**Background**

- **Problem**: Translate Quenya, an Elvish language created by author J.R.R. Tolkien, to English
- **Low-resource setting**: As Elvish is a fictional language, we have very limited, mostly fan-sourced data
  - Existing approaches: transfer learning, universal representation, meta learning, data augmentation

**Data**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Pair</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bible</td>
<td>Elvish</td>
<td>7,135</td>
<td>396</td>
<td>396</td>
<td>Main training</td>
</tr>
<tr>
<td>Misc</td>
<td>English</td>
<td>-</td>
<td>-</td>
<td>215</td>
<td>Test</td>
</tr>
<tr>
<td>Dictionary</td>
<td></td>
<td>5,107</td>
<td>-</td>
<td>-</td>
<td>Augmentation</td>
</tr>
<tr>
<td>Europarl</td>
<td>Finnish</td>
<td>1M</td>
<td>1,000</td>
<td>3,000</td>
<td>Transfer, LM</td>
</tr>
<tr>
<td>Newstest</td>
<td>English</td>
<td>-</td>
<td>-</td>
<td>3,000</td>
<td>Scarcity</td>
</tr>
</tbody>
</table>

**Model**

![Diagram of the model](image)

**Data Augmentation**

- **Idea**: generate new examples that use vocabulary which do not appear in our training set
- **Process**: 
  1. Use forward and backward language models to generate new words in sentences
  2. Use dictionary to make appropriate substitutions, making new examples

**Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Loss</th>
<th>Train PPL</th>
<th>Bible BLEU</th>
<th>Misc BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>35.71</td>
<td>4.33</td>
<td>60.11</td>
<td>9.89</td>
</tr>
<tr>
<td>Transfer</td>
<td>44.18</td>
<td>6.14</td>
<td>61.35</td>
<td>10.41</td>
</tr>
<tr>
<td>Augmented</td>
<td>39.65</td>
<td>11.68</td>
<td>61.08</td>
<td>11.38</td>
</tr>
</tbody>
</table>

**Translation Example**

Elvish: Mernelye fire ar harya alcar
Gold Translation: You wanted to die and have glory
Our Translation: You wanted to die and possess glory

**Augmentation Example**

Original: For you do not know what day your Lord will come
Augment: For you do not grasp what day your Lord will come

Original: An ualde ista mi mana ré Herulda tuluva
Augment: An ualde mapa mi mana ré Herulda tuluva

**Conclusion & Challenges**

- Lack of data makes training and evaluation difficult
- It is difficult to balance augmentation and repetition
- Transfer learning does help by providing a good initial model
- Quality over quantity: better augmented examples more helpful than large number of examples

**Future Work**

- Different ways to combine forward and backward language models to generate new candidates
- Named entity recognition to prevent mistranslation
- Better parsing (normalize case, stemming)

**References**