
</tr>
</thead>
<tbody>
<tr>
<td>advance and progress of the region</td>
<td>(0.8) (0.2)</td>
</tr>
<tr>
<td>development and progress of this province</td>
<td>(0.5) (0.3)</td>
</tr>
<tr>
<td>progress of this region</td>
<td>(0.3) (0.5)</td>
</tr>
</tbody>
</table>
Expected BLEU Training

\[
\text{Human: development and progress of the region}
\]

<table>
<thead>
<tr>
<th>advance and progress of the region</th>
<th>0.8</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>development and progress of this province</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>progress of this region</td>
<td>0.3</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
x\text{BLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e})p(e \mid f; \theta) = 0.5
\]
Expected BLEU Training

本地区的发展和进步

Human: development and progress of the region

| Test                                                                 | sBLEU | $p_t (\hat{e} | f; \theta)$ | $\delta_t$ |
|----------------------------------------------------------------------|-------|------------------------------|------------|
| 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       |

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

Human: development and progress of the region

|                          | sBLEU | $p_t(e|f;\theta)$ | $\delta_t$ | $p_{t+1}(e|f;\theta)$ |
|--------------------------|-------|-------------------|------------|-----------------------|
| 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                   |

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

本地区的发展和进步

Human: development and progress of the region

| Expression                                             | sBLEU | \( p_t(e|f;\theta) \) | \( \delta_t \) | \( p_{t+1}(e|f;\theta) \) |
|--------------------------------------------------------|-------|-------------------------|----------------|--------------------------|
| 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                      |

\[
\text{xBLEU} = \sum_{e \in E(f)} \text{sBLEU}(e, \tilde{e}) p(e|f;\theta) = 0.5
\]
Expected BLEU Training

Human: development and progress of the region

|                              | sBLEU | $p_t(e|f;\theta)$ | $\delta_t$ | $p_{t+1}(e|f;\theta)$ |
|------------------------------|-------|-------------------|------------|-----------------------|
| 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                   |

$$x\text{BLEU} = \sum_{e \in E(f)} s\text{BLEU}(e, \tilde{e})p(e|f;\theta) = 0.5 \rightarrow 0.6$$
Results

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Phrase-based Koehn et al. (2003)</th>
<th>RNN Cross Entropy</th>
<th>RNN +xBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>28.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Phrase-based Koehn et al. (2003)  

RNN Cross Entropy +0.8  

RNN +xBLEU +1.4
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

 BLEU

Baseline  4.5K  9K  900K  3M

Features

+2.0

Training set size

news2011

Auli et al., EMNLP 2014

+0.7
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


Marcel
NP :
marcel

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

completeness
NP :
completeness

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

S :
proved(marcel, completeness)

Combinatory Categorial Grammar (CCG; Steedman 2000)
How is this useful?

User query → (Semantic) parsing → Knowledge Base → Answer

Marcel proved completeness

$\lambda y. \lambda x. proved(x, y)$

$\lambda x. proved(x, completeness)$

$\lambda y. proved(x, y)$

$proved(marcel, completeness)$
How is this useful?

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

with belief propagation, dual decomposition and softmax-margin training

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

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

Marcel
proved
completeness

parsing factor
super tagging factor
inside-outside
forward-backward
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.

Fowler & Penn (2010)
Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+$\text{DecF}_1$).
- Hamming loss for supertagging sub-model (+$\text{Tagger}$).
- 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.
Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+DecF1).
- Hamming loss for supertagging sub-model (+Tagger).
- Belief propagation for inference.

Fowler & Penn (2010)

Auli & Lopez EMNLP 2011

best CCG parsing results to date
Integrated Parsing & Tagging

- F-measure loss for parsing sub-model (+DecF₁).
- 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

Marcel  NP  proved  \textit{S\(\backslash NP/NP\)}
Recurrent nets for CCG supertagging & parsing

Marcel \textit{NP} proved \textit{S\backslash NP/NP}
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