A Tulu Resource for Machine Translation

NOEMI AEPLI, MANU NARAYANAN

LREC-COLING 2024

Objective

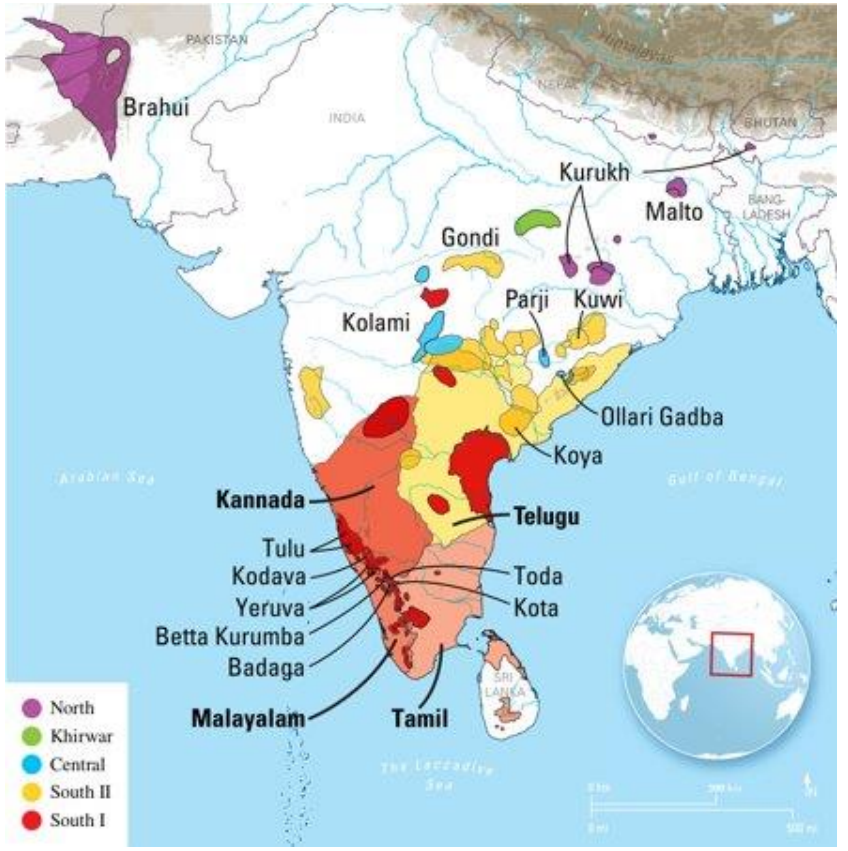
- Create a parallel dataset for Tulu (*TCY*) based on FLORES-200
- Apply low-resource translation methods based on transfer learning to translate English (*EN*) to Tulu (and vice versa)
  - Without using parallel training data in Tulu
  - Using only monolingual data in Tulu
- Use the newly created Tulu parallel dataset for evaluating the translation model

Motivation

- Majority of the world's 7000 languages are resource-scarce: absence of data or absence of domain-diverse data
- Low resource translation methods without parallel data are showing promising results
- Tulu, spoken by ~2.5 million in India:
  - Has very little digital resources (parallel data)
  - Is culturally relevant, though not a recognized official language

Dravidian Languages

- One of the four main language families in India
- Approx. 4500 years old
- Spoken by ~250 million people in southern and central India and surrounding countries
- Consists of ~80 languages

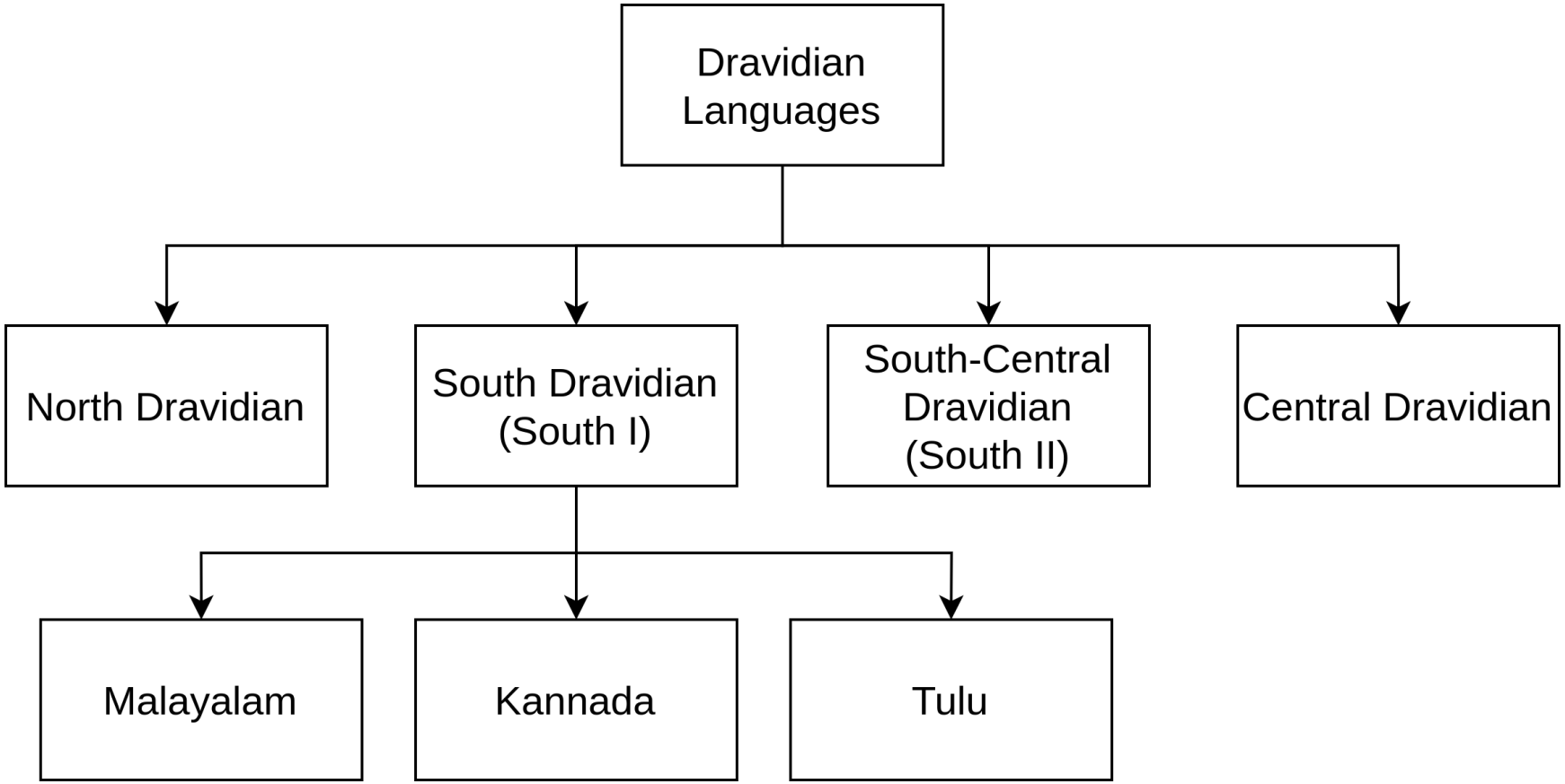


Map of Dravidian Languages

Dravidian Languages

- **Vowels**: 10-vowel system; 5 short and 5 long ones
- **Consonants:** Retroflex consonants are present
- **Cases:** Between 5 and 8 cases
- **Morphology:** Agglutinative; suffixation and compounding used to express voice and tense
- **Syntax:** Word order is flexible subject-object-verb; verb always final
- **Writing:** Major scripts used are Kannada (*KN*), Malayalam (*ML*), Tamil and Telugu

Dravidian Languages



Tulu

- ~2.5 million speakers, mainly in:
  - Dakshina Kannada and Udupi districts of Karnataka
  - Kasargod district of Kerala
- Most speakers are bilingual, speaking Malayalam or Kannada as well
- Many dialects; no standardization yet
- Traditionally used Tigalari script (no Unicode script available); rapidly replaced by Kannada script due to disuse

Tulu

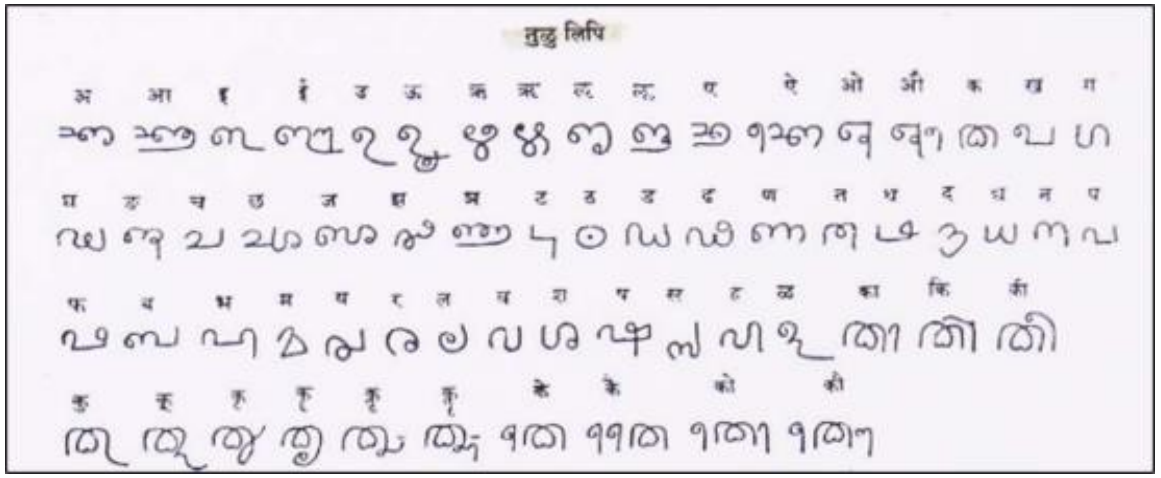
Kannada	Malayalam
ಆಯೆ ಬರ್ಪೆ	ആയെ ബര്പെ

(aa-ye ba-rpe) 'e' in 'french'
He will come

Kannada	Malayalam
ಯಾನ್ ಬರ್ಪೆ'	യാന് ബര്പെ'

(yaa-n ba-rpé) 'e' in 'end'
I will come

The two '/e/' sounds in *Tcy*



Tigalari alphabet

Kannada

- ~43.7 million speakers, mainly in Karnataka
- One of the 22 official ("Scheduled") languages of India
- Standardized script; Unicode characters available
- Increasingly, more loan words from Kannada being adopted into Tulu
- Higher influence of Sanskrit (an Indo-Aryan language)

ಎಲ್ಲಾ ಮಾನವರೂ ಸ್ವತಂತ್ರರಾಗಿಯೇ ಜನಿಸಿದ್ದಾರೆ. ಹಾಗೂ ಘನತೆ ಮತ್ತು ಹಕ್ಕು ಗಳಲ್ಲಿ ಸಮಾನರಾಗಿದ್ದರೆ. ವಿವೇಕ ಮತ್ತು ಅಂತಃಕರಣ ಗಳನ್ನು ಪಡೆದವರಾದ್ದರಿಂದ ಅವರು ಪರಸ್ಪರ ಸಹೋದರ ಭಾವದಿಂದ ವರ್ತಿಸಬೇಕು.

Kn script sample

Dataset for Tulu

- FLORES-200: 2009 sentences (dev: 997, test: 1012) in 200 languages
- Human translated to Tulu: 1300 sentences (dev:647, test:653)
- Team of native Tulu speakers based in Mangaluru

Translation Guidelines

Adapted from NLLB and FLORES-200 developers:

- Neutral, informative and clear translation
- No assistance from machine translation tools
- Proper nouns, abbreviations, idioms and metaphors as per common use in Tulu
- Consensus from experts in the team in the case of queries
- Review by experts in the team for typos and errors

Translation Challenges

- Loss of vocabulary over time; Kannada loanwords
- Passive voice not commonly used
- Dialectical variations; Mangaluru (Central Tulunad) dialect was followed for consistency
- Ambiguity in translating 'you': Tulu has singular (*ee)*, plural (*nikulu*) and formal (*eeru*) forms for 'you'
- Phonetic variations of */e/* sound cannot be represented in Kannada script

Machine Translation

- Transfer learning approach based on NMT-Adapt paper
- Leverages lexical and syntactic similarities between a high-resource and a low-resource language
- Without using any parallel data
- Utilizes monolingual data in low-resource language
- Combines back-translation and denoising autoencoding
- Multi-step and iterative

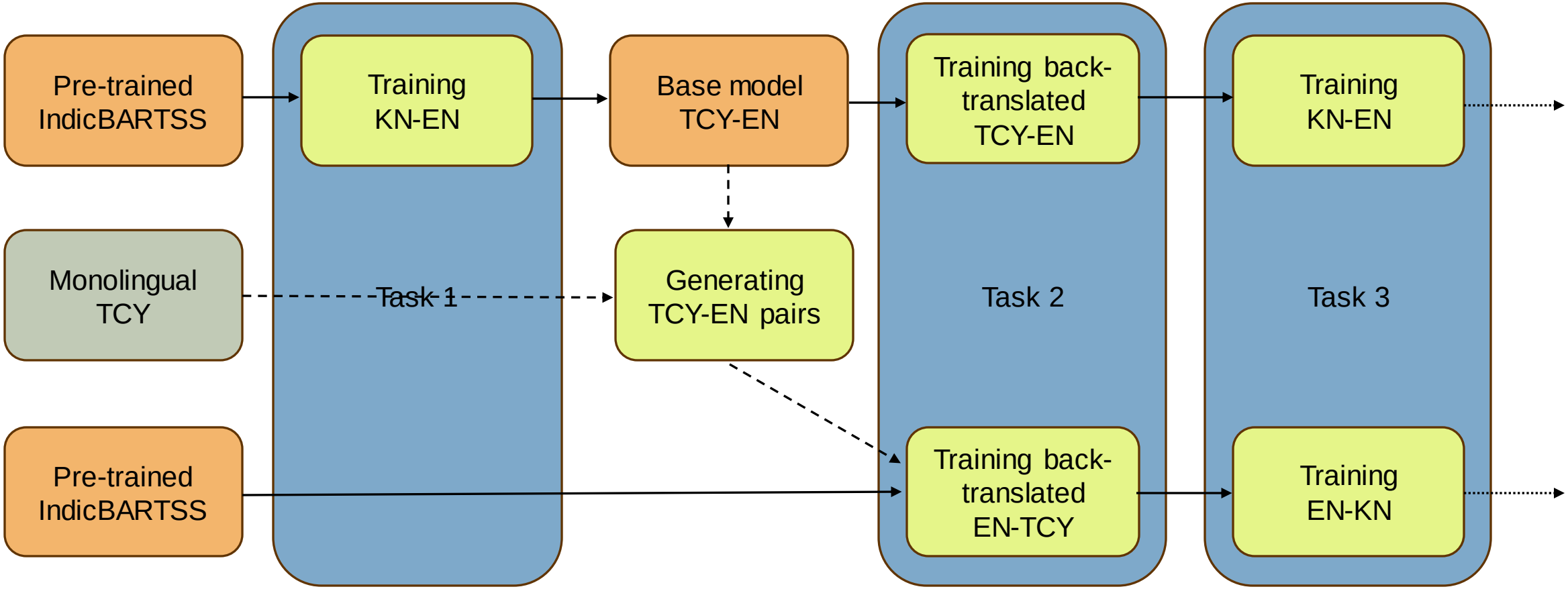
Experimental Setup

- Limitations to implementing NMT-Adapt steps:
  - mBART and mBART-50 do not support Kannada
  - Adversarial training step was omitted
- Pre-trained model: IndicBARTSS
- Training toolkit: YANMTT
- Tokenizer: ALBERT
- Evaluation: BLEU score using sacreBLEU

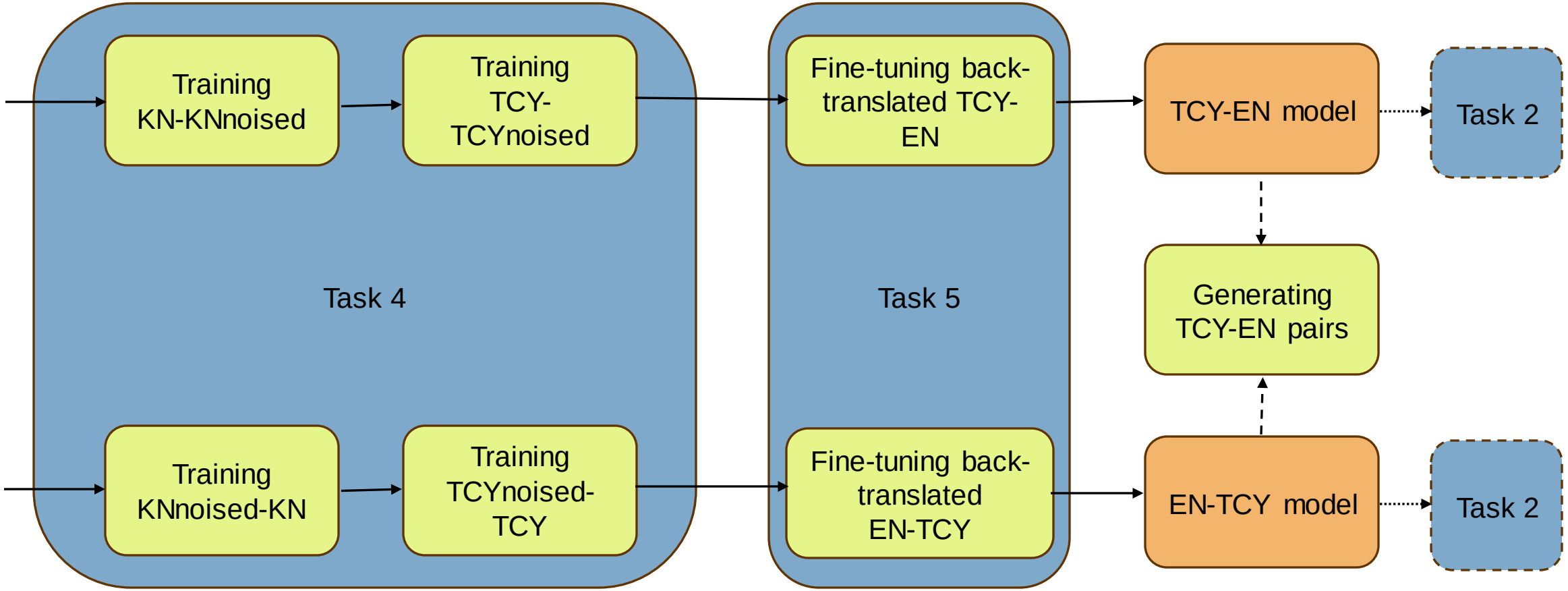
Datasets

Type	Source	Size (sentences)	Details
Training data: *EN-KN*	Samanantar	4,093,524	From Indian websites, govt documents ...
Monolingual data: *TCY*	Scraped from TCY WIkipedia	40,124	1,894 articles
Test data: *EN-TCY*	Human translation	1,300	Dev set: 647 Test set: 653
Additional EN-TCY training data	DravidianLangTech-22	8,300	Manually translated from digital *TCY* data

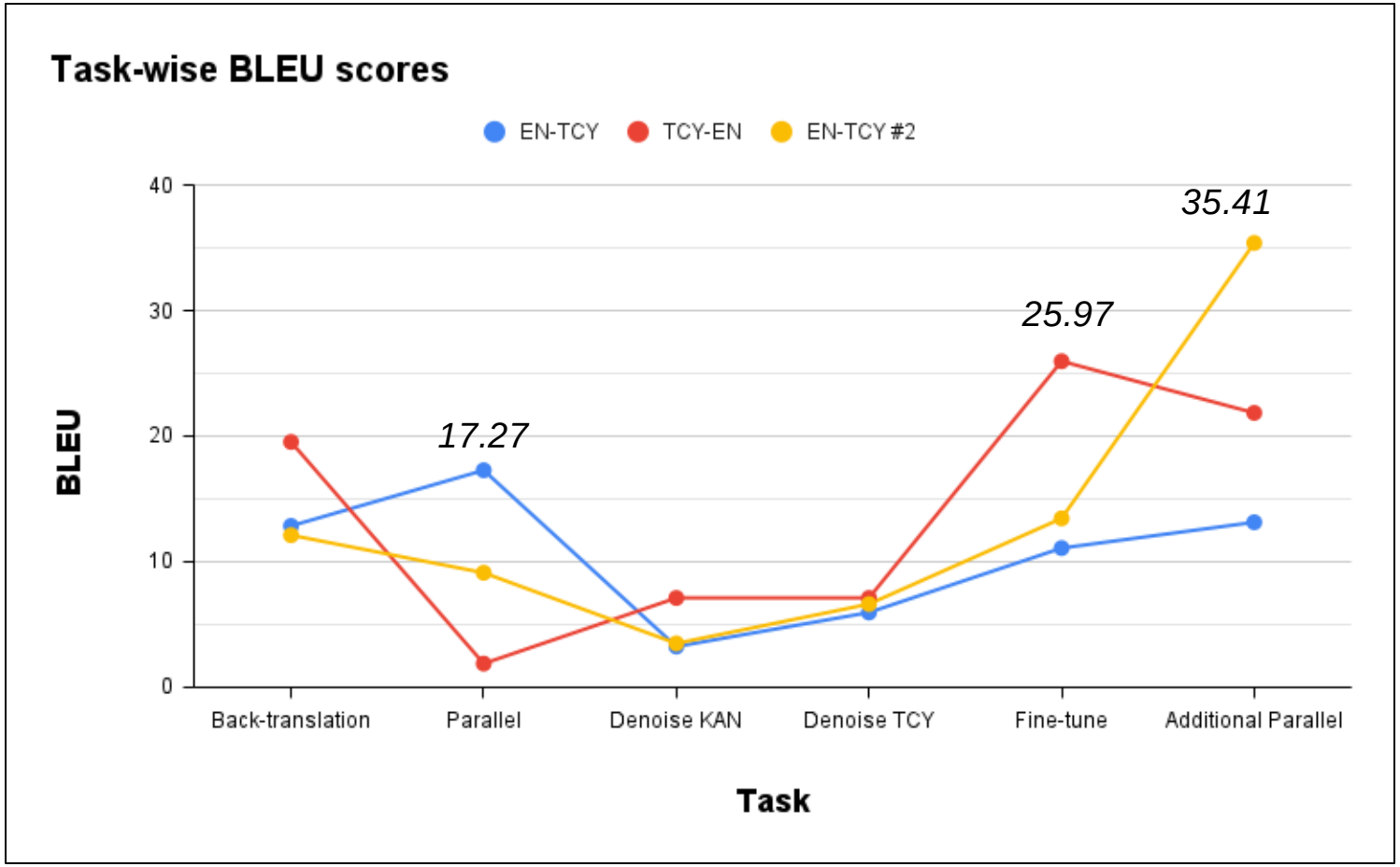
Experiment


Pre-trained IndicBARTSS
Training KN-EN
Base model TCY-EN
Training back-translated TCY-EN
Training KN-EN
Monolingual TCY
Task 1
Generating TCY-EN pairs
Task 2
Task 3
Pre-trained IndicBARTSS
Training back-translated EN-TCY
Training EN-KN

Experiment


Training KN-KNnoised
Training TCY-TCYnoised
Training KNnoised-KN
Training TCYnoised-TCY
Task 4
Fine-tuning back-translated TCY-EN
Fine-tuning back-translated EN-TCY
Task 5
TCY-EN model
EN-TCY model
Generating TCY-EN pairs
Task 2
Task 2

Results


Task-wise BLEU scores
EN-TCY
TCY-EN
EN-TCY #2
BLEU
40
30
20
10
0
17.27
25.97
35.41
Back-translation
Parallel
Denoise KAN
Denoise TCY
Fine-tune
Additional Parallel
Task

Results

Insights from qualitative analysis of translations:

- Transliteration of TCY words to EN prevalent in the model with highest BLEU
- Some common TCY words still retain the KN meaning:
  - 'uppuna' (together) translated as salt
  - 'tenkāyi amērikā' (South America) is translated as United States, whereas 'dakṣiṇa āphrikā' (South Africa) is correctly translated as South Africa
- Presence of KN characters in EN translation (and vice versa) after denoising autoencoding
- Instances of words/sequences repeating
- Translation quality diminishes for longer sentences

Conclusions

- First parallel dataset for English-Tulu, based on FLORES-200
- Transfer learning approach utilizing similarities between KN and TCY achieved a BLEU score opf 25.97 for TCY-EN
- However, overall low score limits useability
- Without adversarial training, denoising autoencoding has limited effectiveness:
  - Encoder does not capture language-agnostic semantic information

Future Research

- Extend the training toolkit (YANMTT), or use a different experimental setup to implement adversarial training
- Acquire more monolingual TCY data:
  - 'CURL' project by Uni Leipzig
  - Scrape 'Bible' data from Joshua Project
- Complete translating the entire FLORES-200 dataset

Thank you

References

W.-J. Ko, A. El-Kishky, A. Renduchintala, V. Chaudhary, N. Goyal, F. Guzmán, P. Fung, P. Koehn, and M. Diab. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 802–812, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.66. URL https://aclanthology.org/2021.acl-long.66.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

A. F. Aji, N. Bogoychev, K. Heafield, and R. Sennrich. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.688. URL https://aclanthology.org/2020.acl-main.688.

References

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Sy2ogebAW.

S. Bala Das, A. Biradar, T. Kumar Mishra, and B. Kr. Patra. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6), jun 2023. ISSN 2375-4699. doi: 10.1145/3587932. URL https://doi.org/10.1145/3587932.

A. Currey and K. Heafield. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5610. URL https://aclanthology.org/D19-5610.

R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. Khapra, and P. Kumar. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.145. URL https://aclanthology.org/2022.findings-acl.145.

References

N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl a 00474. URL https://aclanthology.org/2022.tacl-1.30.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl a 00065. URL https://aclanthology.org/Q17-1024.

D. Kakwani, A. Kunchukuttan, S. Golla, G. N.C., A. Bhattacharyya, M. M. Khapra, and P. Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445.

B. Krishnamurti. *The Dravidian Languages*. Cambridge Language Surveys. Cambridge University Press, 2003. doi: 10.1017/CBO9780511486876.