OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation

Shantipriya Parida[1], Satya Ranjan Dash[2], Ondřej Bojar[3],
Petr Motlicek[1], Priyanka Pattnaik[2], Debasish Kumar Mallick[2],

[1]Idiap Research Institute, Martigny, Switzerland
[2]KIIT University, Odisha, India
[3]Charles 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.

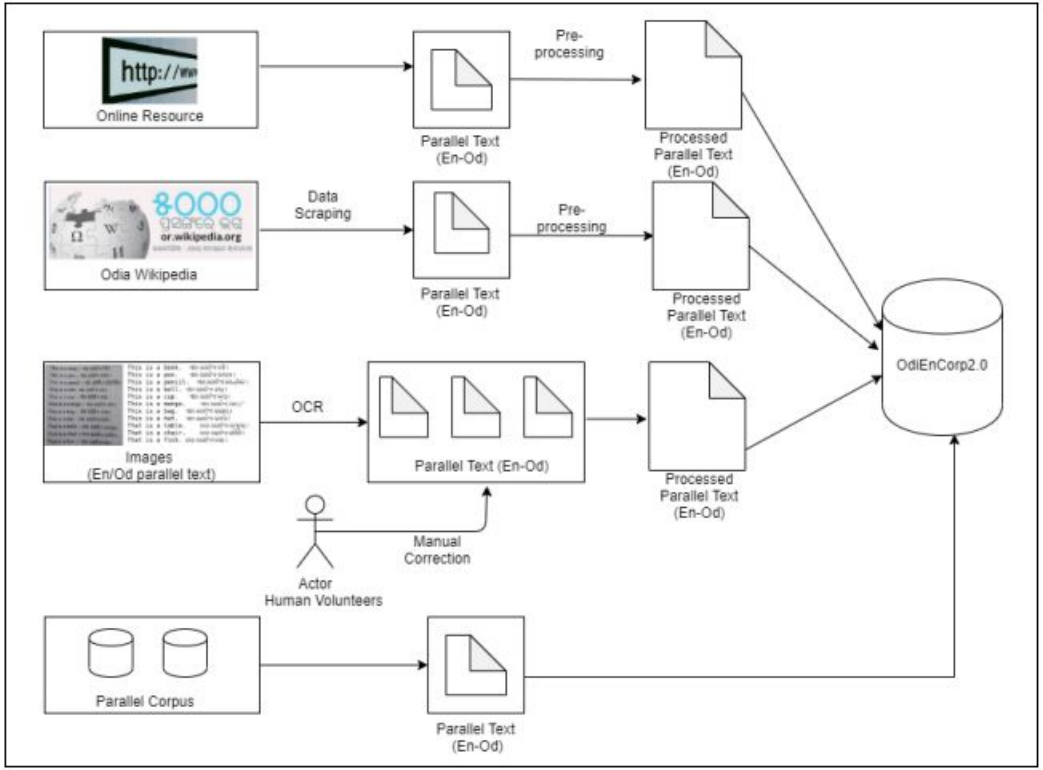
- There is a demand for English↔Odia machine translation system.

- 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.



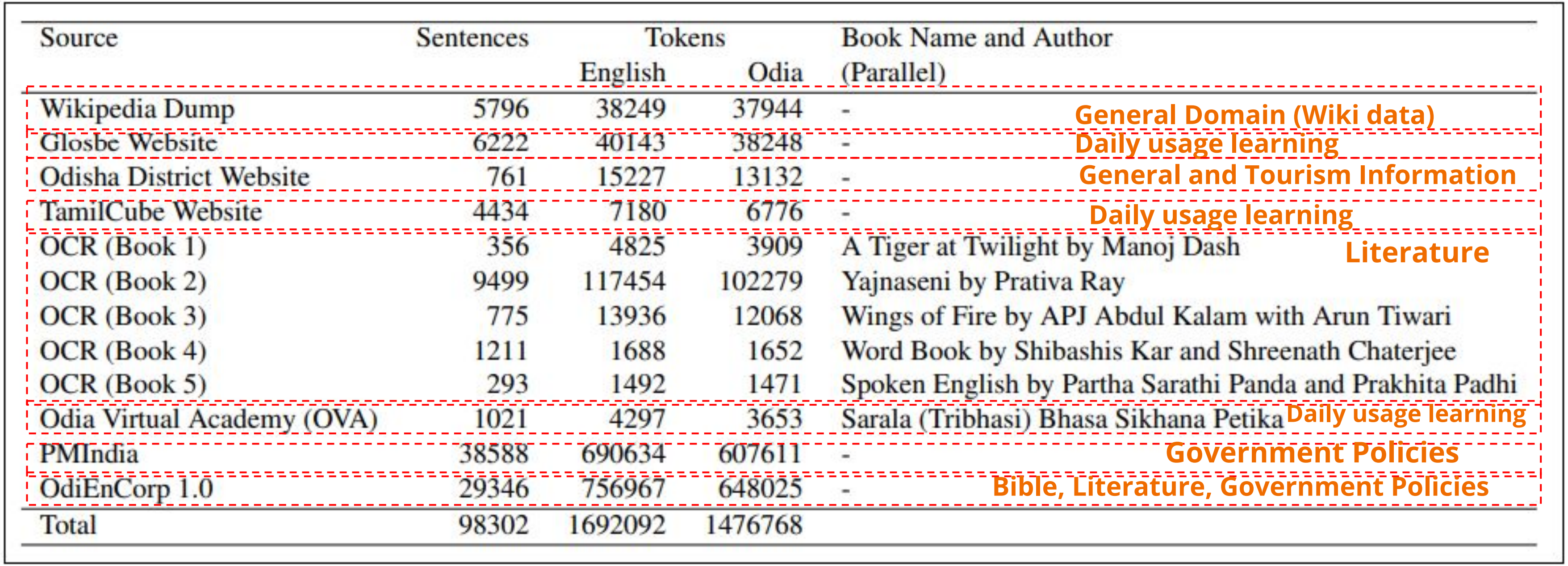
Block diagram of the Corpus building process

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.



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	

General Domain (Wiki data)

Daily usage learning

General and Tourism Information

Daily usage learning

Literature

Daily usage learning

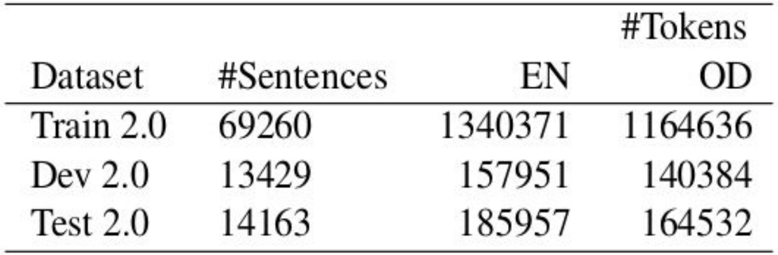
Government Policies

Bible, Literature, Government Policies

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

Baseline (Neural Machine Translation)

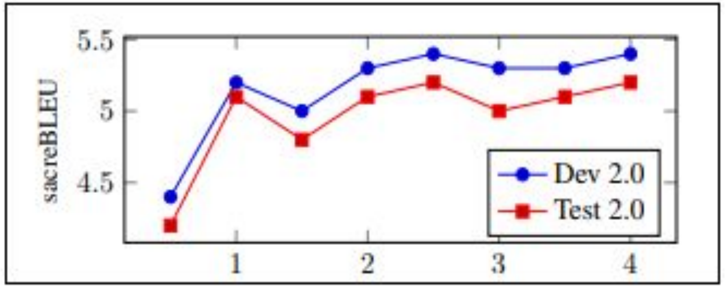
- Dataset
  - Removed duplicated sentence pairs and shuffled.



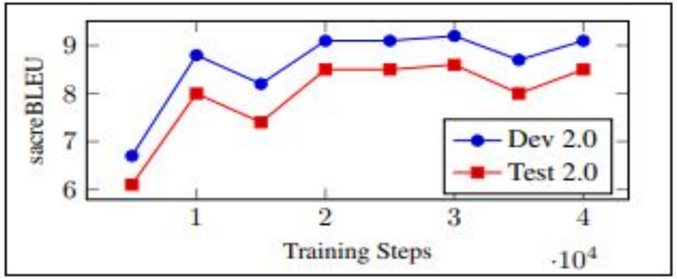
		#Tokens	
Dataset	#Sentences	EN	OD
Train 2.0	69260	1340371	1164636
Dev 2.0	13429	157951	140384
Test 2.0	14163	185957	164532

OdiEnCorp 2.0 processed for NMT experiments.

- 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).

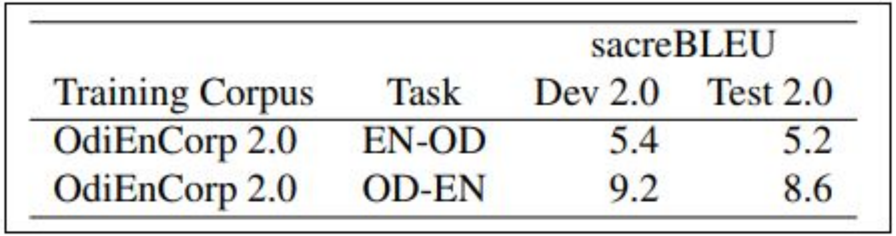


Learning Curve (EN->OD)



Learning Curve (OD->EN)

Result



		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

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