Smaller, faster, deeper: University of Edinburgh MT submission to WMT 2017

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, Philip Williams

University of Edinburgh

July 19 2017
Main collaborators

Rico Sennrich  Barry Haddow
Introduction to Neural MT

Making Models Smaller

Making Training Faster

Making Models Bigger

Using Monolingual Data

Ensembling and Reranking

Results
Linear models in MT

Phrase-based machine translation

- log-linear model: \( p(t) = \exp \sum_{i=1}^{n} \lambda_i h_i(x) \)

Weighted Model

- number of feature functions \( n \)
- random variables \( x = (e, f, \text{start}, \text{end}) \)
- feature functions
  \[
  h_1 = p(e | f) \text{ translation probability} \\
  h_2 = d(\text{start}_e - \text{start}_f) \text{ distortion} \\
  h_3 = d(p_{LM}) \text{ language model}
  \]
- weights \( \lambda \)
If caught early, the progression of the disease can be slowed down.

Si est détectée tôt, sa progression la maladie peut être ralentie.
Neural MT

Of course John has fun with the game.

\[ p(t|s) = f(s) \]

\( f(s) \) is a non-linear function represented by neural network.
Neural versus Phrase-base MT

**Phrase-based SMT**
Learn segment-segment correspondances from bitext
- Training is multistage pipeline of heuristics
- Fixed weights for features
- Limited ability to encode history
- Strong independence assumptions

**Neural MT**
Learn mathematical function on vectors from bitext
- End-to-end trained model
- Output conditioned on full source text and target history
- Non-linear dependence on information sources
Recurrent neural network

\[ h_i = f(x_i, h_{i-1}) \]
\[ y_i = g(h_i) \]
RNN for Language Modelling

- Predict $w_i$ conditioned on $w_1 \ldots w_{i-1}$
- Allows unlimited history
- Outperforms traditional count-based n-gram models
Encoder-Decoder

Decoding

Of course John has fun

Decoder

Encoder

NATürlich hat John Spaß

Alexandra Birch  UEDIN WMT 2017 9 / 36
Encoder-Decoder with Attention

Decoder

Encoder

natürlich  hat  john  spaß

of  course  john  has  fun

0.1  0.1  0.7  0.1

Decoder

Encoder

Barry Haddow (UEDIN) NMT: Success and Challenges Amazon 11 / 45

Alexandra Birch UEDIN WMT 2017 10 / 36
Limitations of Neural MT

Limitations
- Limited memory on GPUs
- Slow training times
- Not really deep deep learning models
- Training is not very stable
1 Introduction to Neural MT
2 Making Models Smaller
3 Making Training Faster
4 Making Models Bigger
5 Using Monolingual Data
6 Ensembling and Reranking
7 Results
Improvements to Subword Segmentation

Byte-pair encoding [Sennrich et al., 2016]
- iterative, frequency-based merging of subword units into larger units
- "joint BPE" on parallel corpus for more consistent segmentation

Problems
- subword unit can be part of (frequent) larger unit, but rare on its own
  - Allergikerzimmer 330
  - Allergiker: 10
- subword unit can be frequent in one language, but rare in the other
  - nationalities 541 (EN) 1 (DE)

Consequences
- model is unlikely to learn good representation for rare subwords
- BPE may even produce subword that is unknown at test time
- having rare subwords in vocabulary is wasteful
Improvements to Subword Segmentation

Solution
- require that subword has been observed in source training corpus
- optionally require minimum frequency
- even in joint BPE, condition is checked for each side individually
- split up (reverse merge of) subwords that don’t meet this requirement

Vocabulary size (EN→DE with joint BPE; 90k merge operations)

<table>
<thead>
<tr>
<th></th>
<th>BPE</th>
<th>EN</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>no filter</td>
<td></td>
<td>83227</td>
<td>91921</td>
</tr>
<tr>
<td>threshold 50</td>
<td></td>
<td>52652</td>
<td>73297</td>
</tr>
</tbody>
</table>

- only minimal change in sequence length and BLEU
- advantage: smaller models; no UNK
- disadvantage: additional hyperparameter: frequency threshold
Parameter Tying

Embedding and Output Layer

- in Nematus, last hidden layer has same size as target-size embedding
  → output matrix: vocabulary size × embedding layer size
  → target embedding matrix: embedding layer size × vocabulary size
- little effect on quality, but smaller models.
Optimizers

Optimizer

- adaptive learning rates tend to speed up training
- this year we used adam [Kingma and Ba, 2015] instead of adadelta [Zeiler, 2012]

Learning Rate Annealing

- our WMT systems use Adam without annealing
- SGD with annealing is popular [Sutskever et al., 2014, Wu et al., 2016]
- Adam with annealing recommended by [Denkowski and Neubig, 2017]
  → "--anneal_restarts 2 --patience 3" in Nematus
if input distribution to NN layer changes, parameters need to adapt to this **covariate shift**.

normalization of layers reduces shift, and improves training stability.

for layer $a$ with $H$ units, re-center and re-scale layer.

normalization changes representation power:

two bias parameters, $g$ and $b$, restore original representation power

$$
\mu = \frac{1}{H} \sum_{i=1}^{H} a_i 
$$

$$
\sigma = \sqrt{\frac{1}{H} \sum_{i=1}^{H} (a_i - \mu)^2}
$$

$$
h = \left[ \frac{g}{\sigma} \odot (a - \mu) + b \right]
$$
Figure: Alternating stacked encoder [Zhou et al., 2016].
Figure: Deep transition network.
we use depth of 4 for submission systems
all models were trained on single GPU
in post-submission experiments [Miceli Barone et al., 2017], BiDeep architecture (combination of deep transition and stack) performed best
Introduction to Neural MT
Making Models Smaller
Making Training Faster
Making Models Bigger
Using Monolingual Data
Ensembling and Reranking
Results
Monolingual Data

Back-translations
- all systems in the news shared task use target-side news data, automatically translated into source language

Copying
- the EN↔TR systems use monolingual data that is paired with a copy on the source

Biomedical task
- pseudo in-domain monolingual data is extracted from commoncrawl as follows:
  - automatically translate in-domain source corpus
  - perform Moore-Lewis data selection in target-language commoncrawl corpus
1. Introduction to Neural MT
2. Making Models Smaller
3. Making Training Faster
4. Making Models Bigger
5. Using Monolingual Data
6. Ensembling and Reranking
7. Results
Ensembling and Reranking: Research Questions

**Ensembling**
- last year, we used checkpoint ensemble (of last 4 checkpoints).
- this year, we contrast this with ensemble of independent models.

**Reranking with right-to-left models**
- last year, reranking with right-to-left models gave significant improvements for three translation directions.
- this year, we evaluate strategy on stronger baseline and more systems.
## News Task: Into English

<table>
<thead>
<tr>
<th>system</th>
<th>CS→EN 2017</th>
<th>DE→EN 2017</th>
<th>LV→EN 2017</th>
<th>RU→EN 2017</th>
<th>TR→EN 2017</th>
<th>ZH→EN 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT-16 single system</td>
<td>25.9</td>
<td>31.1</td>
<td>—</td>
<td>29.6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>baseline</td>
<td>27.5</td>
<td>32.0</td>
<td>16.4</td>
<td>31.3</td>
<td>19.7</td>
<td>21.7</td>
</tr>
<tr>
<td>+layer normalization</td>
<td>28.2</td>
<td>32.1</td>
<td>17.0</td>
<td>32.3</td>
<td>18.8</td>
<td>22.5</td>
</tr>
<tr>
<td>+deep model</td>
<td>28.9</td>
<td>33.5</td>
<td>16.6</td>
<td>32.7</td>
<td>20.6</td>
<td>22.9</td>
</tr>
<tr>
<td>+checkpoint ensemble</td>
<td>29.4</td>
<td>33.8</td>
<td>17.7</td>
<td>33.3</td>
<td>21.0</td>
<td>23.6</td>
</tr>
<tr>
<td>+independent ensemble</td>
<td>30.3</td>
<td>34.4</td>
<td>18.5</td>
<td>33.6</td>
<td>21.6</td>
<td>25.1</td>
</tr>
<tr>
<td>+right-to-left reranking</td>
<td>31.1</td>
<td>35.1</td>
<td>19.0</td>
<td>34.6</td>
<td>22.3</td>
<td>25.7</td>
</tr>
<tr>
<td>WMT-17 submission</td>
<td>30.9</td>
<td>35.1</td>
<td>19.0</td>
<td>30.8</td>
<td>20.1</td>
<td>25.7</td>
</tr>
</tbody>
</table>

We start from stronger baselines (more data, adam, new BPE). Layer normalization and deep models generally help. Checkpoint ensembles help, but independent ensembles are better. Reranking helps lead to large improvements over baseline.
### News Task: Into English

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT-16 single system baseline</td>
<td>25.9</td>
<td>31.1</td>
<td>—</td>
<td>29.6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>WMT-16 single system baseline</td>
<td>27.5</td>
<td>32.0</td>
<td>16.4</td>
<td>31.3</td>
<td>19.7</td>
<td>21.7</td>
</tr>
</tbody>
</table>

- we start from stronger baselines (more data, adam, new BPE)
### News Task: Into English

<table>
<thead>
<tr>
<th>system</th>
<th>CS→EN 2017</th>
<th>DE→EN 2017</th>
<th>LV→EN 2017</th>
<th>RU→EN 2017</th>
<th>TR→EN 2017</th>
<th>ZH→EN 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>27.5</td>
<td>32.0</td>
<td>16.4</td>
<td>31.3</td>
<td>19.7</td>
<td>21.7</td>
</tr>
<tr>
<td>+layer normalization</td>
<td>28.2</td>
<td>32.1</td>
<td>17.0</td>
<td>32.3</td>
<td>18.8</td>
<td>22.5</td>
</tr>
<tr>
<td>+deep model</td>
<td>28.9</td>
<td>33.5</td>
<td>16.6</td>
<td>32.7</td>
<td>20.6</td>
<td>22.9</td>
</tr>
</tbody>
</table>

- we start from stronger baselines (more data, adam, new BPE)
- layer normalization and deep models generally help
### News Task: Into English

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+deep model</td>
<td>28.9</td>
<td>33.5</td>
<td>16.6</td>
<td>32.7</td>
<td>20.6</td>
<td>22.9</td>
</tr>
<tr>
<td>+checkpoint ensemble</td>
<td>29.4</td>
<td>33.8</td>
<td>17.7</td>
<td>33.3</td>
<td>21.0</td>
<td>23.6</td>
</tr>
<tr>
<td>+independent ensemble</td>
<td>30.3</td>
<td>34.4</td>
<td>18.5</td>
<td>33.6</td>
<td>21.6</td>
<td>25.1</td>
</tr>
</tbody>
</table>

- We start from stronger baselines (more data, adam, new BPE)
- Layer normalization and deep models generally help
- Checkpoint ensembles help, but independent ensembles are better
## News Task: Into English

<table>
<thead>
<tr>
<th>system</th>
<th>CS→EN 2017</th>
<th>DE→EN 2017</th>
<th>LV→EN 2017</th>
<th>RU→EN 2017</th>
<th>TR→EN 2017</th>
<th>ZH→EN 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>+independent ensemble</td>
<td>30.3</td>
<td>34.4</td>
<td>18.5</td>
<td>33.6</td>
<td>21.6</td>
<td>25.1</td>
</tr>
<tr>
<td>+right-to-left reranking</td>
<td>31.1</td>
<td>35.1</td>
<td>19.0</td>
<td>34.6</td>
<td>22.3</td>
<td>25.7</td>
</tr>
</tbody>
</table>

- We start from stronger baselines (more data, adam, new BPE)
- Layer normalization and deep models generally help
- Checkpoint ensembles help, but independent ensembles are better
- Reranking helps
### News Task: Into English

<table>
<thead>
<tr>
<th>system</th>
<th>CS→EN 2017</th>
<th>DE→EN 2017</th>
<th>LV→EN 2017</th>
<th>RU→EN 2017</th>
<th>TR→EN 2017</th>
<th>ZH→EN 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT-16 single system</td>
<td>25.9</td>
<td>31.1</td>
<td>—</td>
<td>29.6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>baseline</td>
<td>27.5</td>
<td>32.0</td>
<td>16.4</td>
<td>31.3</td>
<td>19.7</td>
<td>21.7</td>
</tr>
<tr>
<td>+layer normalization</td>
<td>28.2</td>
<td>32.1</td>
<td>17.0</td>
<td>32.3</td>
<td>18.8</td>
<td>22.5</td>
</tr>
<tr>
<td>+deep model</td>
<td>28.9</td>
<td>33.5</td>
<td>16.6</td>
<td>32.7</td>
<td>20.6</td>
<td>22.9</td>
</tr>
<tr>
<td>+checkpoint ensemble</td>
<td>29.4</td>
<td>33.8</td>
<td>17.7</td>
<td>33.3</td>
<td>21.0</td>
<td>23.6</td>
</tr>
<tr>
<td>+independent ensemble</td>
<td>30.3</td>
<td>34.4</td>
<td>18.5</td>
<td>33.6</td>
<td>21.6</td>
<td>25.1</td>
</tr>
<tr>
<td>+right-to-left reranking</td>
<td>31.1</td>
<td>35.1</td>
<td>19.0</td>
<td>34.6</td>
<td>22.3</td>
<td>25.7</td>
</tr>
<tr>
<td>WMT-17 submission</td>
<td>30.9</td>
<td>35.1</td>
<td>19.0</td>
<td>30.8</td>
<td>20.1</td>
<td>25.7</td>
</tr>
</tbody>
</table>

- We start from stronger baselines (more data, adam, new BPE)
- Layer normalization and deep models generally help
- Checkpoint ensembles help, but independent ensembles are better
- Reranking helps
- Large improvements over baseline
## News Task: Out of English

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT16 single system</td>
<td>19.7</td>
<td>24.9</td>
<td>—</td>
<td>26.7</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>baseline</td>
<td>20.5</td>
<td>26.1</td>
<td>14.6</td>
<td>28.0</td>
<td>15.6</td>
<td>31.3</td>
</tr>
<tr>
<td>+layer normalization</td>
<td>20.5</td>
<td>26.1</td>
<td>14.9</td>
<td>28.7</td>
<td>15.7</td>
<td>32.3</td>
</tr>
<tr>
<td>+deep model</td>
<td>21.1</td>
<td>26.6</td>
<td>15.1</td>
<td>29.9</td>
<td>16.2</td>
<td>33.4</td>
</tr>
<tr>
<td>+checkpoint ensemble</td>
<td>22.0</td>
<td>27.5</td>
<td>16.1</td>
<td>31.0</td>
<td>16.7</td>
<td>33.5</td>
</tr>
<tr>
<td>+independent ensemble</td>
<td>22.8</td>
<td>28.3</td>
<td>16.7</td>
<td>31.6</td>
<td>17.6</td>
<td>35.8</td>
</tr>
<tr>
<td>+right-to-left reranking</td>
<td>22.8</td>
<td>28.3</td>
<td>16.9</td>
<td>–</td>
<td>18.1</td>
<td>36.3</td>
</tr>
<tr>
<td>WMT-17 submission</td>
<td>22.8</td>
<td>28.3</td>
<td>16.9</td>
<td>29.8</td>
<td>16.5</td>
<td>36.3</td>
</tr>
</tbody>
</table>
### Biomedical Task

<table>
<thead>
<tr>
<th>system</th>
<th>EN→PL</th>
<th>EN→RO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coch</td>
<td>NHS24</td>
</tr>
<tr>
<td>baseline</td>
<td>26.2</td>
<td>18.2</td>
</tr>
<tr>
<td>+ layer normalization</td>
<td>25.5</td>
<td>20.2</td>
</tr>
<tr>
<td>+ deep model</td>
<td>25.9</td>
<td>20.2</td>
</tr>
<tr>
<td>+ checkpoint ensemble</td>
<td>28.4</td>
<td>21.3</td>
</tr>
<tr>
<td>+ independent ensemble</td>
<td>28.1</td>
<td>21.6</td>
</tr>
<tr>
<td>+ right-to-left reranking</td>
<td>28.6</td>
<td>22.5</td>
</tr>
<tr>
<td>WMT17 submission</td>
<td>29.0</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Copied Monolingual Data

Table: BLEU scores for EN↔TR when adding copied monolingual data.

<table>
<thead>
<tr>
<th>system</th>
<th>TR→EN</th>
<th>EN→TR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td>baseline</td>
<td>20.0</td>
<td>13.2</td>
</tr>
<tr>
<td>+copied</td>
<td>20.2</td>
<td>13.8</td>
</tr>
<tr>
<td></td>
<td>2016</td>
<td>2017</td>
</tr>
<tr>
<td></td>
<td>19.7</td>
<td>14.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.6</td>
</tr>
</tbody>
</table>
## Biomedical Task: Domain Adaptation

<table>
<thead>
<tr>
<th>system</th>
<th>EN→PL</th>
<th>EN→RO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coch</td>
<td>NHS24</td>
</tr>
<tr>
<td>generic (single)</td>
<td>22.8</td>
<td>16.6</td>
</tr>
<tr>
<td>generic (ensemble 4)</td>
<td>23.6</td>
<td>19.9</td>
</tr>
<tr>
<td>fine-tuned (single)</td>
<td>27.2</td>
<td>19.5</td>
</tr>
<tr>
<td>fine-tuned (ensemble 4)</td>
<td>27.4</td>
<td>20.9</td>
</tr>
</tbody>
</table>
Thank you!
Stronger Baselines for Trustable Results in Neural Machine Translation.
ArXiv e-prints.

In The International Conference on Learning Representations, San Diego, California, USA.

Deep Architectures for Neural Machine Translation.
Association for Computational Linguistics.

Using the Output Embedding to Improve Language Models.

Neural Machine Translation of Rare Words with Subword Units.
Association for Computational Linguistics.

Sequence to Sequence Learning with Neural Networks.
