idiap
RESEARCH INSTITUTE

- **Areas of activity**: multilingual and multimodal interaction and multimedia information management, including human behavior modeling.
- **Staff**: 120+ (+50 across 16 start-ups)

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Odia Machine Learning Conference
October 04 2020

Odia Natural Language Processing Resource Development

Shantipriya Parida, Ondřej Bojar, Satya Ranjan Dash, Petr Motlicek,
Priyanka P. Pattnaik, Debasish Kumar Mallick, Biranchi Narayan Nayak,
Satya Prakash Biswal, and Amulya Ratna Dash

Idiap Research Institute
Martigny, Switzerland

Agenda

- Overview
- Current Scenario
- Need for Odia NLP Resource Development
- OdiEnCorp (Odia-English Corpus)
- Odia NLP Resource Catalog
- Conclusion

Overview

- Natural language processing (NLP) helps computers communicate with humans in their own language and scales other language-related tasks.
- NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.

Current Scenario

- Although the Odia language has a rich cultural heritage, this is not completely digitized or accessible, resulting in a lack of resources.

- In context to NLP research and development, the availability of resources are limited and not available online.

- Developing such NLP resources shown below required the attention of all
  - Language corpus,
  - Language models,
  - Dataset for
    - Summarization
    - Topic Detection
    - Named entity recognition (NER)
    - Fake news detection
    - Aggresivnes/hate speech detection
    - Codemix detection
    - Dialect detection
    - Treebank

Need for Odia NLP Resource Development

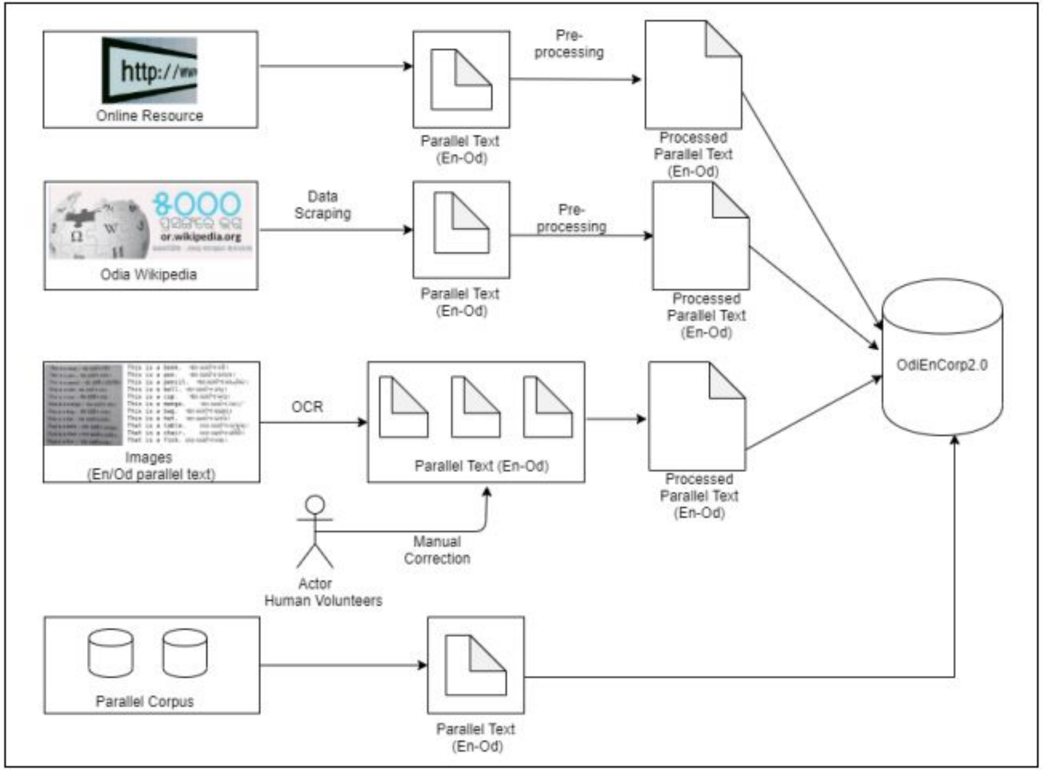
- It will help the researchers for NLP research, Government for building services accessible for common people and industry for building products localization and better customer reach.

OdiEnCorp (Odia-English Corpus)

- 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.



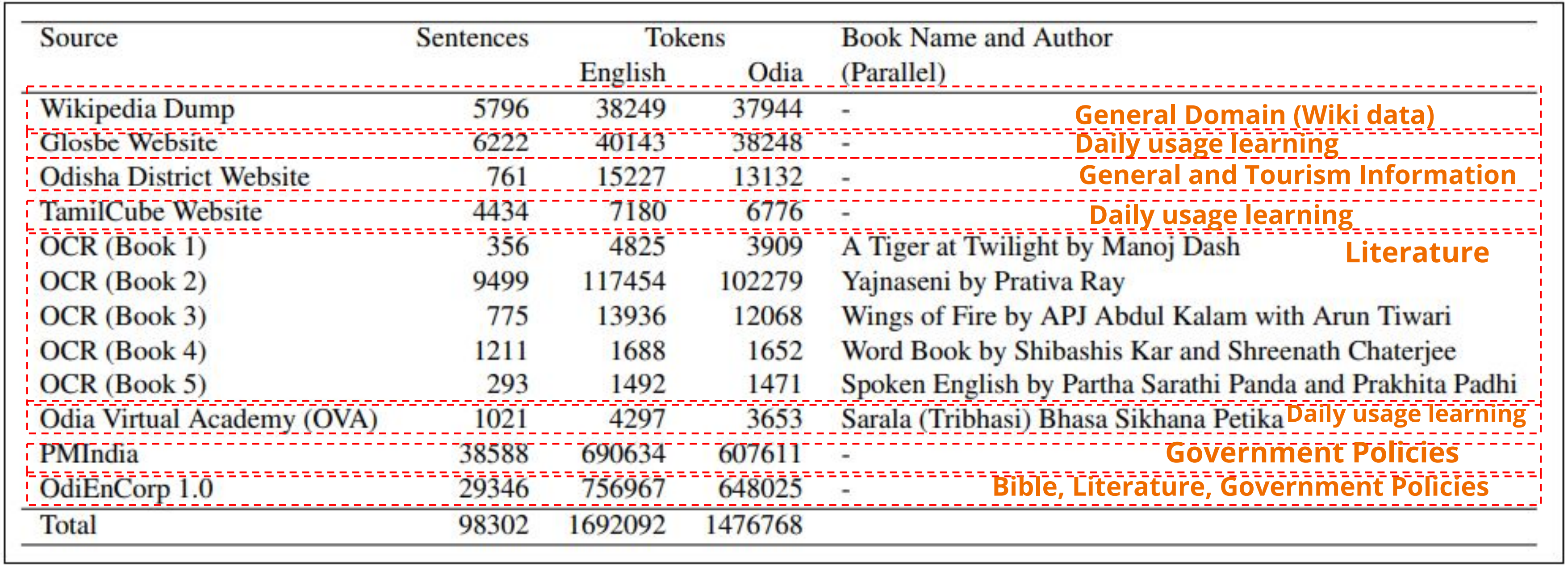
Block diagram of the Corpus building process

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.

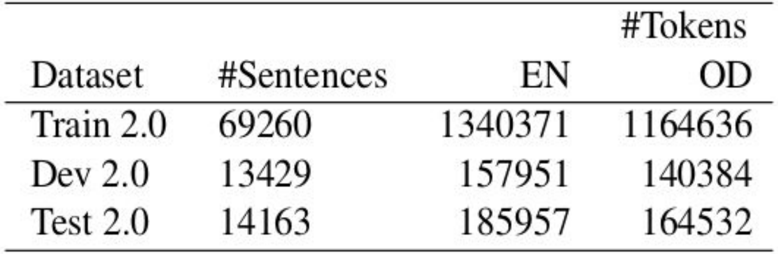


Source	Sentences	Tokens		Book Name and Author	
		English	Odia	(Parallel)	
Wikipedia Dump	5796	38249	37944	-	General Domain (Wiki data)
Glosbe Website	6222	40143	38248	-	Daily usage learning
Odisha District Website	761	15227	13132	-	General and Tourism Information
TamilCube Website	4434	7180	6776	-	Daily usage learning
OCR (Book 1)	356	4825	3909	A Tiger at Twilight by Manoj Dash	Literature
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	Daily usage learning
PMIndia	38588	690634	607611	-	Government Policies
OdiEnCorp 1.0	29346	756967	648025	-	Bible, Literature, Government Policies
Total	98302	1692092	1476768		

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

Baseline (Neural Machine Translation)

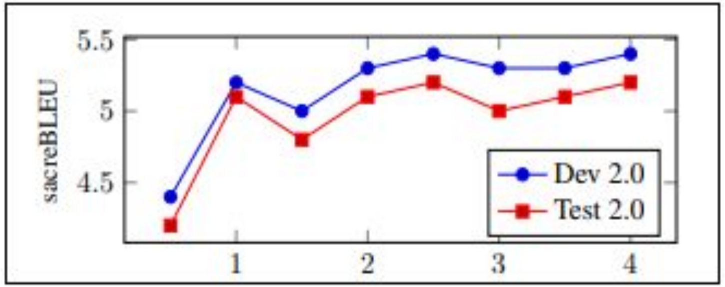
- Dataset
  - Removed duplicated sentence pairs and shuffled.



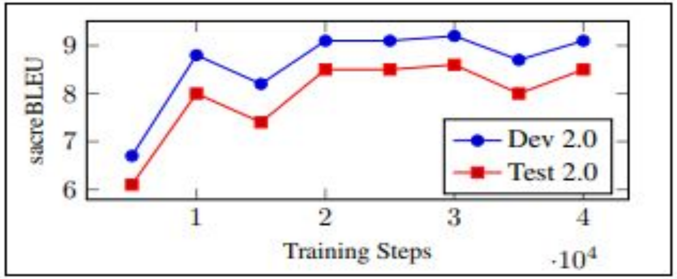
Dataset	#Sentences	#Tokens EN	 OD
Train 2.0	69260	1340371	1164636
Dev 2.0	13429	157951	140384
Test 2.0	14163	185957	164532

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).

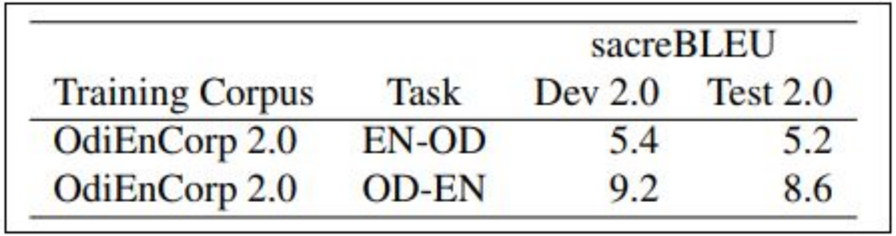


Learning Curve (EN->OD)



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.

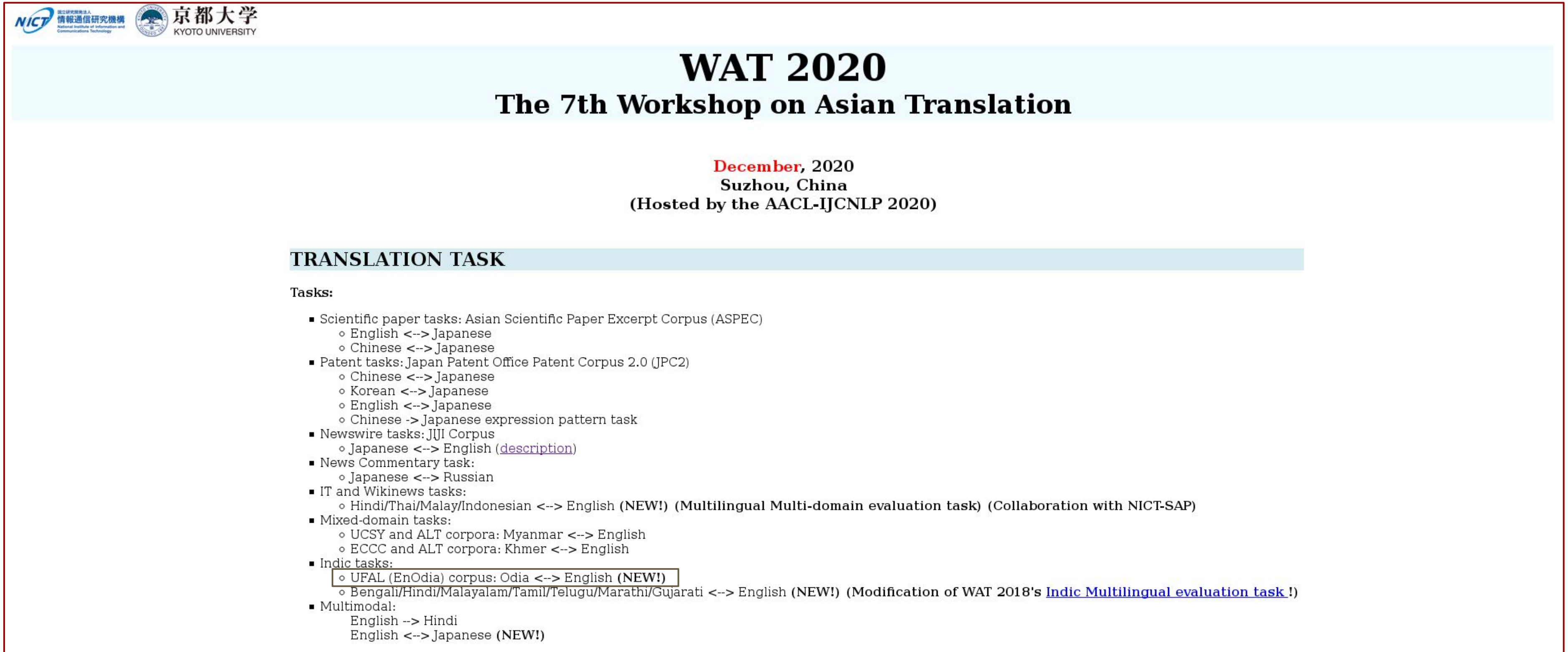
Availability

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

WAT 2020

Website: http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html



NICT 情報通信研究機構 京都大学 KYOTO UNIVERSITY

WAT 2020
The 7th Workshop on Asian Translation

December, 2020
Suzhou, China
(Hosted by the AACL-IJCNLP 2020)

TRANSLATION TASK

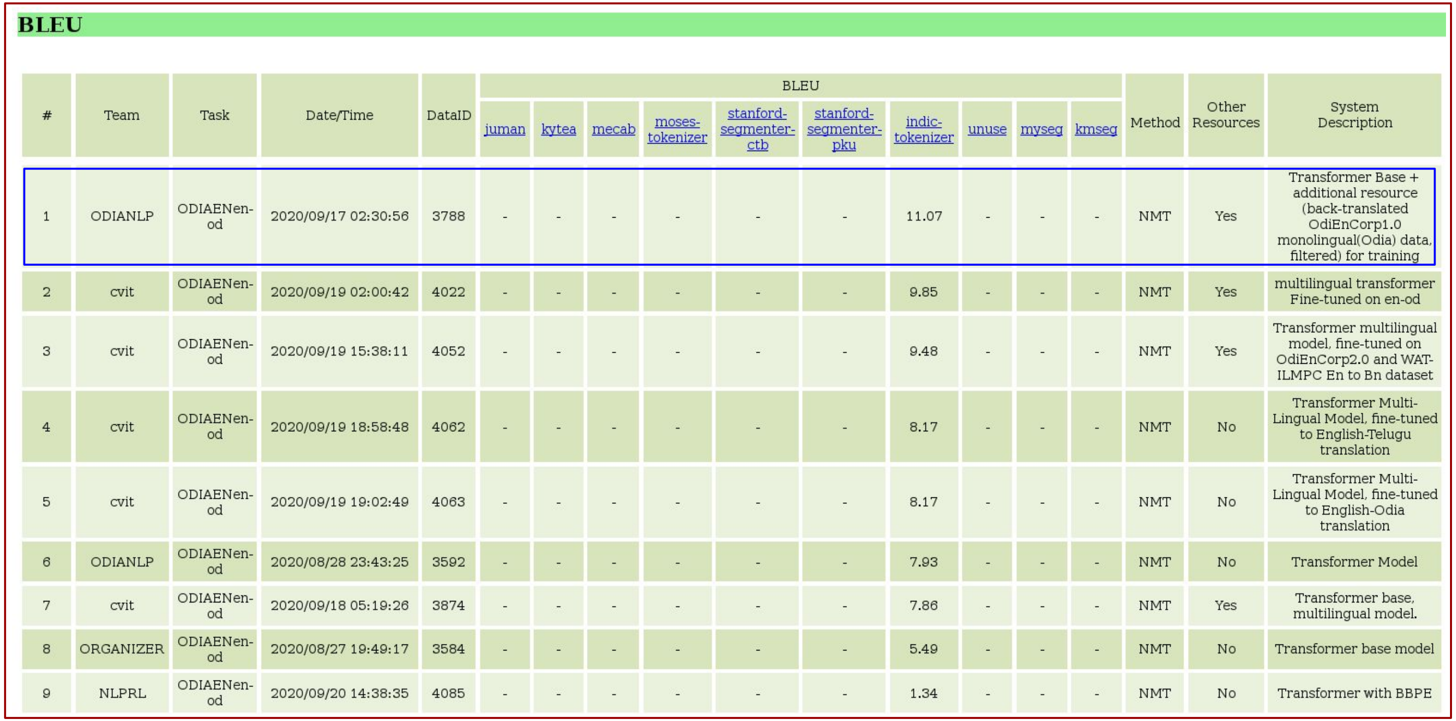
Tasks:

- Scientific paper tasks: Asian Scientific Paper Excerpt Corpus (ASPEC)
  - English <--> Japanese
  - Chinese <--> Japanese
- Patent tasks: Japan Patent Office Patent Corpus 2.0 (JPC2)
  - Chinese <--> Japanese
  - Korean <--> Japanese
  - English <--> Japanese
  - Chinese -> Japanese expression pattern task
- Newswire tasks: JIJI Corpus
  - Japanese <--> English (description)
- News Commentary task:
  - Japanese <--> Russian
- IT and Wikinews tasks:
  - Hindi/Thai/Malay/Indonesian <--> English **(NEW!)** **(Multilingual Multi-domain evaluation task)** **(Collaboration with NICT-SAP)**
- Mixed-domain tasks:
  - UCSY and ALT corpora: Myanmar <--> English
  - ECCC and ALT corpora: Khmer <--> English
- Indic tasks:
  - UFAL (EnOdia) corpus: Odia <--> English **(NEW!)**
  - Bengali/Hindi/Malayalam/Tamil/Telugu/Marathi/Gujarati <--> English **(NEW!)** **(Modification of WAT 2018's Indic Multilingual evaluation task !)**
- Multimodal:

  English --> Hindi
  English <--> Japanese **(NEW!)**

ODIANLP Team Participation at WAT 2020

English to Odia Translation Task (Automatic Evaluation)

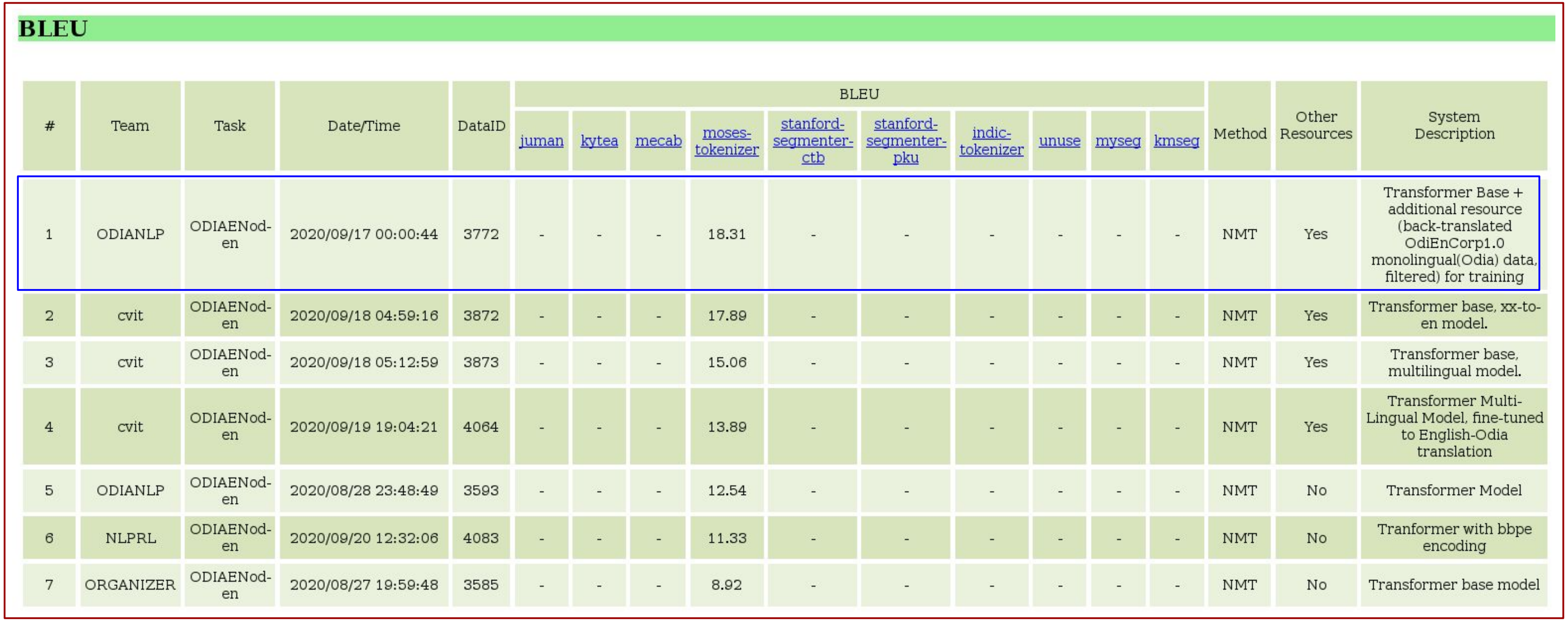


BLEU

#	Team	Task	Date/Time	DataID	BLEU										Method	Other Resources	System Description
					juman	kytea	mecab	moses-tokenizer	stanford-segmenter-ctb	stanford-segmenter-pku	indic-tokenizer	unuse	myseg	kmseg			
1	ODIANLP	ODIAEnen-od	2020/09/17 02:30:56	3788	-	-	-	-	-	-	11.07	-	-	-	NMT	Yes	Transformer Base + additional resource (back-translated OdiEnCorp1.0 monolingual(Odia) data, filtered) for training
2	cvit	ODIAEnen-od	2020/09/19 02:00:42	4022	-	-	-	-	-	-	9.85	-	-	-	NMT	Yes	multilingual transformer Fine-tuned on en-od
3	cvit	ODIAEnen-od	2020/09/19 15:38:11	4052	-	-	-	-	-	-	9.48	-	-	-	NMT	Yes	Transformer multilingual model, fine-tuned on OdiEnCorp2.0 and WAT-ILMPC En to Bn dataset
4	cvit	ODIAEnen-od	2020/09/19 18:58:48	4062	-	-	-	-	-	-	8.17	-	-	-	NMT	No	Transformer Multi-Lingual Model, fine-tuned to English-Telugu translation
5	cvit	ODIAEnen-od	2020/09/19 19:02:49	4063	-	-	-	-	-	-	8.17	-	-	-	NMT	No	Transformer Multi-Lingual Model, fine-tuned to English-Odia translation
6	ODIANLP	ODIAEnen-od	2020/08/28 23:43:25	3592	-	-	-	-	-	-	7.93	-	-	-	NMT	No	Transformer Model
7	cvit	ODIAEnen-od	2020/09/18 05:19:26	3874	-	-	-	-	-	-	7.86	-	-	-	NMT	Yes	Transformer base, multilingual model.
8	ORGANIZER	ODIAEnen-od	2020/08/27 19:49:17	3584	-	-	-	-	-	-	5.49	-	-	-	NMT	No	Transformer base model
9	NLPRL	ODIAEnen-od	2020/09/20 14:38:35	4085	-	-	-	-	-	-	1.34	-	-	-	NMT	No	Transformer with BBPE

ODIANLP Team Participation at WAT 2020

Odia to English Translation Task (Automatic Evaluation)

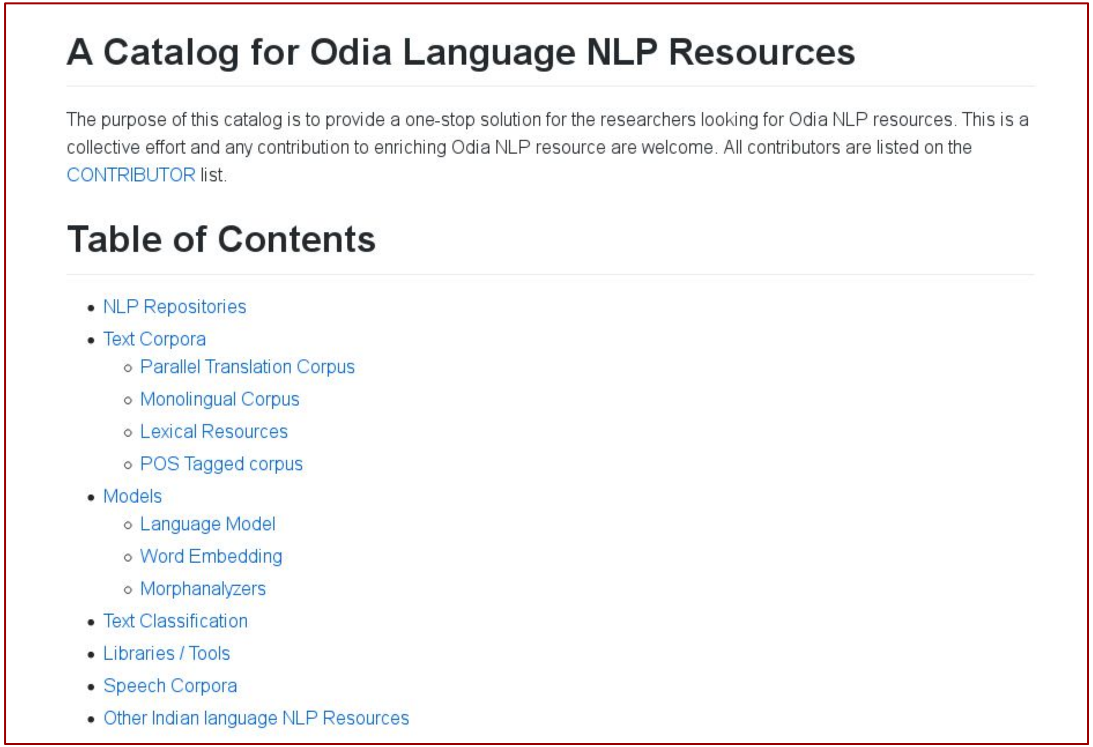


BLEU

#	Team	Task	Date/Time	DataID	BLEU										Method	Other Resources	System Description
					juman	kytea	mecab	moses-tokenizer	stanford-segmenter-ctb	stanford-segmenter-pku	indic-tokenizer	unuse	myseg	kmseg			
1	ODIANLP	ODIAENod-en	2020/09/17 00:00:44	3772	-	-	-	18.31	-	-	-	-	-	-	NMT	Yes	Transformer Base + additional resource (back-translated OdiEnCorp1.0 monolingual(Odia) data, filtered) for training
2	cvit	ODIAENod-en	2020/09/18 04:59:16	3872	-	-	-	17.89	-	-	-	-	-	-	NMT	Yes	Transformer base, xx-to-en model.
3	cvit	ODIAENod-en	2020/09/18 05:12:59	3873	-	-	-	15.06	-	-	-	-	-	-	NMT	Yes	Transformer base, multilingual model.
4	cvit	ODIAENod-en	2020/09/19 19:04:21	4064	-	-	-	13.89	-	-	-	-	-	-	NMT	Yes	Transformer Multi-Lingual Model, fine-tuned to English-Odia translation
5	ODIANLP	ODIAENod-en	2020/08/28 23:48:49	3593	-	-	-	12.54	-	-	-	-	-	-	NMT	No	Transformer Model
6	NLPRL	ODIAENod-en	2020/09/20 12:32:06	4083	-	-	-	11.33	-	-	-	-	-	-	NMT	No	Tranformer with bbpe encoding
7	ORGANIZER	ODIAENod-en	2020/08/27 19:59:48	3585	-	-	-	8.92	-	-	-	-	-	-	NMT	No	Transformer base model

Odia NLP Resource Catalog

Website: https://github.com/shantipriyap/Odia-NLP-Resource-Catalog

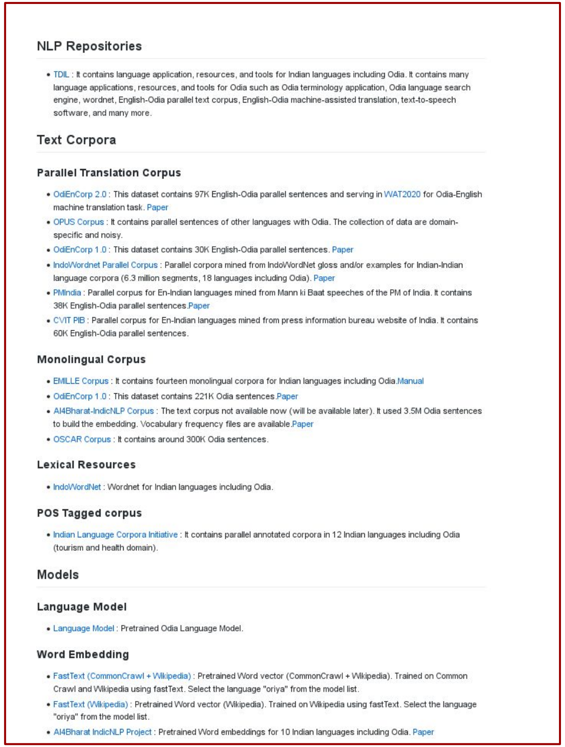


A Catalog for Odia Language NLP Resources

The purpose of this catalog is to provide a one-stop solution for the researchers looking for Odia NLP resources. This is a collective effort and any contribution to enriching Odia NLP resource are welcome. All contributors are listed on the CONTRIBUTOR list.

Table of Contents

- NLP Repositories
- Text Corpora
  - Parallel Translation Corpus
  - Monolingual Corpus
  - Lexical Resources
  - POS Tagged corpus
- Models
  - Language Model
  - Word Embedding
  - Morphanalyzers
- Text Classification
- Libraries / Tools
- Speech Corpora
- Other Indian language NLP Resources



NLP Repositories

- TDIL : It contains language application, resources, and tools for Indian languages including Odia. It contains many language applications, resources, and tools for Odia such as Odia terminology application, Odia language search engine, wordnet, English-Odia parallel text corpus, English-Odia machine-assisted translation, text-to-speech software, and many more.

Text Corpora

Parallel Translation Corpus

- OdiEnCorp 2.0 : This dataset contains 97K English-Odia parallel sentences and serving in WAT2020 for Odia-English machine translation task. Paper
- OPUS Corpus : It contains parallel sentences of other languages with Odia. The collection of data are domain-specific and noisy.
- OdiEnCorp 1.0 : This dataset contains 30K English-Odia parallel sentences. Paper
- IndoWordnet Parallel Corpus : Parallel corpora mined from IndoWordNet gloss and/or examples for Indian-Indian language corpora (6.3 million segments, 18 languages including Odia). Paper
- PMIndia : Parallel corpus for En-Indian languages mined from Mann ki Baat speeches of the PM of India. It contains 38K English-Odia parallel sentences.Paper
- CVIT PIB : Parallel corpus for En-Indian languages mined from press information bureau website of India. It contains 60K English-Odia parallel sentences.

Monolingual Corpus

- EMILLE Corpus : It contains fourteen monolingual corpora for Indian languages including Odia.Manual
- OdiEnCorp 1.0 : This dataset contains 221K Odia sentences.Paper
- AI4Bharat-IndicNLP Corpus : The text corpus not available now (will be available later). It used 3.5M Odia sentences to build the embedding. Vocabulary frequency files are available.Paper
- OSCAR Corpus : It contains around 300K Odia sentences.

Lexical Resources

- IndoWordNet : Wordnet for Indian languages including Odia.

POS Tagged corpus

- Indian Language Corpora Initiative : It contains parallel annotated corpora in 12 Indian languages including Odia (tourism and health domain).

Models

Language Model

- Language Model : Pretrained Odia Language Model.

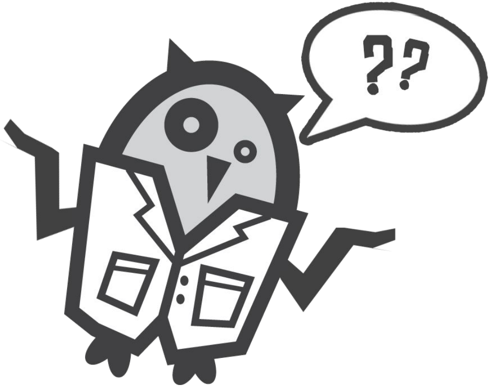
Word Embedding

- FastText (CommonCrawl + Wikipedia) : Pretrained Word vector (CommonCrawl + Wikipedia). Trained on Common Crawl and Wikipedia using fastText. Select the language "oriya" from the model list.
- FastText (Wikipedia) : Pretrained Word vector (Wikipedia). Trained on Wikipedia using fastText. Select the language "oriya" from the model list.
- AI4Bharat IndicNLP Project : Pretrained Word embeddings for 10 Indian languages including Odia. Paper

Conclusions and future work

- Extending OdiEnCorp 2.0 with more parallel data, again by finding various new sources.
- Build the Odia-English machine translation system using the (WAT2020 model) and release for research and non-commercial purposes.
- Building NLP resources for Odia language for research and development.

Q&A



Contact information:

- Email: shantipriya.parida@idiap.ch
- Twitter: @Shantipriyapar3
- Web : https://www.idiap.ch/~sparida/

References

[1] Parida, S., Dash, S. R., Bojar, O., Motlıcek, P., Pattnaik, P., & Mallick, D. K. OdiEnCorp 2.0: Odia-English Parallel Corpus for Machine Translation. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020* (p. 14).

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

Dr. Shantipriya Parida
Idiap Research Institute, Switzerland

Assoc. Prof. Ondřej Bojar
Charles University, Czech Republic

Assoc. Prof. Satya Ranjan Dash
KIIT University, India

Dr. Petr Motlicek
Idiap Research Institute, Switzerland

Thanks to all of our collaborators

Priyanka P. Pattnaik
COE AI Lab, India

Debasish Kumar Mallick
KIIT University, India

Biranchi Narayan Nayak
Vettons, Malaysia

Satya Prakash Biswal
University Of Chicago, USA

Amulya Ratna Dash
IQVIA RDS, India

Thank You