OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation

Shantipriya Parida¹, Satya Ranjan Dash², Ondřej Bojar³, Petr Motlicek¹, Priyanka Pattnaik², Debasish Kumar Mallick²,

¹Idiap Research Institute, Martigny, Switzerland
²KIIT University, Odisha, India
³Charles University, Prague, Czech Republic
Agenda

● Overview
● Data Sources
● Data Processing
● Final Data Size and Domain Coverage
● Baseline
● Availability
● Conclusion
Overview

● Odia is an Indian language belonging to the Indo-Aryan branch of the Indo-European language family.

● Odia is one of 22 official languages of India and sixth Indian language to be designated as a Classical language.

● There is a demand for English↔Odia machine translation system.

● There is lack of Odia resources, particularly parallel corpora.

● Existing few English-Odia corpora are small in size, cover few domains not very suitable for machine translation, which motivates us for OdiEnCorp 2.0.
Data Sources

- Data extracted from other online resources.
- Data extracted from Odia Wikipedia.
- Data extracted using Optical Character Recognition (OCR).
- Data reused from existing corpora.

Block diagram of the Corpus building process
Data Processing

- Extraction of plain text.
  - Python script to scrape plain text from HTML page.
- Manual processing.
  - Correction of noisy text extracted using OCR-based approach.
- Sentence segmentation.
  - Paragraph segmented into sentences based on English full stop (.) and Odia Danda (|) or Purnaviram.
- Sentence alignment.
  - Manual sentence alignment for Odia Wikipedia articles where text in two language are independent of each other.
Final Datasize and Domain Coverage

- The composition of OdiEnCorp 2.0 with statistics for individual sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sentences</th>
<th>Tokens</th>
<th>Book Name and Author</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Parallel)</td>
</tr>
<tr>
<td>Wikipedia Dump</td>
<td>5796</td>
<td>38249</td>
<td>37944</td>
</tr>
<tr>
<td>Glosbe Website</td>
<td>6222</td>
<td>40143</td>
<td>38248</td>
</tr>
<tr>
<td>Odisha District Website</td>
<td>761</td>
<td>15227</td>
<td>13132</td>
</tr>
<tr>
<td>TamilCube Website</td>
<td>4434</td>
<td>7180</td>
<td>6776</td>
</tr>
<tr>
<td>OCR (Book 1)</td>
<td>356</td>
<td>4825</td>
<td>3909</td>
</tr>
<tr>
<td>OCR (Book 2)</td>
<td>9499</td>
<td>117454</td>
<td>102279</td>
</tr>
<tr>
<td>OCR (Book 3)</td>
<td>775</td>
<td>13936</td>
<td>12068</td>
</tr>
<tr>
<td>OCR (Book 4)</td>
<td>1211</td>
<td>1688</td>
<td>1652</td>
</tr>
<tr>
<td>OCR (Book 5)</td>
<td>293</td>
<td>1492</td>
<td>1471</td>
</tr>
<tr>
<td>Odia Virtual Academy (OVA)</td>
<td>1021</td>
<td>4297</td>
<td>3653</td>
</tr>
<tr>
<td>PMIndia</td>
<td>38588</td>
<td>690634</td>
<td>607611</td>
</tr>
<tr>
<td>OdiEnCorp 1.0</td>
<td>29346</td>
<td>756967</td>
<td>648028</td>
</tr>
<tr>
<td>Total</td>
<td>98302</td>
<td>1692092</td>
<td>1476768</td>
</tr>
</tbody>
</table>

OdiEnCorp 2.0 parallel corpus details. Training, dev and test sets together
Baseline (Neural Machine Translation)

● Dataset
  ○ Removed duplicated sentence pairs and shuffled.

● NMT Setup
  ○ We used Transformer model as implemented in OpenNMT-py.
  ○ Generated vocabulary of 32K sub-word type jointly for source and target language.
  ○ Train using single GPU (learning rate: 0.2, 8000 warm-up steps).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Sentences</th>
<th>EN</th>
<th>OD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train 2.0</td>
<td>69260</td>
<td>1340371</td>
<td>1164636</td>
</tr>
<tr>
<td>Dev 2.0</td>
<td>13429</td>
<td>157951</td>
<td>140384</td>
</tr>
<tr>
<td>Test 2.0</td>
<td>14163</td>
<td>185957</td>
<td>164532</td>
</tr>
</tbody>
</table>

OdiEnCorp 2.0 processed for NMT experiments.
Result

<table>
<thead>
<tr>
<th>Training Corpus</th>
<th>Task</th>
<th>sacreBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>OdiEnCorp 2.0</td>
<td>EN-OD</td>
<td>5.4</td>
</tr>
<tr>
<td>OdiEnCorp 2.0</td>
<td>OD-EN</td>
<td>9.2</td>
</tr>
</tbody>
</table>

Results for baseline NMT on Dev and Test sets for OdiEnCorp 2.0.

Availability

OdiEnCorp 2.0 is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, CC-BY-NC-SA at:

http://hdl.handle.net/11234/1-3211
Conclusion

- The corpus will be used for low resource machine translation shared tasks. The first such task is *Workshop on Asian Translation (WAT 2020)* Indic shared task on *Odia↔English* machine translation.
- Extending *OdiEnCorp 2.0* with more parallel data, again by finding various new sources.
- Building an English ↔ Odia translation system by:
  - Utilizing the developed *OdiEnCorp 2.0* corpus.
  - Other techniques (back translation, domain adaptation)
  - Releasing it to users for non-commercial purposes.
Any Questions ?
Thank You