Empirical Experiments on Low-resource Translation using mBART Pre-Training: Performance Analysis of Translating Kanada/Sinhala to English using Intermediate Task Fine-Tuning

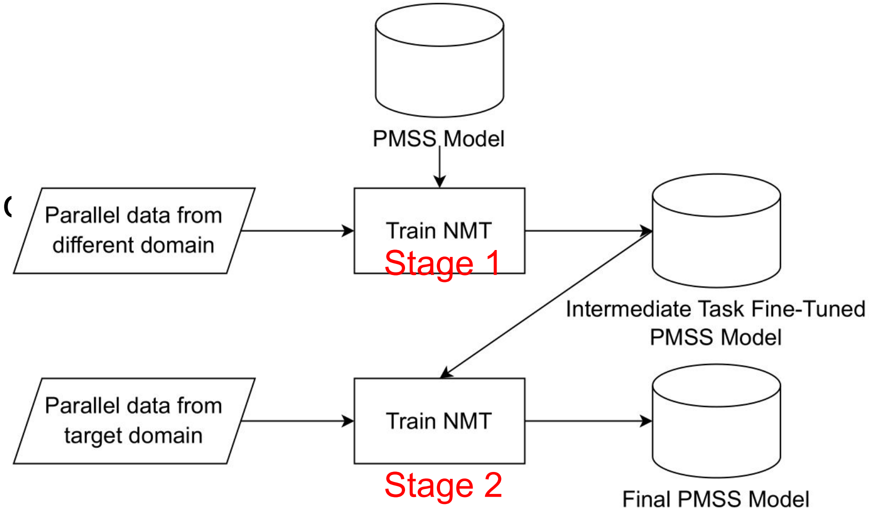
CSC495
Yuchen(Rachel) Zeng

Table of Content

1. Background
2. Purpose
3. Results & Conclusion
    a. Single-domain
    b. Mixed-domain
    c. Mixed-all-domain
    d. Mixed-domain FT vs. Mixed-domain ITFT
    e. Mixed-domain FT vs. Single-domain ITFT
    f. Mixed-domain ITFT vs. Single-domain ITFT
4. Conclusion
5. Next Step

Background

- Using auxiliary parallel data to build domain-specific Neural Machine Translation (NMT) systems for low-resource languages (LRLs) is an under-explored problem.
- We previously conducted a large-scale study o two ways to utilize this parallel data
  - Intermediate Task Fine-Tuning (ITFT)
  - Additional pre-training
- Was focusing on EN->XX.
- **The presentation will expand upon previous research and provide additional evaluations on the reverse direction(XX->EN) using ITFT.**
  - Sinhala
  - Kanada (Not seen in pre-training)


PMSS Model
Parallel data from different domain
Train NMT
Stage 1
Intermediate Task Fine-Tuned PMSS Model
Parallel data from target domain
Train NMT
Stage 2
Final PMSS Model

Purpose

- Verify the robustness and consistency of our previous findings
  - Before, our experiments only cover limited domains, datasets, and directions
  - Provide additional evaluations in the reverse direction(XX->EN)
- Compare various ITFT techniques to find the optimal approach for different scenarios
- Investigate the impact of dataset divergence on model performance
  - Obtain a more straightforward observation in mixed-all-domain ITFT experiments.

Single-domain: ITFT Outperforms Baseline

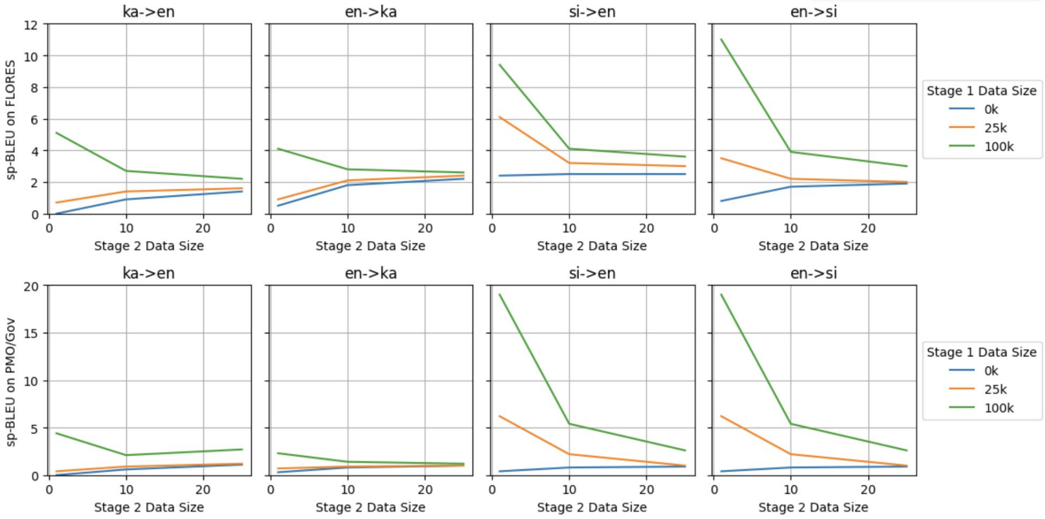
Domains used:
Stage 1: CCAlign(Intermediate Task)
Stage 2: Bible
Test: PMO/Gov, Flore

Baseline: Only trained by Bible(Stage 2)->Blue Line

Conclusion:

- Results using ITFT **outperforms baseline**
- As stage 2 domain size increase, gain from ITFT diminished and results tend to converge to the baseline.
- More beneficial when task has less than 10k data points
- For Kanada, better performance for KA->EN comparing to EN->KA
- Consistent with the result from EN->XX(Previous work)

New Old New Old



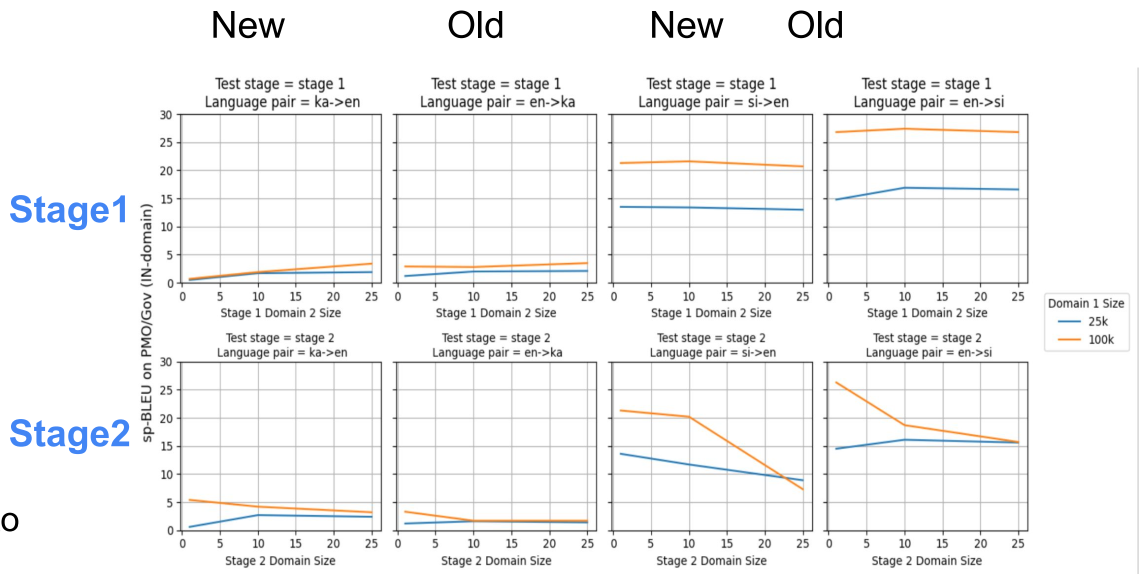
Mixed-domain: Performance Subject to Data Size and Domain Divergence

Domains used:
Stage 1: CCAlign + Bible(Intermediate Task)
Stage 2: Bible
Using same data size for Bible in Stage 1 & 2
Test: PMO/Gov

Baseline:
Only train by CCAlign + Bible(Only Stage 1)

Findings:

- As stage 2 domain size increase, gain from ITFT diminished and results tend to converge to the baseline.
  - Can even underperform(SI->EN)



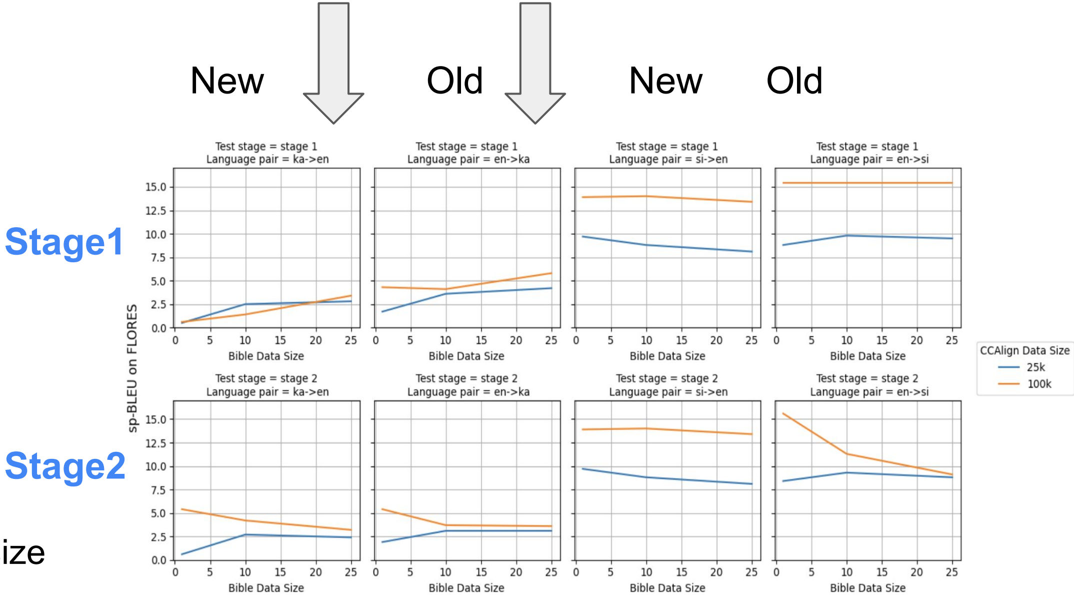
Mixed-domain: Performance Subject to Data Size and Domain Divergence

Domains used:
Stage 1: CCAlign + Bible(Intermediate Task)
Stage 2: Bible
Using same data size for Bible in Stage 1 & 2
Test: Flore

Baseline:
Only train by CCAlign + Bible(Only Stage 1)

Finding:

- Similar pattern as tested using PMO/Gov
- Mixed domain FT for Kanada
  - Performance is better as Bible data size grow
  - Comparing to when use 25k CCAlign data, using 100k CCAlign data underperformed when stage 2 data size is 10k, overperform when stage 2 data size is 25k.



New Old New Old

Stage1

Stage2

Conclusion:

- Mixed-domain FT is subject to data size and domain divergence
- Have a better performance for Kanada since it is unseen in pre-training

Mixed-domain with All Domains: Influenced by Domain Divergence

Domains used:
Stage 1: CCAlign + PMO/Gov + Bible(Intermediate Task)
Stage 2:

- Bible
- PMO/Gov

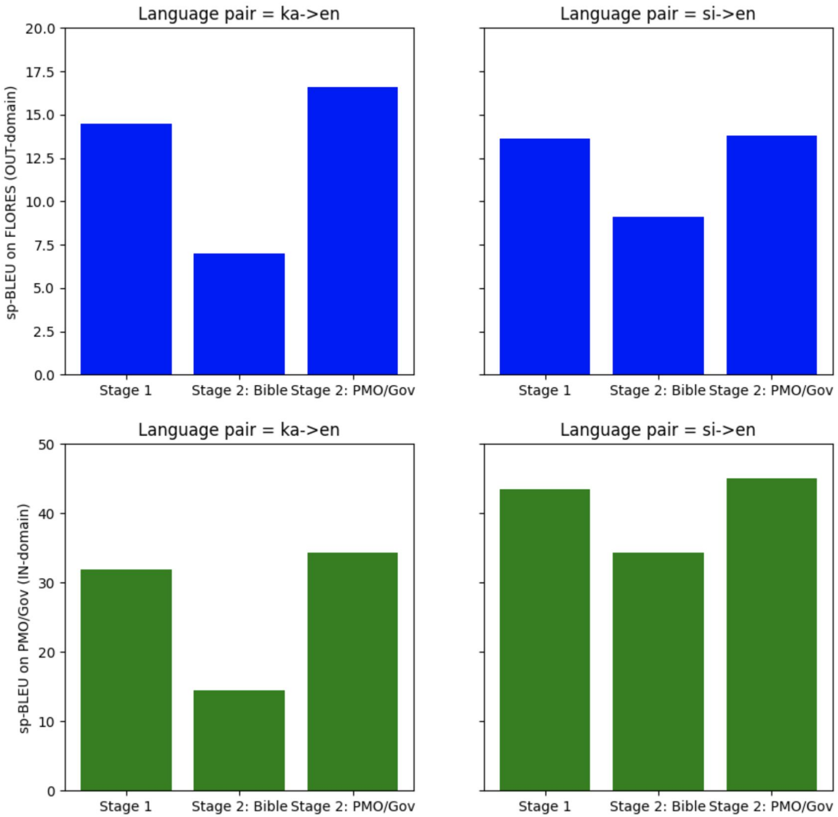
Data Size: All using 25k
Test: Flore

Finding:

- Underperform when using Bible in Stage 2, overperform when using PMO/Gov in Stage 2.

Conclusion:

- **Domain divergence has a major influence**

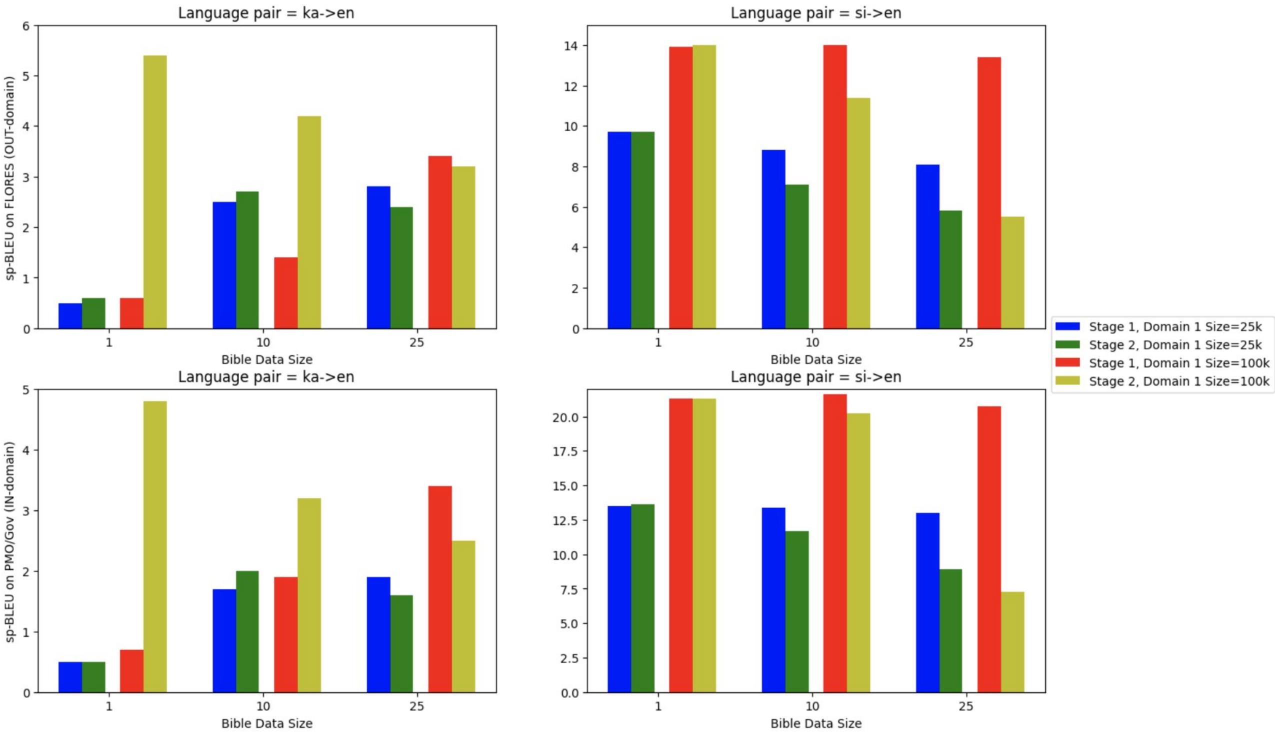


Mixed-domain FT v.s. Mixed-domain ITFT: Better Performance for Kanada

Domains used:
Stage 1:
CCAlign + Bible(Intermediate Task)
Stage 2: Bible
Using same data size for Bible in Stage 1 & 2
Test: Flore

Conclusion:

- Huge gain when Bible data size is small
- Domain Divergence
- Clearer pattern for Kanada since it is not seen in pre-training



Mixed-domain FT v.s. Single-domain ITFT: Mixed-domain outperforms for Sinhala

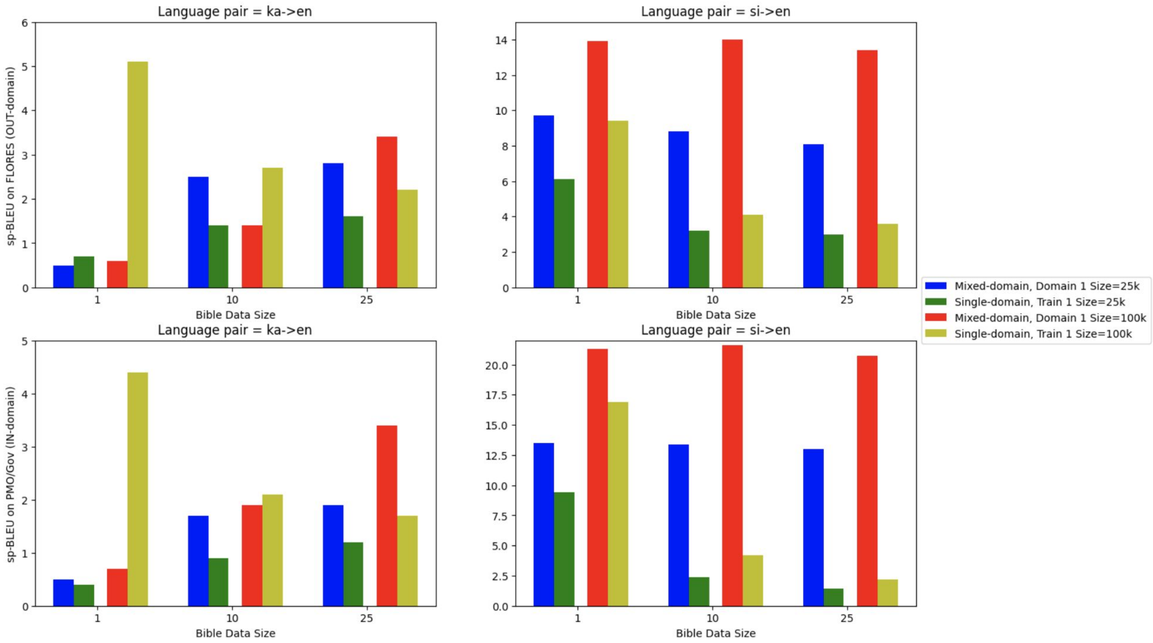
Mixed-domain:
Only Stage 1: CCAlign + Bible

Single domain:
Stage 1: CCAlign(Intermediate Task)
Stage 2: Bible

Test: Flore

Conclusion:

- Kanada: Single-domain ITFT sometimes perform better
- Sinhala: Mixed-domain FT is a clear winner



Mixed-domain ITFT v.s. Single-domain ITFT: Mixed-domain ITFT Outperforms

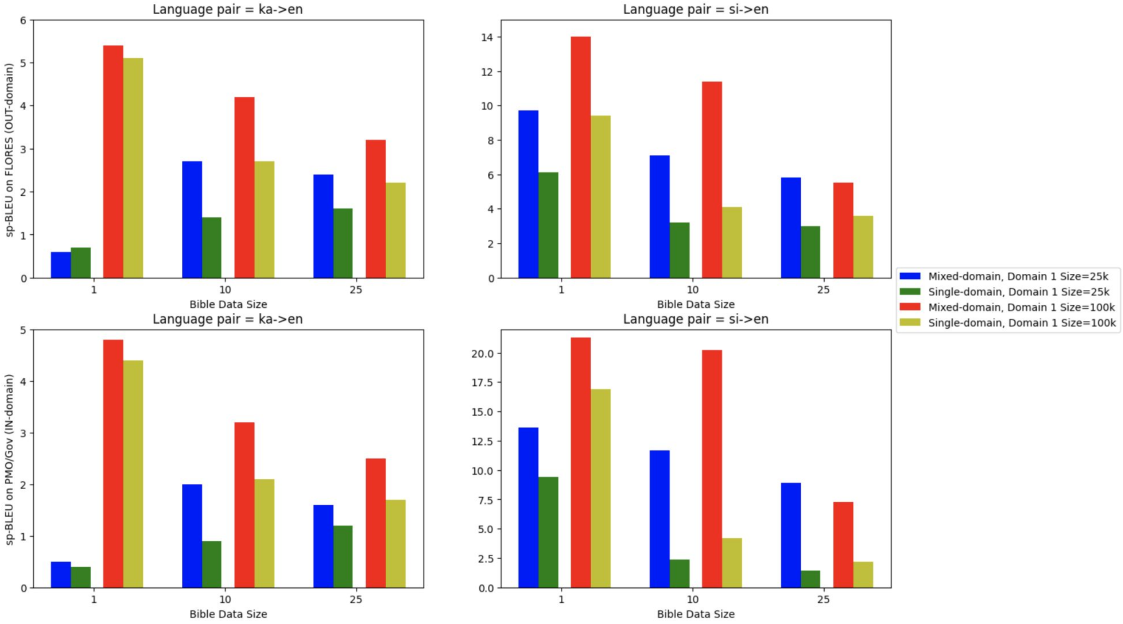
Domains used:
Stage 1:
CCAlign + Bible(Intermediate Task)
Stage 2: Bible
Using same data size for Bible in Stage 1 & 2

Single domain:
Stage 1: CCAlign(Intermediate Task)
Stage 2: Bible

Test: Flore

Conclusion:

- Mixed-domain ITFT is a clear winner
- Previously we found that single-domain ITFT sometimes perform better, but this pattern is unseen here
- The only exception is KA->EN, when domain size for bible is 1k and single-domain ITFT overperform for 0.1 spBELU



Conclusion

- ITFT has noticeable impact on performance of LRT for XX->EN
- Mixed-domain ITFT results in better performance than Single-domain ITFT
- Domain divergence affects performance
- Consistent with results from previous experiments for EN->XX, even with better performance

Next Step

1. Multilingual ITFT
    a. [May 5] Finish the settings
    b. [May 12] Complete the experiments
2. [May 19] Multilingual and multi-domain ITFT
3. [Ray] Explore new pre-training objectives on parallel data

Thank you