SrcMix: Mixing of Related Source Languages Benefits Extremely Low-resource Machine Translation

Sanjeev Kumar, Preethi Jyothi, Pushpak Bhattacharyya

{sanjeev, pjyothi}@cse.iitb.ac.in

EACL 2026 RABAT · MOROCCO
Mars · March 24-29, 2026 · مارس

Introduction

Problem Definition

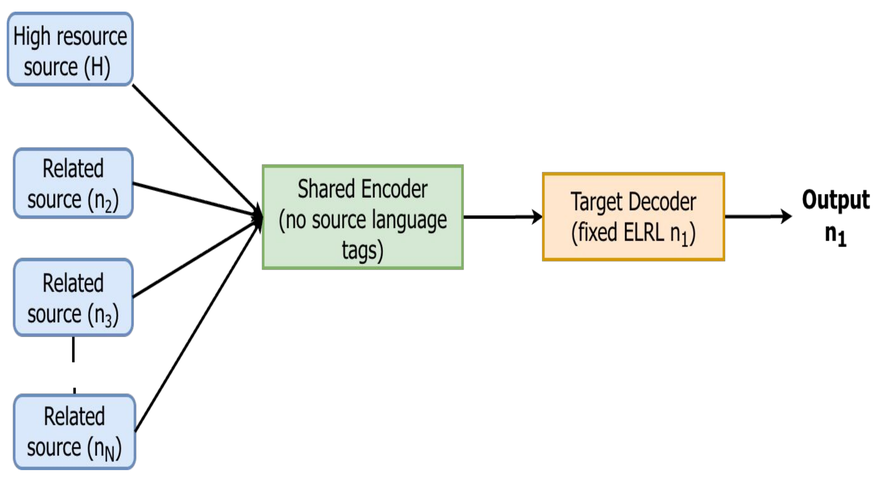
- Extremely low-resource languages (ELRLs) have <10K parallel sentences
- Zero-shot transfer performs poorly on machine translation (MT)
- Naive multilingual training for MT causes negative transfer
- Millions of speakers remain digitally excluded

Can structured multilinguality help?

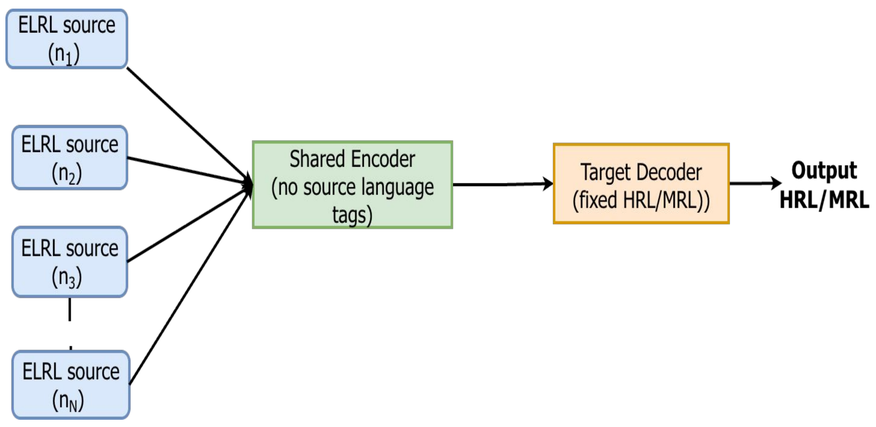
Source-Mix (SrcMix)

Key Idea: Mix multiple linguistically related source languages while keeping the target language fixed

Only linguistically related source languages are mixed.

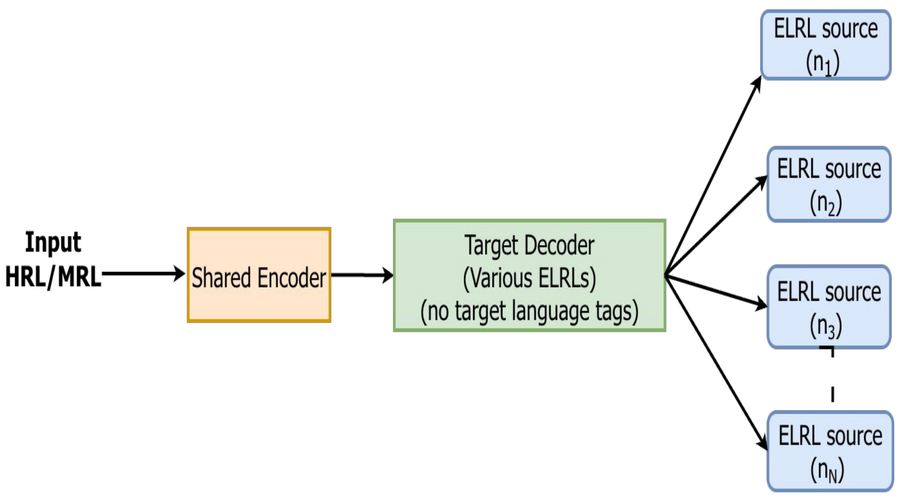


- **Reverse Direction** (ELRL to HRL/MRL):



Target-Mix (TgtMix)

- **Forward Direction** (HRL/MRL to ELRL):



- **Reverse Direction** (ELRL to HRL): **Not applicable**
  - Reason: Only one target HRL (English) and one MRL (Hindi).

Introducing New Resources for Angika MT

- Angika (anp): An Indo-Aryan ELRL spoken in India and Nepal.
- Written in the Devanagari script and follows SOV word order.
- Significant speaker base (15M), but limited MT resources.
- We release **first** public dataset for Angika MT.

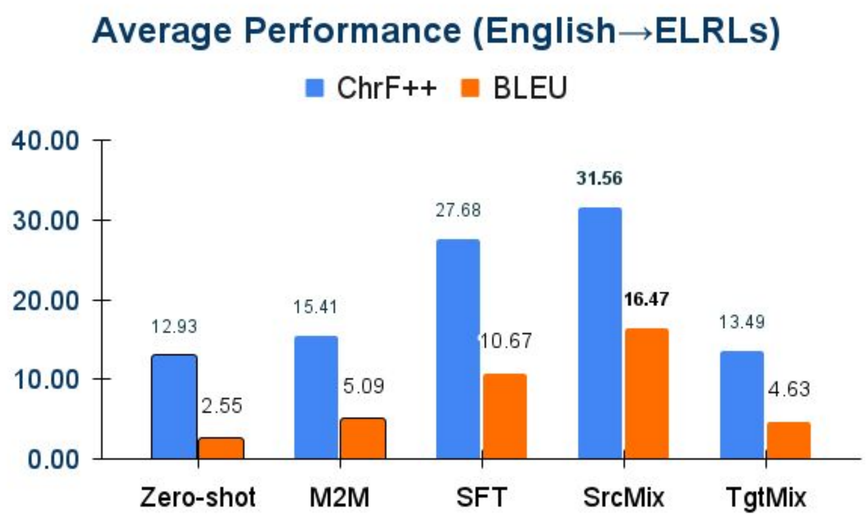
Experimental Setup

- **Datasets**:
  - NLLB Seed Corpus (6,192 sentences each)
  - FLORES-200 (dev: 997, dev-test: 1,012 samples)
  - Custom-created Angika dataset
- **Language Groups** (14 ELRLs):
  - 4 African: Nigerian Fulfulde, Nuer, Bambara, Tamasheq
  - 4 Romance: Friulian, Ligurian, Limburgish, Sardinian
  - 3 Indic: Angika, Bhojpuri, Magahi
  - 3 Arabic: Dari, Kashmiri_Arab, Southern Pashto
- **Models:**
  - Aya-101 (13B): LLM based on mT5 architecture
  - mT5-large (1.2B): Traditional NMT model
  - **Decoder-only** models such as LLaMa-3.1-8B, and Gemma-7B perform **poorly** in ELRLs.

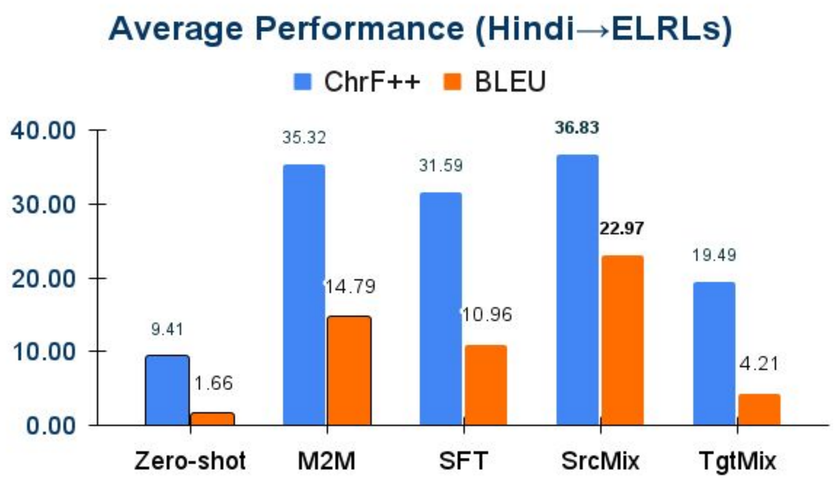
Results

- Key Result:
  - **Zero-shot:** very low performance
  - **M2M (Many-to-Many)**: Degrades ELRL translation
  - **SFT (Supervised finetuning)**: Strong improvement
  - **SrcMix**: **Best performance (BLEU ↑ & ChrF++ ↑)**

SrcMix improves performance by **+6 BLEU and +4 ChrF++** over **SFT**.



SrcMix improves performance by **+12 BLEU and +5 ChrF++** over **SFT**.



Why Does SrcMix Work?

- Related languages share morphology and syntax.
- Shared scripts improve lexical alignment.
- Decoder learns from multiple related sources while keeping target language fixed.

Multilinguality helps only when related languages are mixed in a structured way.

Discussion & Conclusions

- **Directionality matters:** Source-side mixing (SrcMix) consistently outperforms Target-side mixing (TgtMix)
- Linguistic relatedness enables positive transfer
- Mixing related source languages yields consistent gains
- Naive multilingual training introduces negative transfer in ELRL MT
- Structured multilingual transfer outperforms naive multilinguality

Future Work

- Investigate adaptive mixing ratios instead of uniform mixing
- Explore token-level collaboration during decoding
- Apply SrcMix to speech and multimodal translation

Acknowledgements

The first author gratefully acknowledges support in the form of a Ph.D. scholarship from the TCS Research Foundation. The second author gratefully acknowledges financial support from the Amazon–IITB AI/ML Initiative.

References

- Üstün et al., 2024. Aya-101: A massively multilingual LLM for translation.
- Xue et al., 2021. mT5: A massively multilingual text-to-text transformer.
- Maillard et al., 2023. NLLB Seed Corpus.
- Goyal et al., 2022. FLORES-200 Benchmark.

Code

Dataset

Let's connect:

@snjev310