DeMuX: Data-efficient Multilingual Learning

Simran Khanuja[1], Srinivas Gowriraj[2], Lucio Dery[1], Graham Neubig[1]

[1]Carnegie Mellon University [2]TikTok

Correspondence: skhanuja@andrew.cmu.edu

1 Research Question

Given:

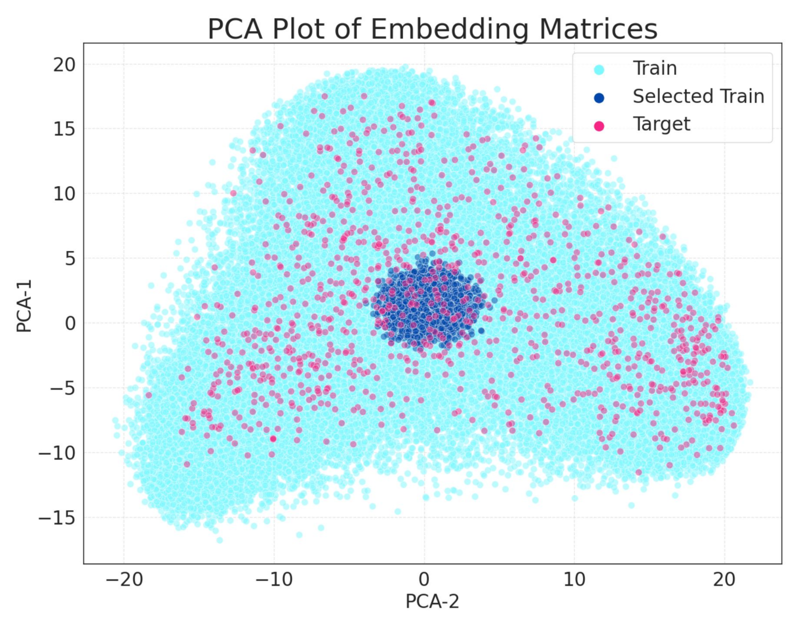
1. Pre-trained **multilingual model**
2. (*Large amounts of*) unlabelled multilingual **source data**
3. (*Small amounts of*) unlabelled multilingual **target data**
4. Annotation **budget**

How do we select the **exact data points** to give to annotators for best performance in their **domain and target** languages, under a **fixed budget**, from a **multilingual source** data pool?

2 Overview and benefits of proposed framework: DeMuX

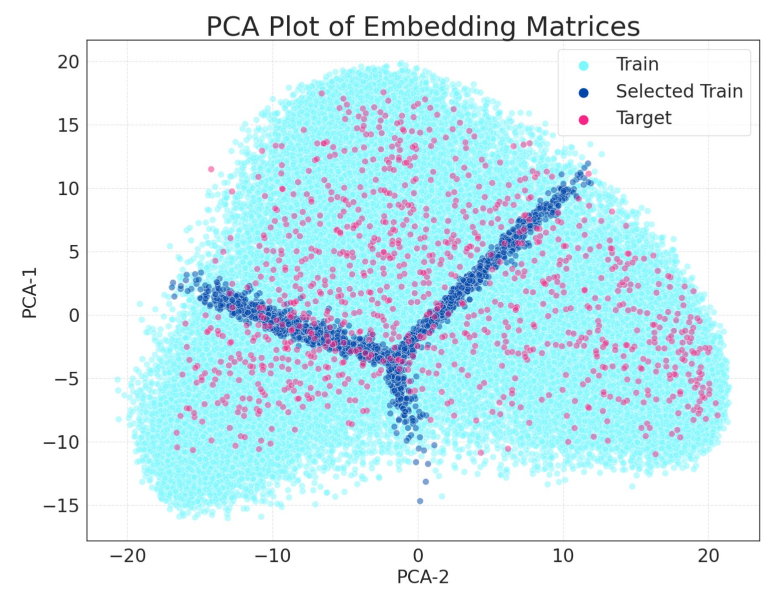
Benefits: DeMuX works with…

1. **No** language identification ✅
2. **No** linguistic feature information ✅
3. **No** past model performance ✅
4. **Disjoint** source/target languages ✅



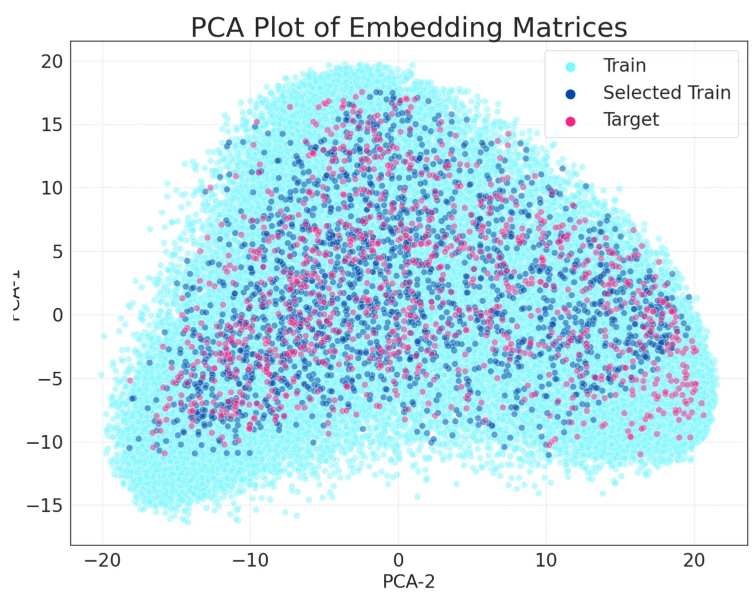
Average Distance

Picks points having a **minimum average distance** w/ the **unlabelled target pool**.



Uncertainty

Picks points that the model would **potentially misclassify**



KNN-Uncertainty

Picks **most uncertain** points from the union of **top-k neighbors**

3 Experimental Setup and Key Results

Dataset	Single Target			Multi-Target	
	High-perf.	Mid-perf.	Low-perf.	Geo Pool	Low-Performing Pool
UDPOS	French	Turkish	Urdu	Telugu, Marathi, Urdu	Arabic, Hebrew, Japanese, Korean, Chinese, Persian, Tamil, Vietnamese, Urdu
NER	French	Turkish	Urdu	Indonesian, Malay, Vietnamese	Arabic, Indonesian, Malay, Hebrew, Japanese, Kazakh, Malay, Tamil, Telugu, Thai, Yoruba, Chinese, Urdu
XNLI	French	Turkish	Urdu	Bulgarian, Greek, Turkish	Arabic, Thai, Swahili, Urdu, Hindi
TyDiQA	Finnish	Arabic	Bengali	Bengali, Telugu	Swahili, Bengali, Korean

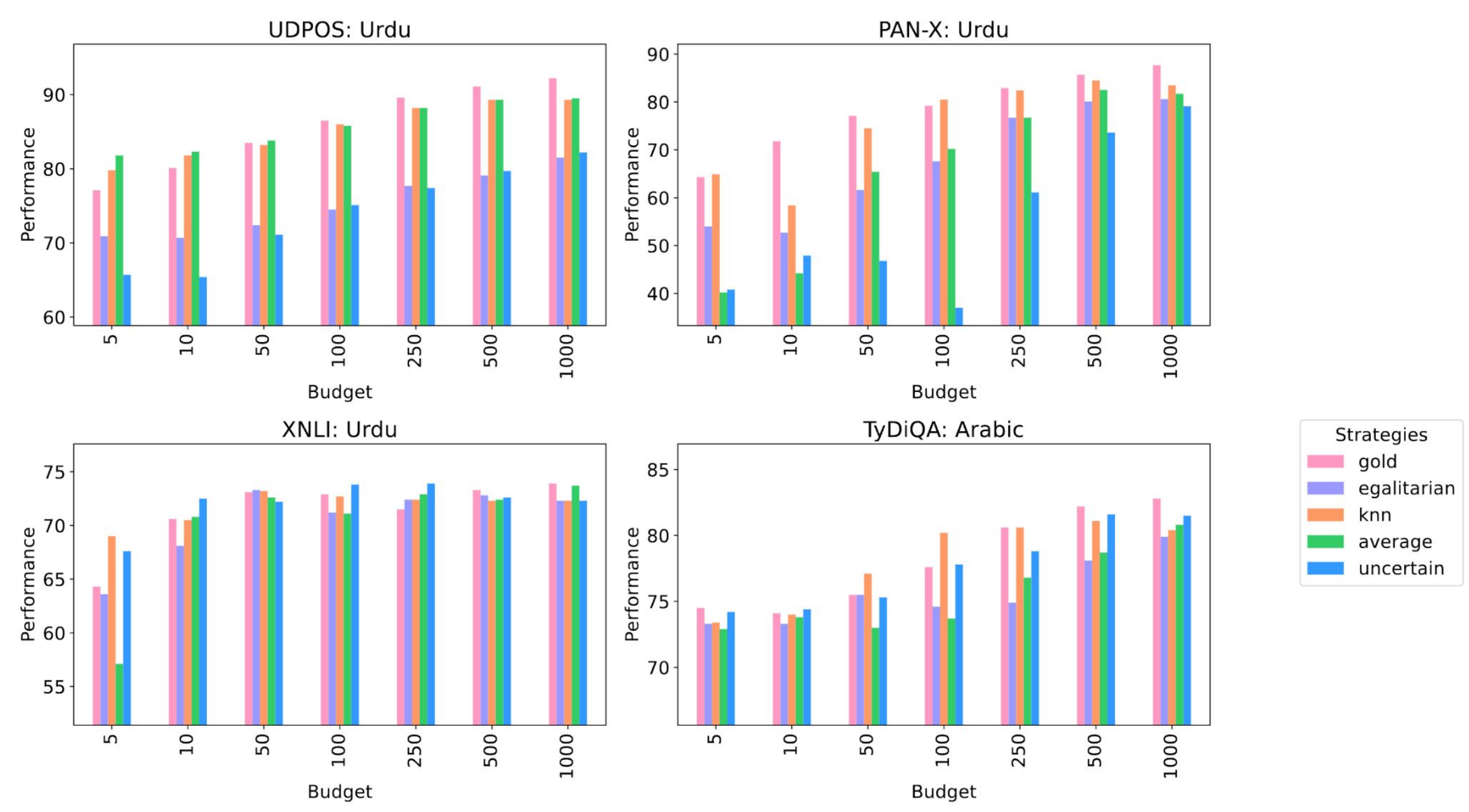
Models
XLM-R
RemBERT
InfoXLM

10,000 Budget, Five AL Rounds (XLM-R)

Dataset	Strategy	High-perf.	Mid-perf.	Low-perf.	Geo Pool	Low-perf. Pool
NER	EN-FT	80	79.5	65.6	61	45.8
	GOLD	90.1	92.8	94.5	81.2	73.7
	$BASE_{EGAL}$	85.4	87.6	84	80.6	62.8
	$DeMuX_{KNN}$	87.8	89.2	85.8	82.4	62.3
	Δ_{BASE}	**2.4**	**1.6**	**1.8**	**1.8**	**-0.5**
XNLI	EN-FT	81.8	77.3	69.9	80.1	73.4
	GOLD	81.6	79.5	70.3	81.6	76
	$BASE_{EGAL}$	81.6	78.8	73	80.9	75.6
	$DeMuX_{AVG}$	83.7	79.9	75.3	82.2	77.1
	Δ_{BASE}	**2.1**	**1.1**	**2.3**	**1.3**	**1.5**
TyDiQA	EN-FT	78.9	73.2	79.9	80.7	78.5
	GOLD	81.2	83.8	83.7	84.7	81
	$BASE_{EGAL}$	79.9	81.7	79.6	81.1	78.7
	$DeMuX_{UNC}$	80.8	82.9	80.3	81	77.8
	Δ_{BASE}	**0.9**	**1.2**	**0.7**	**-0.1**	**-0.9**

Key Takeaway: DeMuX beats baseline in **84%** cases; proximity to target **(lower distance)** matters more for token-level tasks; informativeness **(higher uncertainty)** matters more for QA

Multiple Budgets, One AL Round (XLM-R)



Key Takeaway: Benefits are higher for lower budgets with diminishing returns

4 Further Analysis and Discussion

How does DeMuX fare on multilingual target pools?

- gains over baseline, but smaller on average than single target

Does the model select data from the same languages across tasks?

- No, eg.: for Urdu, data chosen from Hindi for NLI/POS; and Farsi/Arabic (script similarity) for NER.

What is the minimum budget for which we can observe gains in one AL round?

- Gains of up to 8-11 F1 for token-level, and 2-5 F1 points for NLI and QA
- Gains diminish as the budget increases

Do the selected data points matter or does following the language distribution suffice?

- performance declines when you replace selected data points with random data points while following the language distribution of selected points.

Extended to generation tasks like MT (supported in github repo)

Testing on **unseen languages** during pre-training

- Model: MuRIL (trained on Indian languages)
- Languages:
  - Afrikaans (seen script)
  - Bulgarian (unscreen script but characters present in vocab)

Target	Top-3 langs selected by DeMuX	Baseline	DeMuX F1
Afrikaans	German:28%, Estonian:19%, Finnish:13%	77.0	**78.1**
Bulgarian	Russian:81%, Greek:4.8%, Georgian:3.4%	39.9	**51.9**

5 Contact & Resources

Paper **Code**

Thanks!

Please contact skhanuja@cs.cmu.edu to follow up!

https://github.com/simran-khanuja/demux