University of Zurich[UZH]

Swiss National Science Foundation

A Tulu Resource for Machine Translation

Manu Narayanan & **Noëmi Aepli**, Department of Computational Linguistics, University of Zurich

ಯನನ್ ಬೆಂಬಲ ಮನ್ಪುನ ಜನಕುಲ್ ಯನ್ನೊಟ್ಟು ಉಲ್ಲೆರ್ ಪಂದ್ ಯೆಂಕ್ ಕೃಷಿವುಂಡ್.

> **TCY:** *I am happy that there are people willing to support me*

Summary

> we introduce the **first MT dataset for Tulu**: human translations of the benchmark dataset *FLORES-200* (2k sentences)

> **Tulu** (**TCY**) is a **south Dravidian** language with ca. 2.5 million speakers in southwestern India, closely related to **Kannada** (**KAN**)

> we develop a **MT system for English-Tulu**: leveraging the resources of related Dravidian languages & employing **transfer learning**

> we collaborated with the volunteer organization *Jai Tulunad* that is dedicated to preserving and promoting Tulu language and culture

> our MT system achieved a **BLEU** score of **26** for TCY-EN, outperforming Google Translate by 7 BLEU points (September 2023)

The First Dataset for Tulu MT

> **data:** extension of the existing multilingual ***FLORES-200***

> **translators: 15 native Tulu speakers** translating from English and Kannada to Tulu

> **community contact**: volunteer organization ***Jai Tulunad***

> **dialect:** Mangaluru (Central Tulunad) dialect

> **guidelines:** adapted from the original FLORES-200 guidelines

> **challenges:** vocabulary, passive voice, dialectal variations, script (2 version of /e/), ambiguity of "*you*"

Datasets

Dataset	Source	#sents
EN–KN training	Samanantar	4,093,524
EN–KN test	FLORES-200	2,009
TCY monolingual	Wikipedia	40,124
EN–TCY test	Human transl. FLORES	1,300
EN–TCY training	DravidianLangTech-22	8,300

Kannada ಆಯೆ ಬರ್ಪೆ — *Malayalam* ആയെ ബർപെ
(aa-ye ba-rpe) 'e' in 'french'
He will come

Kannada ಯಾನ್ ಬರ್ಪೆ' — *Malayalam* യാന് ബർപെ'
(yaa-n ba-rpé) 'e' in 'end'
I will come

> examples of the 2 /e/ sounds in Tulu

Method: NMT-Adapt

> **Task 1: Fine-tuning for back-translation**

> fine-tuning IndicBARTSS LM with EN-KN Samanatar dataset

> **Task 2: Training with back-translation**

> fine-tuning IndicBARTSS LM with back-translation pairs from Task 1

> **Task 3: Training with parallel data**

> train model from Task 2 with parallel EN-KN Samanatar dataset

> **Task 4: Denoising autoencoding**

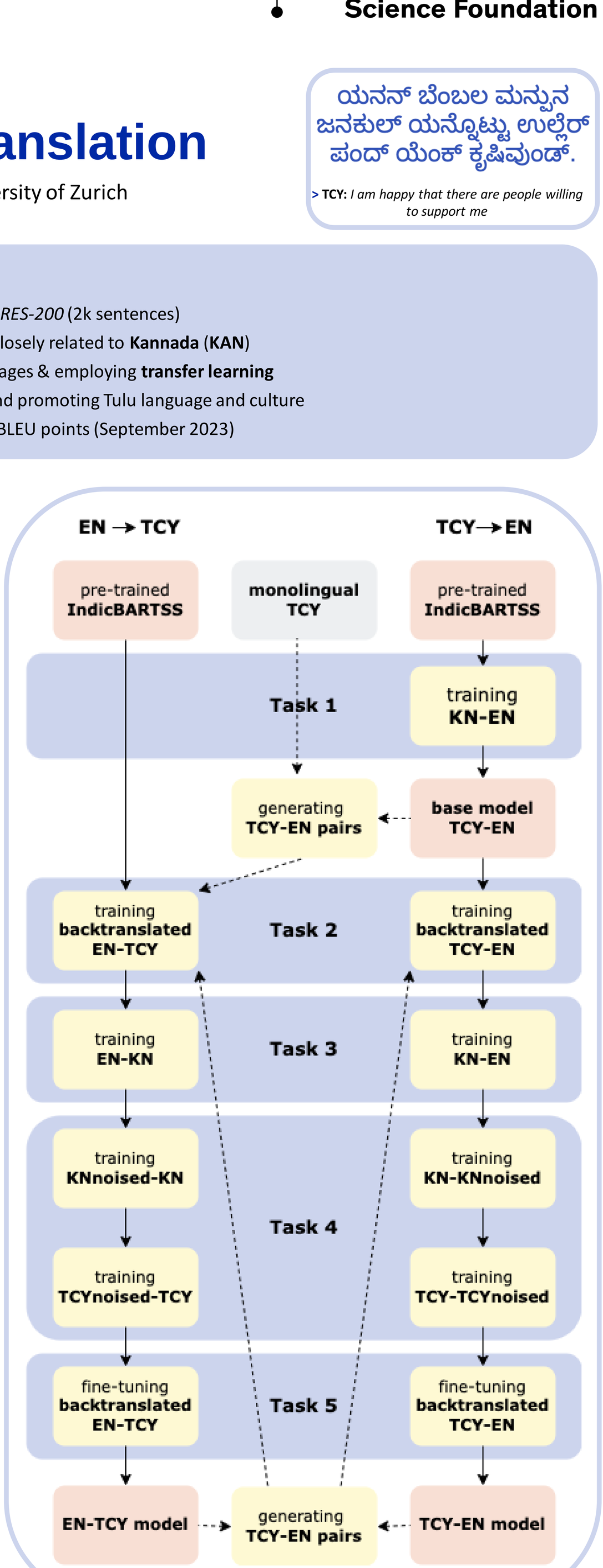
> training EN-TCY base model with noised TCY & KN sentences as source and non-noised as target

> noising: random shuffling and word masking

> **Task 5: Fine-tuning with back-translation**

> fine-tuning EN-TCY model with back-translated pairs

> these sentence pairs are used to repeat Tasks 2-5 on the TCY-EN base model from Task 1



Results

Iteration	Direction	Task no.	Task	Lannguages	BLEU
1	TCY–EN	1	fine-tuning with	KN–EN	1.84
	EN–TCY	2	back-translation with	EN–TCY	12.83
	EN–TCY	3	training with parallel	EN–KN	17.27
	EN–TCY	4a	denoising autoencoding with	KN	3.20
	EN–TCY	4b	denoising autoencoding with	TCY	5.92
	EN–TCY	5	fine-tuning with back-translation data		11.06
	TCY–EN	2	back-translation with	TCY–EN	19.53
	TCY–EN	4a	denoising autoencoding with	KN	7.08
	TCY–EN	4b	denoising autoencoding with	TCY	7.08
	TCY-EN	5	fine-tuning with back-translation data		**25.97**
2	EN–TCY	2	back-translation with	EN–TCY	12.09
	EN–TCY	3	training with	EN–KN	9.09
	EN–TCY	4a	denoising autoencoding with	KN	3.45
	EN–TCY	4b	denoising autoencoding with	TCY	6.59
	EN–TCY	5	fine-tuning with back-translation data		13.43

Conclusion

> we created the first dataset for Tulu

> we trained a MT system without parallel TCY-EN data

> using Kannada data & the NMT-Adapt method

> our MT system achieved a BLEU score of 26 for TCY-EN

> we collaborated with the local organization *Jai Tulunad*

> dataset and code are available for research