The University of Edinburgh’s Bengali↔Hindi Submissions to the WMT21 News Translation Task

Proyag Pal  Alham Fikri Aji  Pinzhen Chen  Sukanta Sen

Overview

- **UEdin’s** Bengali→Hindi (bn→hi) and Hindi→Bengali (hi→bn) systems submitted to the News Translation task at WMT21.
- **Top (tied)** systems among all constrained submissions for both directions, according to human evaluation.
- Our approach mainly focuses on cleaning, back-translation, and fine-tuning to the target domain.
- All models are trained with parallel and synthetic data, fine-tuned on retrieved in-domain data, further fine-tuned on dev set. Models fine-tuned in different ways are ensembled.

Data and Cleaning

Constrained condition:

- 3.3M parallel sentence pairs from CCAligned
- NewsCrawl monolingual: 10.1M lines of bn, 46.1M lines of hi
- CommonCrawl monolingual: 49.6M lines of bn, 202M lines of hi

Training with Synthetic Data

Use back-translation and forward translation using models trained only on parallel data to generate synthetic data. Use this synthetic data in different ways:

- Tagged back-translation
- Train models on all back-translated data, then continue training with parallel data only
- Train on parallel, back-translated, and forward translated data, then continue training with parallel data only

Decoding and Post-processing

- Ensemble many fine-tuned models to decode.
- Run sentence splitter on test source: rejoin outputs.
- Transliterate all numerals to Latin script for consistency.

Model and Training Configuration

- **32k subword** SentencePiece vocabulary shared between source and target sides.
- **Transformer-Big** architecture - 6 encoder + 6 decoder layers, 16 heads, embedding size 1024, unit size 4096.
- 32GB dynamic batch size. Adam optimizer with learning rate 0.0003, optimizer delay 3, early stop when dev set BLEU doesn’t improve for 20k updates.

Fine-tuning

Adapt to the target domain by retrieving sentences similar to the dev/test set and fine-tuning the models on those subsets of sentences. Finally, fine-tune to the dev set, since that’s the most in-domain data available.

<table>
<thead>
<tr>
<th>Language Direction</th>
<th>Ave.</th>
<th>Ave.</th>
<th>z Score</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) bn→hi</td>
<td>82.1</td>
<td>0.202</td>
<td>GTCOM</td>
<td>79.1</td>
</tr>
<tr>
<td></td>
<td>77.5</td>
<td>0.080</td>
<td>TRANSSION</td>
<td>78.0</td>
</tr>
<tr>
<td></td>
<td>78.0</td>
<td>0.054</td>
<td>UEdin</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>75.7</td>
<td>-0.080</td>
<td>HuaweiTSC</td>
<td>75.7</td>
</tr>
<tr>
<td></td>
<td>70.8</td>
<td>-0.373</td>
<td>Online-G</td>
<td>70.8</td>
</tr>
<tr>
<td>(b) hi→bn</td>
<td>95.0</td>
<td>0.245</td>
<td>HuaweiTSC</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>94.5</td>
<td>0.233</td>
<td>GTCOM</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>92.3</td>
<td>0.080</td>
<td>Online-Y</td>
<td>92.0</td>
</tr>
<tr>
<td></td>
<td>91.3</td>
<td>0.029</td>
<td>Online-B</td>
<td>90.9</td>
</tr>
<tr>
<td></td>
<td>73.5</td>
<td>-1.100</td>
<td>Online-G</td>
<td>73.5</td>
</tr>
</tbody>
</table>

Ave. Ave. z System

We produced the best constrained systems (tied) for both directions.

Human Evaluation

Our submissions are in bold. Systems within a cluster are considered tied.