Moses in a Neural MT World

Hieu Hoang
MT Marathon 2017
Dayton, Ohio
SMT Moses in a Neural MT World

Hieu Hoang
MT Marathon 2017
Dayton, Ohio
Why SMT still matters in an NMT World?

Hieu Hoang
MT Marathon 2017
Dayton, Ohio
Why SMT still matters?

● It’s still used
● Translation quality
   ○ Low resource languages
   ○ Out-of-domain
   ○ Hierarchical/Syntax models
● Translation speed
   ○ Speed / Quality trade-off
   ○ Modern hardware
● Hybridization
   ○ NMT in SMT
   ○ SMT in NMT
Why SMT still matters?

● It’s still used

● Translation quality
  ○ Low resource languages
  ○ Out-of-domain
  ○ Hierarchical/Syntax models

● Translation speed
  ○ Speed / Quality trade-off
  ○ Modern hardware

● Hybridization
  ○ NMT in SMT
  ○ SMT in NMT
SMT is still used

- EU, UN, WIPO, AutoDesk, Adobe, Amazon, eBay, Unbabel, ...
  - Use Moses
  - Looking at NMT
- Google
  - 20 language pairs uses NMT
  - 82 more to go!
- Bing
  - 11 language pairs uses NMT
  - 50 more to go!
- Why?
  - .....
Why SMT still matters?

● It’s still used
● Translation quality
  ○ Low resource languages
  ○ Out-of-domain
  ○ Hierarchical models
● Translation speed
  ○ Speed / Quality trade-off
  ○ Modern hardware
● Hybridization
  ○ NMT in SMT
  ○ SMT in NMT
## Translation Quality

### ‘NMT beats SMT’

<table>
<thead>
<tr>
<th>en-de</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>34.2</td>
</tr>
<tr>
<td>metamind</td>
<td>32.3</td>
</tr>
<tr>
<td>NYU-UMontreal</td>
<td>30.8</td>
</tr>
<tr>
<td>cambridge</td>
<td>30.6</td>
</tr>
<tr>
<td>uedin-syntax</td>
<td>30.6</td>
</tr>
<tr>
<td>KIT/LIMSI</td>
<td>29.1</td>
</tr>
<tr>
<td>KIT</td>
<td>29.0</td>
</tr>
<tr>
<td>uedin-pbmt</td>
<td>28.4</td>
</tr>
<tr>
<td>jhu-syntax</td>
<td>26.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>de-en</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>uedin-nmt</td>
<td>38.6</td>
</tr>
<tr>
<td>uedin-pbmt</td>
<td>35.1</td>
</tr>
<tr>
<td>jhu-pbmt</td>
<td>34.5</td>
</tr>
<tr>
<td>uedin-syntax</td>
<td>34.1</td>
</tr>
<tr>
<td>KIT</td>
<td>33.9</td>
</tr>
<tr>
<td>jhu-syntax</td>
<td>31.0</td>
</tr>
</tbody>
</table>

*Pure NMT*  
*Pure SMT*

(Sennrich, MTMA 2016)
## Translation Quality

‘NMT beats SMT’

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pb</td>
<td>53.07</td>
<td>49.77</td>
<td>42.80</td>
<td>36.00</td>
<td>31.58</td>
<td>41.96</td>
<td>61.26</td>
<td>50.09</td>
<td>43.25</td>
<td>37.84</td>
</tr>
<tr>
<td>NMT</td>
<td><strong>55.97</strong></td>
<td><strong>52.16</strong></td>
<td><strong>44.67</strong></td>
<td><strong>38.77</strong></td>
<td><strong>39.73</strong></td>
<td><strong>45.07</strong></td>
<td><strong>62.12</strong></td>
<td><strong>51.46</strong></td>
<td><strong>45.27</strong></td>
<td><strong>46.82</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pb</td>
<td>38.13</td>
<td>59.89</td>
<td>49.76</td>
<td>39.69</td>
<td>31.27</td>
<td>34.43</td>
<td>52.22</td>
<td><strong>52.44</strong></td>
<td>36.48</td>
<td>29.98</td>
</tr>
<tr>
<td>NMT</td>
<td><strong>39.69</strong></td>
<td><strong>61.13</strong></td>
<td><strong>49.75</strong></td>
<td><strong>40.96</strong></td>
<td><strong>41.21</strong></td>
<td><strong>35.53</strong></td>
<td><strong>52.38</strong></td>
<td><strong>51.78</strong></td>
<td><strong>37.26</strong></td>
<td><strong>37.37</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pb</td>
<td>34.43</td>
<td>52.59</td>
<td>49.61</td>
<td>43.37</td>
<td>32.63</td>
<td>28.02</td>
<td>42.97</td>
<td>39.64</td>
<td>34.42</td>
<td>29.57</td>
</tr>
<tr>
<td>NMT</td>
<td><strong>35.73</strong></td>
<td><strong>52.35</strong></td>
<td><strong>50.45</strong></td>
<td><strong>44.14</strong></td>
<td><strong>39.42</strong></td>
<td><strong>35.38</strong></td>
<td><strong>51.77</strong></td>
<td><strong>48.29</strong></td>
<td><strong>42.00</strong></td>
<td><strong>36.41</strong></td>
</tr>
</tbody>
</table>

(Junczys-Dowmunt et al, 2016)
Why SMT still matters?

● It’s still used
● Translation quality
  ○ Low resource languages
  ○ Out-of-domain
  ○ Hierarchical/Syntax models
● Translation speed
  ○ Speed / Quality trade-off
  ○ Modern hardware
● Hybridization
  ○ NMT in SMT
  ○ SMT in NMT
# Translation Quality

## Low Resource Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>SMT (Syntax)</th>
<th>NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausa</td>
<td>23.7</td>
<td>16.8</td>
</tr>
<tr>
<td>Turkish</td>
<td>20.4</td>
<td>11.4</td>
</tr>
<tr>
<td>Uzbek</td>
<td>17.9</td>
<td>10.7</td>
</tr>
<tr>
<td>Urdu</td>
<td>17.9</td>
<td>5.2</td>
</tr>
</tbody>
</table>

(Zoph et al, 2016)

<table>
<thead>
<tr>
<th>Language</th>
<th>SMT (PB)</th>
<th>NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausa</td>
<td>29.48</td>
<td>27.32</td>
</tr>
</tbody>
</table>

(Wu et al, 2016)
Why SMT still matters?

- It’s still used
- Translation quality
  - Low resource languages
  - Out-of-domain
  - Hierarchical/Syntax models
- Translation speed
  - Speed / Quality trade-off
  - Modern hardware
- Hybridization
  - NMT in SMT
  - SMT in NMT
## Translation Quality

### Out-of-domain translation

<table>
<thead>
<tr>
<th>Catalan-Spanish</th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>75.20</td>
<td>50.53</td>
</tr>
<tr>
<td>Phrase-based</td>
<td>81.80</td>
<td><strong>57.20</strong></td>
</tr>
<tr>
<td>Neural MT</td>
<td><strong>83.01</strong></td>
<td>52.10</td>
</tr>
</tbody>
</table>

(Costa-jussa, 2017)
Translation Quality

Learning curves

(Anonymous, under review)
Translation Quality

Human Knowledge

- SMT
  - Reordering
    - Distortion penalty
    - Reordering limit
  - Word penalty
  - Coverage
    - Translate each word once and only once
  - Rule-based translation
    - Named entities, numbers
  - Syntax

- NMT
  - Topology
    - Sequence-to-sequence
    - LSTM/GRU
    - Attention
Why SMT still matters?

● It’s still used
● Translation quality
  ○ Low resource languages
  ○ Out-of-domain
  ○ Hierarchical/Syntax models
● Translation speed
  ○ Speed / Quality trade-off
  ○ Modern hardware
● Hybridization
  ○ NMT in SMT
  ○ SMT in NMT
## Translation Quality

### Hierarchical/Syntax Translation

<table>
<thead>
<tr>
<th></th>
<th>en-jp</th>
<th></th>
<th>zh-en</th>
<th>en-zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase-based</td>
<td>29.80</td>
<td>43.0</td>
<td>37.84</td>
<td></td>
</tr>
<tr>
<td>Hierarchical</td>
<td><strong>32.56</strong></td>
<td><strong>47.3</strong></td>
<td><strong>41.70</strong></td>
<td></td>
</tr>
<tr>
<td>NMT</td>
<td>32.19</td>
<td>51.8</td>
<td>46.82</td>
<td></td>
</tr>
</tbody>
</table>

(Chen et al., 2015)
Why SMT still matters?

- It’s still used
- Translation quality
  - Low resource languages
  - Out-of-domain
  - Hierarchical/Syntax models
- Translation speed
  - Speed / Quality trade-off
  - Modern hardware
- Hybridization
  - NMT in SMT
  - SMT in NMT
Translation Speed

- NMT
  - Specialized hardware
  - Efficient software
    - Batching

<table>
<thead>
<tr>
<th></th>
<th>Words / sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses (16 cores)</td>
<td>455.3</td>
</tr>
<tr>
<td>Nematus (1 GPU)</td>
<td>268.4</td>
</tr>
<tr>
<td>AmuNMT (16 cores)</td>
<td>140.9</td>
</tr>
<tr>
<td>AmuNMT (1 GPU)</td>
<td><strong>864.7</strong></td>
</tr>
</tbody>
</table>

(Junczys-Dowmunt et al, 2016)
Translation Speed

- **SMT**
  - Larger servers
  - Efficient software
    - Multi-threading

<table>
<thead>
<tr>
<th>Europarl data</th>
<th>Words / sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses (32 cores)</td>
<td>836</td>
</tr>
<tr>
<td>Moses2 (32 cores)</td>
<td>18,764</td>
</tr>
<tr>
<td>AmuNMT (1 GPU)</td>
<td>3,488</td>
</tr>
</tbody>
</table>
Translation Speed

Speed / Quality Trade-off

Varying cube-pruning pop-limit (SMT), beam size (NMT)
Why SMT still matters?

- It’s still used
- Translation quality
  - Low resource languages
  - Out-of-domain
  - Hierarchical/Syntax models
- Translation speed
  - Speed / Quality trade-off
    - Modern hardware
- Hybridization
  - NMT in SMT
  - SMT in NMT
Translation Speed
Moses decoder

- General purpose
- Feature rich
- Started 10 years ago
- Poor multi-core scalability
Translation Speed

Moses2 decoder

● Prioritize
  ○ Decoding Speed
  ○ Multi-threaded
  ○

● Today’s servers
  ○ Many cores
  ○ Lots of RAM

● Efficient memory management
● Phrase-table optimization
● Lexicalized reordering model optimization
Translation Speed
Moses2 decoder

Phrase-Based

Hierarchical

(words/sec vs. # threads)

Moses  Moses2

(words/sec vs. # threads)

Moses  Moses2
Why SMT still matters?

- It’s still used
- Translation quality
  - Low resource languages
  - Out-of-domain
  - Hierarchical/Syntax models
- Translation speed
  - Speed / Quality trade-off
  - Modern hardware
- Hybridization
  - NMT in SMT
  - SMT in NMT
Hybridization

● Neural networks in SMT
  ○ N-best reranking
  ○ Neural feature functions
  ○ Encoder-decoder RNN

● SMT in NMT
  ○ Inspired by SMT
    ■ Attention models
    ■ Coverage models
  ○ Language models
  ○ Phrase-tables

● Systems combination
Conclusion

Why SMT still matters?

- It’s still used
- Translation quality
  - Low resource languages
  - Out-of-domain
  - Hierarchical/Syntax models
- Translation speed
- Hybridization