IDIAP RESEARCH REPORT

# ODIENCORP 2.0: ODIA-ENGLISH PARALLEL CORPUS FOR MACHINE TRANSLATION

Shantipriya Parida[a] Satya Ranjan Dash
Ondrej Bojar Petr Motlicek Priyanka Pattnaik
Debasish Kumar Mallick

Idiap-RR-08-2020

MAY 2020

---

[a]Idiap Research Institute

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

# OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation

**Shantipriya Parida**[1] **Satya Ranjan Dash**[2] **Ondřej Bojar**[3*]
**Petr Motlíček**[1] **Priyanka Pattnaik**[2] **Debasish Kumar Mallick**[2]
[1]Idiap Research Institute, Martigny, Switzerland
{shantipriya.parida, petr.motlicek}@idiap.ch
[2]KIIT University, Bhubaneswar, India
sdashfca@kiit.ac.in, priyankapattanaik2013@gmail.com,
mdebasishkumar@gmail.com
[3]Charles University, Prague, Czech Republic
bojar@ufal.mff.cuni.cz

## Abstract

The preparation of parallel corpora is a challenging task, particularly for languages that suffer from under-representation in the digital world. In a multi-lingual country like India, the need for such parallel corpora is stringent for several low-resource languages. In this work, we provide an extended English-Odia parallel corpus, OdiEnCorp 2.0, aiming particularly at Neural Machine Translation (NMT) systems which will help translate English↔Odia. OdiEnCorp 2.0 includes existing English-Odia corpora and we extended the collection by several other methods of data acquisition: parallel data scraping from many websites, including Odia Wikipedia, but also optical character recognition (OCR) to extract parallel data from scanned images. Our OCR-based data extraction approach for building a parallel corpus is suitable for other low resource languages that lack in online content. The resulting OdiEnCorp 2.0 contains 98,302 sentences and 1.69 million English and 1.47 million Odia tokens. To the best of our knowledge, OdiEnCorp 2.0 is the largest Odia-English parallel corpus covering different domains and available freely for non-commercial and research purposes.

**Keywords:** Parallel Corpus, Machine Translation (MT), Optical Character Recognition (OCR)

## 1. Introduction

Odia (also called Oriya) is an Indian language belonging to the Indo-Aryan branch of the Indo-European language family. It is the predominant language of the Indian state of Odisha. Odia is one of the 22 official languages and 14 regional languages of India. Odia is the sixth Indian language to be designated a Classical Language in India based on having a long literary history and not having borrowed extensively from other languages.[1] Odia is written in Odia script, which is a Brahmic script. Odia has its origins pinned to the 10th century. In the 16th and 17th centuries, as in the case of other Indian languages, Odia too suffered changes due to the influence of Sanskrit.[2] Odia is nowadays spoken by 50 million speakers.[3] It is heavily influenced by the Dravidian languages as well as Arabic, Persian, English. Odias inflectional morphology is rich with a three-tier tense system. The prototypical word order is subject-object-verb (SOV).

In today's digital world, there has been a demand for machine translation systems for English↔Odia translation for a long time which couldn't have been fulfilled due to the lack of Odia resources, particularly a parallel corpus. Parallel corpora are of great importance in language studies, teaching and many natural language processing applications such as machine translation, cross-language information retrieval, word sense disambiguation, bilingual terminology extraction as well as induction of tools across languages. The Odia language is not available in many machine translation systems. Several researchers explored these goals, developing Odia resources and prototype machine translation systems but these are not available online and benefitting users (Das et al., 2018; Balabantaray and Sahoo, 2013; Rautaray et al., 2019).

We have analysed the available English-Odia parallel corpora (OdiEnCorp 1.0, PMIndia) and their performance (BLEU score) for machine translation (Parida et al., 2020; Haddow and Kirefu, 2020). OdiEnCorp 1.0 contains Odia-English parallel and monolingual data. The statistics of OdiEnCorp 1.0 are shown in Table 1. In OdiEnCorp 1.0, the parallel sentences are mostly derived from the English-Odia parallel Bible and the size of the parallel corpus (29K) is not sufficient for neural machine translation (NMT) as documented by the baseline results (Parida et al., 2020) as well as attempts at improving them using NMT techniques such as transfer learning (Kocmi and Bojar, 2019).

The recently released PMIndia corpus (Haddow and Kirefu, 2020) contains 38K English-Odia parallel sentences but it is mostly collected from the prime minister of India's official portal[4] containing text about government policies in 13 official languages of India.

These points motivate us for building OdiEnCorp 2.0 with more data, covering various domains suitable for various tasks of language processing, but particularly for the building of an English↔Odia machine translation system which will be useful for the research community as well as general users for non-commercial purposes.

---

* Corresponding author

[1] `https://infogalactic.com/info/Odia_language`

[2] `https://www.indianmirror.com/languages/odiya-language.html`

[3] `https://www.britannica.com/topic/Oriya-language`

[4] `https://www.pmindia.gov.in/en/`

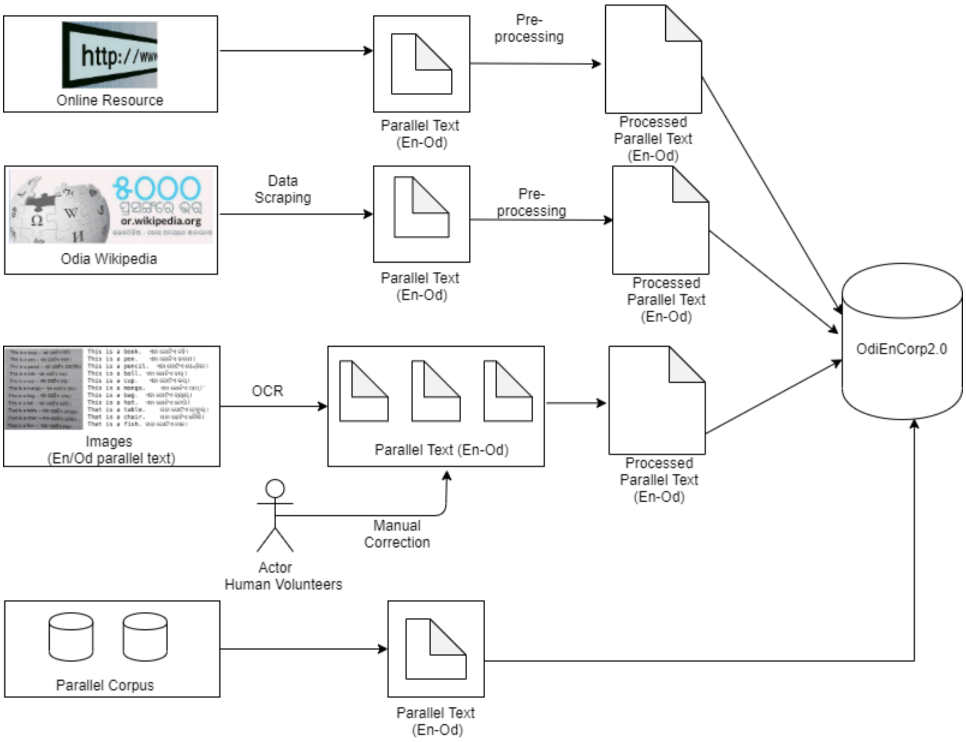


Figure 1: Block diagram of the Corpus building process. The parallel data collected from various sources (online/offline) and processed using both automatic and manual processing to build the final Corpus OdiEnCorp 2.0.

| Source | Sentences | Tokens | |
|---|---|---|---|
| | (Parallel) | English | Odia |
| English-Odia Parallel Bible | 29069 | 756861 | 640157 |
| Odisha Government Portal | 122 | 1044 | 930 |
| Odisha Govt Home Department Portal | 82 | 367 | 327 |
| Odia Digital Library (Odia Bibhaba) | 393 | 7524 | 6233 |
| Odia Digital Library (Odia Virtual Academy) | 31 | 453 | 378 |
| Total | 29697 | 766249 | 648025 |

Table 1: Statistics of OdiEnCorp 1.0.

## 2. Data Sources

As there is a very limited number of online Odia resources available, we have explored several possible ways to collect Odia-English parallel data. Although these methods need a considerable amount of manual processing, we opted for them, to achieve the largest possible data size. In sum, we used these sources:

- Data extracted using OCR,
- Data extracted from Odia Wikipedia,
- Data extracted from other online resources,
- Data reused from existing corpora.

The overall process of the OdiEnCorp 2.0 is shown in Figure 1.

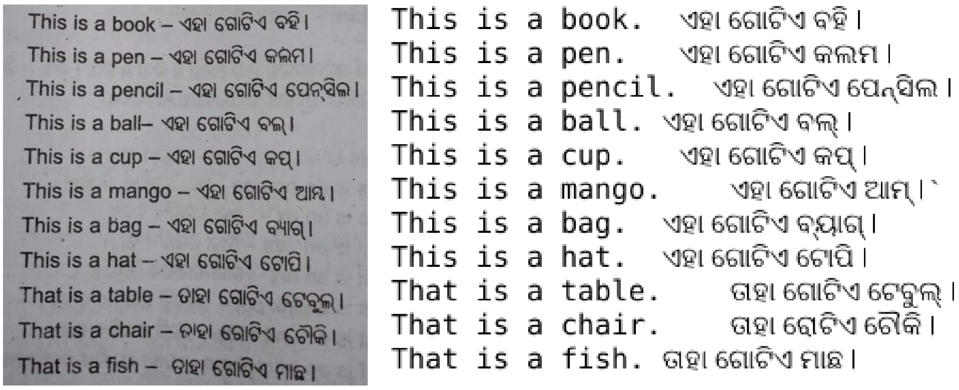
### 2.1. OCR-Based Text Extraction

Many books are translated in more than one language and they could serve as a reliable source to obtain parallel sentences, but they are unfortunately not digitized (Bakliwal et al., 2016; Premjith et al., 2016). OCR technology has improved substantially, which has allowed for large-scale digitization of textual resources such as books, old newspapers, ancient hand-written documents (Dhondt et al., 2017). That said, it should be kept in mind that there are often mistakes in the scanned texts as OCR system occasionally misrecognizes letters or falsely identifies text regions, leading to misspellings and linguistics errors in the output text (Afli et al., 2016).

Odia language has a rich literary heritage and many books are available in printed form. We have explored books having either English and Odia parallel text together or books having both versions (English and Odia). We have used the study, translation, grammar, literature, and motivational books for this purpose, obtaining the source images either from the web, or directly scanning them ourselves.

We start with the image containing the Odia language text represented in the RGB color space. For the Odia text recognition, we use the “Tesseract OCR engine” (Smith, 2007) with several improvements in the pre-processing phase.

First, we move from the traditional method which converts RGB to grayscale by taking the simple average of the three channels. We convert the RGB image into a grayscale image



This is a book – ଏହା ଗୋଟିଏ ବହି ।
This is a pen – ଏହା ଗୋଟିଏ କଲମ ।
This is a pencil – ଏହା ଗୋଟିଏ ପେନ୍ସିଲ ।
This is a ball– ଏହା ଗୋଟିଏ ବଲ୍ ।
This is a cup – ଏହା ଗୋଟିଏ କପ୍ ।
This is a mango – ଏହା ଗୋଟିଏ ଆମ୍ବ ।
This is a bag – ଏହା ଗୋଟିଏ ବ୍ୟାଗ୍ ।
This is a hat – ଏହା ଗୋଟିଏ ଟୋପି ।
That is a table – ତାହା ଗୋଟିଏ ଟେବୁଲ୍ ।
That is a chair – ତାହା ଗୋଟିଏ ଚୌକି ।
That is a fish – ତାହା ଗୋଟିଏ ମାଛ ।

(a) Sample scanned image of parallel (English-Odia) data.

This is a book. ଏହା ଗୋଟିଏ ବହି ।
This is a pen. ଏହା ଗୋଟିଏ କଲମ ।
This is a pencil. ଏହା ଗୋଟିଏ ପେନ୍ସିଲ ।
This is a ball. ଏହା ଗୋଟିଏ ବଲ୍ ।
This is a cup. ଏହା ଗୋଟିଏ କପ୍ ।
This is a mango. ଏହା ଗୋଟିଏ ଆମ୍ ।`
This is a bag. ଏହା ଗୋଟିଏ ବ୍ୟାଗ୍ ।
This is a hat. ଏହା ଗୋଟିଏ ଟୋପି ।
That is a table. ତାହା ଗୋଟିଏ ଟେବୁଲ୍ ।
That is a chair. ତାହା ରୋଟିଏ ଚୌକି ।
That is a fish. ତାହା ଗୋଟିଏ ମାଛ ।

(b) Extracted parallel data.

Figure 2: An illustration of the scanned image containing parallel English-Odia data and extracted data.

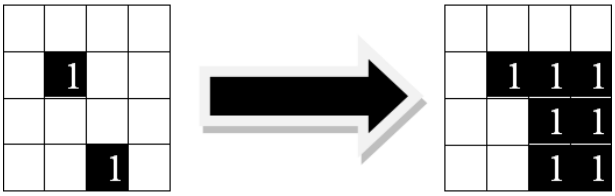


Figure 3: Dilation

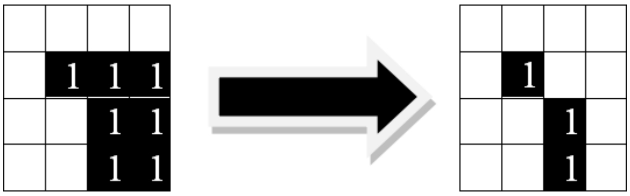


Figure 4: Erosion

by applying the luminosity method which also averages the values, but it takes a weighted average to account for human perception (Joshi, 2019):

$$\text{Grayscale} = 0.299R + 0.587G + 0.114B \tag{1}$$

where $R$ is the amount of red, $G$ green and $B$ blue color in a pixel.

To change the image further to black and white only, "Tesseract" uses the traditional binarization algorithm called "Otsu". We use instead the "Niblack and Sauvola threshold algorithm" which we found to give better results. The advantage of the Niblack algorithm is that it slides a rectangular window through the image (Smith, 2007). The center pixel threshold $T$ is derived from the mean $m$ and variance $s$ values inside the window.

$$T = m + k \cdot s, \tag{2}$$

where $k$ is a constant set to 0.8.
"Niblack" can create noise in some areas of the image, so we further improve it by including the "Sauvola" algorithm. Thus the modified formula is:

$$T = m \cdot \big(1 - k \cdot (1 - \frac{s}{R})\big), \tag{3}$$

where $R$ is the dynamics of standard deviation, a constant set to 128.
This formula will not detect all the images of documents. So the normalized formula we have implemented is:

$$T = m - k \cdot (1 - \frac{s}{R}) \cdot (m - M), \tag{4}$$

where $R$ is the maximum standard deviation of all the windows and $M$ is the gray level of the current image.

However, some black pixels vanish during these processes which may lead to erroneous character recognition, so we use Dilation (Gaikwad and Mahender, 2016) to join areas which got accidentally disconnected, see Figure 3 for an illustration.
Because dilation sometimes produces too many black pixels, we further apply "Erosion" (Alginahi, 2010) as illustrated in Figure 4.
Finally, Figure 2 illustrates a sample of the scanned image containing parallel Odia-English data and the extracted text.

## 2.2. Odia Wikipedia

The Odia Wikipedia started in 2002 and serves as a good source for Odia-English parallel data. The following steps were performed to obtain parallel sentences from Odia Wikipedia, with more details provided in the sections below:

1. Collect Wikipedia dump (20th October 2019) for the language pair Odia-English.
2. Clean the text by removing references, URLs, instructions, or any unnecessary contents.
3. Segment articles into sentences, relying on English/Odia full stop mark.
4. Align sentences between Odia and English.

| Source | Sentences | Tokens | | Book Name and Author |
|---|---|---|---|---|
| | | English | Odia | (Parallel) |
| Wikipedia Dump | 5796 | 38249 | 37944 | - |
| Glosbe Website | 6222 | 40143 | 38248 | - |
| Odisha District Website | 761 | 15227 | 13132 | - |
| TamilCube Website | 4434 | 7180 | 6776 | - |
| OCR (Book 1) | 356 | 4825 | 3909 | A Tiger at Twilight by Manoj Dash |
| OCR (Book 2) | 9499 | 117454 | 102279 | Yajnaseni by Prativa Ray |
| OCR (Book 3) | 775 | 13936 | 12068 | Wings of Fire by APJ Abdul Kalam with Arun Tiwari |
| OCR (Book 4) | 1211 | 1688 | 1652 | Word Book by Shibashis Kar and Shreenath Chaterjee |
| OCR (Book 5) | 293 | 1492 | 1471 | Spoken English by Partha Sarathi Panda and Prakhita Padhi |
| Odia Virtual Academy (OVA) | 1021 | 4297 | 3653 | Sarala (Tribhasi) Bhasa Sikhana Petika |
| PMIndia | 38588 | 690634 | 607611 | - |
| OdiEnCorp 1.0 | 29346 | 756967 | 648025 | - |
| Total | 98302 | 1692092 | 1476768 | |

Table 2: OdiEnCorp 2.0 parallel corpus details. Training, dev and test sets together.

### 2.3. Additional Online Resources

Finding potential parallel texts in a collection of web documents is a challenging task, see e.g. (Antonova and Misyurev, 2011; Kúdela et al., 2017; Schwenk, 2018; Artetxe and Schwenk, 2019).

We have explored websites and prepared a list of such websites which are potential for us to collect Odia-English parallel data. The websites were then crawled with a simple Python script.

We found Odisha's government portals of each district (e.g. Nayagarh district[5]) of Odisha containing general information about the district in both English and Odia version. Analyzing extracted text, we found a few cases where English was repeated in both sides of the website. We have aligned the extracted text manually to obtain the parallel text.

We also extracted parallel data from the Odia digital library "Odia Virtual Academy",[6] an Odisha government-initiated portal to store treasures of Odia language and literature for seamless access to Odia people staying across the globe. The web page provides tri-lingual books (tribal dictionary[7] containing common words and their translations in English and Odia) and we extracted the English-Odia sentence pairs from it.

### 2.4. Reusing Available Corpora

Finally, we included parallel data from OdiEnCorp 1.0 and PMIndia (Parida et al., 2020; Haddow and Kirefu, 2020). Both corpora contain pre-processed English-Odia parallel sentences. The statistics of these corpora are available in Table 2.

## 3. Data Processing

The data collected from different sources were processed to achieve a unified format.

### 3.1. Extraction of Plain Text

When utilizing online resources, we used a Python script to scrape plain text from HTML pages.

### 3.2. Manual Processing

After analyzing the raw data extracted using the OCR-based approach, we found a few errors such as unnecessary characters and symbols, missing words, etc. One of the reasons of poor OCR performance was the fact that some images were taken using mobile phones. In later processing, we always used a proper scanner.

We decided for manual correction of such entries by volunteers whose mother tongue is Odia. Although this task is time-consuming and tedious, the result should be of considerably better quality and much more suitable for machine translation and other NLP tasks.

Four volunteers worked part-time (2-3 hours daily) for four months on scanning of books, extracting data from the scanned images using OCR techniques, collecting data from online as well as offline sources, and post-editing all the data collected from different sources.

### 3.3. Sentence Segmentation

All sources that come in paragraphs (e.g. Wikipedia articles or books) had to be segmented into sentences. We considered full stop (.) as of the end of the sentence for English and Odia Danda or Purnaviram (|) as of the end of the sentence for Odia language.

### 3.4. Sentence Alignment

For some sources, the alignment between English and Odia sentences was straightforward. Sources like Odia Wikipedia posed a bigger challenge, because the texts in the two languages are often created or edited independently of each other.

To achieve the best possible parallel corpus, we relied on manual sentence alignment. In this process, we had to truncate or remove several few sentences in either of the languages in order to reach exactly 1-1 aligned English-Odia sentence pairs.

### 3.5. Domain Coverage

The resulting corpus OdiEnCorp 2.0 covers a wide variety of domains, esp. compared to similar corpora. Our corpus covers the bible, literature, government policies, daily usage, learning, general domain (Wikipedia).

---

[5] `https://nayagarh.nic.in`

[6] `https://ova.gov.in/en/`

[7] `https://ova.gov.in/de/odisha-tribal-dictionary-and-language/`

|  |  | #Tokens |  |
|---|---|---|---|
| Dataset | #Sentences | EN | OD |
| Train 2.0 | 69260 | 1340371 | 1164636 |
| Dev 2.0 | 13429 | 157951 | 140384 |
| Test 2.0 | 14163 | 185957 | 164532 |

Table 3: OdiEnCorp 2.0 *processed for NMT experiments*.

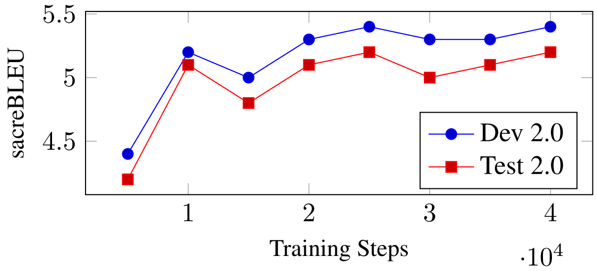


Figure 5: Learning curve (EN→OD)

## 4. Final Data Sizes

The composition of OdiEnCorp 2.0 with statistics for individual sources is provided in Table 2.

The release designates which parts of the corpus should be used for training, which for development and which for final testing. This division of OdiEnCorp 2.0 respects the dev and test sets of OdiEnCorp 1.0, so that models trained on v.2.0 training set can be directly tested on the older v.1.0 dev and test sets.

## 5. Baseline Neural Machine Translation

For future reference, we provide a very baseline experiment with neural machine translation using OdiEnCorp 2.0 data.

### 5.1. Dataset Description

For the purpose of NMT training, we removed duplicated sentence pairs and shuffled the segments The training, dev and test set sizes after this processing are shown in Table 3.

### 5.2. Neural Machine Translation Setup

We used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017).[8] Subword units were constructed using the word pieces algorithm (Johnson et al., 2017). Tokenization is handled automatically as part of the pre-processing pipeline of word pieces.

We generated the vocabulary of 32k sub-word types jointly for both the source and target languages, sharing it between the encoder and decoder. To train the model, we used a single GPU and followed the standard "Noam" learning rate decay,[9] see (Vaswani et al., 2017) or (Popel and Bojar, 2018) for more details. Our starting learning rate was 0.2 and we used 8000 warm-up steps. The learning curves are shown in Figure 5 and Figure 6,

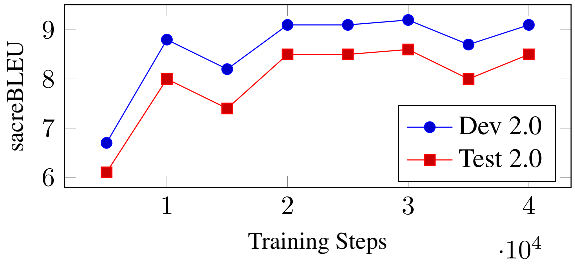


Figure 6: Learning curve (OD→EN)

|  |  | sacreBLEU |  |
|---|---|---|---|
| Training Corpus | Task | Dev 2.0 | Test 2.0 |
| OdiEnCorp 2.0 | EN-OD | 5.4 | 5.2 |
| OdiEnCorp 2.0 | OD-EN | 9.2 | 8.6 |

Table 4: Results for baseline NMT on Dev and Test sets for OdiEnCorp 2.0.

### 5.3. Results

We use sacreBLEU[10,11] for estimating translation quality. Based on the Dev 2.0 best score, we select the model at iteration 40k for EN→OD and at 30k for OD→EN to obtain the final test set scores.

Table 4 reports the performance on the Dev and Test sets of OdiEnCorp 2.0. Table 5 uses the Dev and Test sets belonging to OdiEnCorp 1.0. The results in Table 5 thus allow us to observe the gains compared to the scores reported in Parida et al. (2020).

## 6. 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[12] at:

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

## 7. Conclusion and Future Work

We presented OdiEnCorp 2.0, an updated version of Odia-English parallel corpus aimed for linguistic research and applications in natural language processing, primarily machine translation.

The corpus will be used for low resource machine translation shared tasks. The first such task is WAT 2020[13] Indic shared task on Odia↔English machine translation.

Our plans for future include:

- Extending OdiEnCorp 2.0 with more parallel data, again by finding various new sources.
- Building an English↔Odia translation system utilizing the developed OdiEnCorp 2.0 corpus and other techniques (back translation, domain adaptation) and releasing it to users for non-commercial purposes.

---

[8] `http://opennmt.net/OpenNMT-py/quickstart.html`

[9] `https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html`

[10] `https://github.com/mjpost/sacreBLEU`

[11] Signature: BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.3

[12] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

[13] `https://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html`

| | | sacreBLEU | |
|---|---|---|---|
| Training Corpus | Task | Dev 1.0 | Test 1.0 |
| OdiEnCorp 1.0 | EN-OD | 4.3 | 4.1 |
| OdiEnCorp 2.0 | EN-OD | 4.9 | 4.0 |
| OdiEnCorp 1.0 | OD-EN | 9.4 | 8.6 |
| OdiEnCorp 2.0 | OD-EN | 12.0 | 9.3 |

Table 5: Scores on Dev and Test sets of OdiEnCorp 1.0 for the baseline NMT models trained on OdiEnCorp 1.0 vs. OdiEnCorp 2.0.

- Promoting the corpus in other reputed machine translation campaigns focusing on low resource languages.

## 8. Acknowledgements

The work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German (Title: SM2: Extracting Semantic Meaning from Spoken Material funding application no. 29814.1 IP-ICT). In part, it was also supported by the EU H2020 project "Real-time network, text, and speaker analytics for combating organized crime" (ROXANNE, grant agreement: 833635) and by the grant 18-24210S of the Czech Science Foundation.

This work has been using language resources and tools stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

Afli, H., Barrault, L., and Schwenk, H. (2016). Ocr error correction using statistical machine translation. *Int. J. Comput. Linguistics Appl.*, 7(1):175–191.

Alginahi, Y. (2010). Preprocessing techniques in character recognition. *Character recognition*, 1:1–19.

Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597610, Mar.

Bakliwal, P., Devadath, V., and Jawahar, C. (2016). Align me: A framework to generate parallel corpus using ocrs and bilingual dictionaries. In *Proc. of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 183–187.

Balabantaray, R. and Sahoo, D. (2013). An experiment to create parallel corpora for odia. *International Journal of Computer Applications*, 67(19).

Das, A. K., Pradhan, M., Dash, A. K., Pradhan, C., and Das, H. (2018). A constructive machine translation system for english to odia translation. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0854–0857. IEEE.

Dhondt, E., Grouin, C., and Grau, B. (2017). Generating a training corpus for ocr post-correction using encoder-decoder model. In *Proc. of IJCNLP (Volume 1: Long Papers)*, pages 1006–1014.

Gaikwad, D. K. and Mahender, C. N. (2016). A review paper on text summarization. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(3):154–160.

Haddow, B. and Kirefu, F. (2020). Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Joshi, N. (2019). Text image extraction and summarization. *Asian Journal For Convergence In Technology (AJCT)*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Kocmi, T. and Bojar, O. (2019). Transfer learning across languages from someone else's NMT model. *arXiv preprint arXiv:1909.10955*.

Kúdela, J., Holubová, I., and Bojar, O. (2017). Extracting parallel paragraphs from common crawl. *The Prague Bulletin of Mathematical Linguistics*, (107):36–59.

Parida, S., Bojar, O., and Dash, S. R. (2020). OdiEnCorp: Odia–English and Odia-Only Corpus for Machine Translation. In *Smart Intelligent Computing and Applications*, pages 495–504. Springer.

Popel, M. and Bojar, O. (2018). Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Premjith, B., Kumar, S. S., Shyam, R., Kumar, M. A., and Soman, K. (2016). A fast and efficient framework for creating parallel corpus. *Indian J. Sci. Technol*, 9:1–7.

Rautaray, J., Hota, A., and Gochhayat, S. S. (2019). A shallow parser-based hindi to odia machine translation system. In *Computational Intelligence in Data Mining*, pages 51–62. Springer.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proc. of ACL (Volume 2: Short Papers)*, pages 228–234. Association for Computational Linguistics, July.

Smith, R. (2007). An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., and Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. In *Proc. of AMTA (Volume 1: Research Papers)*, pages 193–199.