ABSTRACT
Transfer learning is a promising direction for low-resource NMT, but introduces many new variables. The effects of these variables, such as auxiliary task choice and dataset sizes, are typically analysed independently due to computational cost. We hypothesise that these factors are not independent and demonstrate initial evidence via a 3-way (7x8x2) systematic study which reveals statistically significant (p-value < 0.0018) non-trivial interactions between main and auxiliary dataset sizes and task relatedness.

EXPERIMENTAL SETUP
- Main task: French -> English (Fr -> En) translation
- Aux. tasks: Portuguese -> English (Pt -> En) & Russian to English (Ru -> En).
- Main task dataset sizes: [4096, 8192, 16384, 32768, 65536, 131072].
- Aux. task dataset sizes: [0, 4096, 8192, 16384, 32768, 65536, 131072].
- Data: screened language learning sentence pairs from the Tatoeba project [1].
- Networks: Transformer models mostly following details from the original transformer paper [2].
- Metric: BLEU scores [3] on auxiliary and main test sets after pre-training and again after fine-tuning.

We conduct 8-fold cross-validation (CV) over all combinations of the 2 aux. tasks, 6 main dataset sizes, and 7 aux. dataset sizes. This results in a total of 672 networks over 84 conditions.

BASELINE AND ZERO-SHOT RESULTS

More (similar quality) data is always better for baseline (bilingual) models. However, returns are diminishing.

Closely-related data is always better for zero-shot performance.

MEAN BLEU SCORES

Transfer learning is a promising direction for low-resource NMT, but introduces many new variables. The effects of these variables, such as auxiliary task choice and dataset sizes, are typically analysed independently due to computational cost. We hypothesise that these factors are not independent and demonstrate initial evidence via a 3-way (7x8x2) systematic study which reveals statistically significant (p-value < 0.0018) non-trivial interactions between main and auxiliary dataset sizes and task relatedness.

```
<table>
<thead>
<tr>
<th>Pt-&gt;En</th>
<th>Ru-&gt;En</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Experimental Setup

- Main task: French -> English (Fr -> En) translation
- Aux. tasks: Portuguese -> English (Pt -> En) & Russian to English (Ru -> En).
- Main task dataset sizes: [4096, 8192, 16384, 32768, 65536, 131072].
- Aux. task dataset sizes: [0, 4096, 8192, 16384, 32768, 65536, 131072].
- Data: screened language learning sentence pairs from the Tatoeba project [1].
- Networks: Transformer models mostly following details from the original transformer paper [2].
- Metric: BLEU scores [3] on auxiliary and main test sets after pre-training and again after fine-tuning.

We conduct 8-fold cross-validation (CV) over all combinations of the 2 aux. tasks, 6 main dataset sizes, and 7 aux. dataset sizes. This results in a total of 672 networks over 84 conditions.

Baseline and Zero-Shot Results

More (similar quality) data is always better for baseline (bilingual) models. However, returns are diminishing.

Closely-related data is always better for zero-shot performance.

```
<table>
<thead>
<tr>
<th>Pt-&gt;En</th>
<th>Ru-&gt;En</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Future Work

- Extend studies to more main and auxiliary tasks (particularly to true low-resource languages) to investigate how results generalise.
- Replicate study to confirm results.
- Balance datasets before tokenization and limit effects of negative transfer for larger datasets.

Related Work

Other work by the authors to improve low-resource NMT is aimed at facilitating easier, faster, and less error-prone research, and facilitating independent NMT research by language researchers:

- graphicalNMT: A system for code-free design, creation, and evaluation of NMT systems
- NLPDataTools: A system for code-free evaluation, analysis, and manipulation of NLP datasets.

References


The University of Queensland, Australia