Summary

- Neural Machine Translation models perform best when an abundance of parallel data is available
- Acquiring human translations for low-resource language pairs is costly
- Hence translation of low-frequency words is difficult and often inaccurate
- Our approach
  - alters existing parallel sentences targeting low-frequency words
  - augments the data by generating new diverse context for low-frequency words and the corresponding translations
  - As a result training with the augmented bitext achieves significant BLEU improvements in a simulated low-resource English-German translation setting

Approach: Translation Data Augmentation (TDA)

![Diagram showing the approach of TDA](image)

Choosing the best translation of $s'_i$ ...

\[ t'_j = \arg \max_{t \in \text{trans}(s'_i)} p_{\text{lex}}(t|s'_i) p_{\text{lexin}}(s'_i|t) p_{\text{LM}}(t|t_1^{-1}) \]

New sentence pair:

I had been told that you would voluntarily be speaking today.
mir wurde signalisiert, sie würden heute freiwillig sprechen.

NMT Results (BLEU)

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>testset2014</th>
<th>testset2015</th>
<th>testset2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>371K</td>
<td>10.6</td>
<td>11.3</td>
<td>13.1</td>
</tr>
<tr>
<td>Back-trans</td>
<td>731K</td>
<td>11.4 (+0.8)</td>
<td>12.2 (+0.9)</td>
<td>14.6 (+1.5)</td>
</tr>
<tr>
<td>TDA$\rightarrow$</td>
<td>4.5M</td>
<td>11.9 (+1.3)**</td>
<td>13.4 (+2.1)**</td>
<td>15.2 (+2.1)**</td>
</tr>
<tr>
<td>TDA$\rightarrow$</td>
<td>6M</td>
<td>12.6 (+2.0)**</td>
<td>13.7 (+2.4)**</td>
<td>15.4 (+2.5)**</td>
</tr>
</tbody>
</table>

EN$\rightarrow$DE

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>testset2014</th>
<th>testset2015</th>
<th>testset2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>371K</td>
<td>8.2</td>
<td>9.2</td>
<td>11.0</td>
</tr>
<tr>
<td>Back-trans</td>
<td>731K</td>
<td>9.0 (+0.8)</td>
<td>10.4 (+1.2)</td>
<td>12.0 (+1.0)**</td>
</tr>
<tr>
<td>TDA$\rightarrow$</td>
<td>4.5M</td>
<td>10.4 (+2.2)**</td>
<td>11.2 (+2.0)**</td>
<td>13.5 (+2.5)**</td>
</tr>
<tr>
<td>TDA$\rightarrow$</td>
<td>6M</td>
<td>10.7 (+2.5)**</td>
<td>11.5 (+2.3)**</td>
<td>13.9 (+2.9)**</td>
</tr>
</tbody>
</table>

Conclusions

- We present a data augmentation technique to enrich the training data targeting rare words
- Increasing the size of the training data by diversifying the context of rare words yields better translations
- Generation of correct rare words during translation increases
- The attention scores of rare words are on average 8.8% higher than the baseline model
- The generated translation length to reference length ratio is on average 7% higher

Data Augmentation

- Image Processing
  - Flipping, cropping, tilting, altering the RGB channels
- Has not been done in Natural Language Processing
- One possible approach is paraphrasing which is meaning-preserving
- Our approach focuses on non meaning-preserving augmentation
- Closest work is back-translation of monolingual data (Sennrich et al. ACL 2016)

Rare Translation Generation (DE$\rightarrow$EN)

![Graph showing rare translation generation results](image)

- Increasing the size of the training data by diversifying the context of rare words
- The attention scores of rare words are on average 8.8% higher than the baseline model
- The generated translation length to reference length ratio is on average 7% higher