MergeDistill

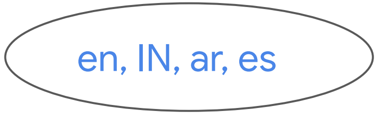
Merging Pre-trained Language Models Using Distillation

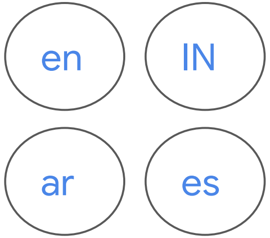
Simran Khanuja, Melvin Johnson, Partha Talukdar

ACL 2021 Findings

Google Research

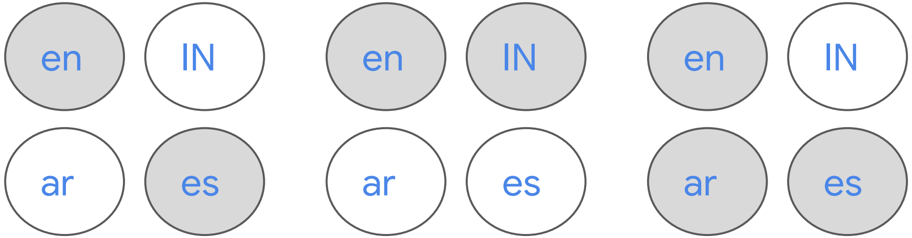
Motivation





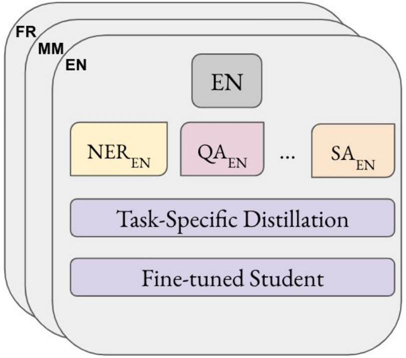
- Zero-shot task transfer
- Positive transfer among related languages
- Ability to handle code-mixed text
- Reduced maintenance costs

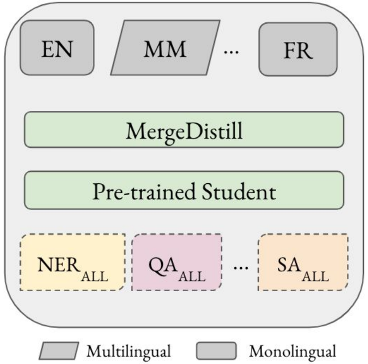
- Large amounts of domain/language specific pre-training data
- Linguistically informed vocabulary
- Custom model architectures, learning objectives



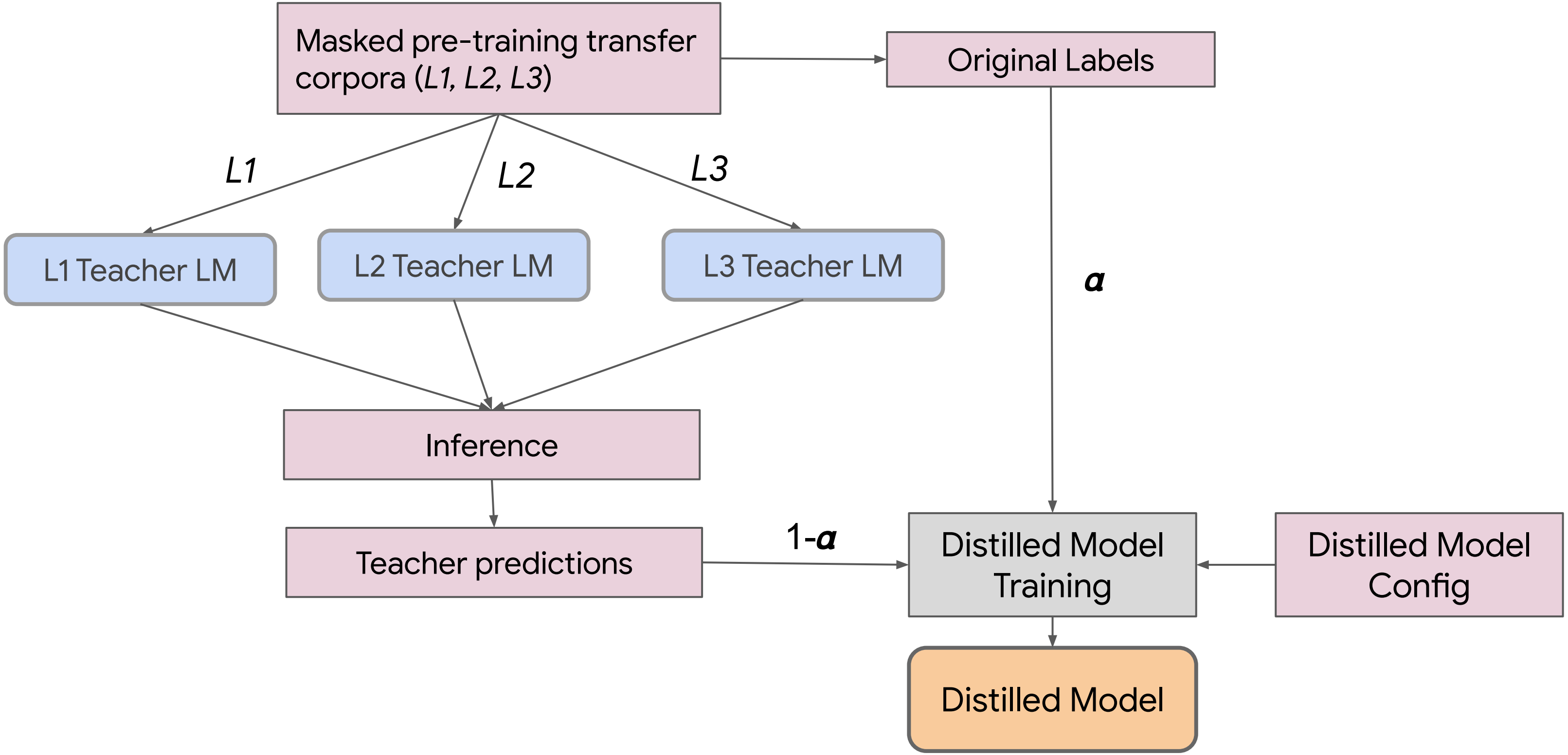
Distillation

Model Stage	Task Type	No. of Teacher LMs	Past Work
Fine-tuning	Task-specific	Single	Tang et al., 2019; Kaliamoorthi et al., 2021
		Multiple	Clark et al., 2019; Turc et al., 2019
Pre-training	Task-agnostic	Single	Sanh et al., 2019; Sun et al., 2020, 2019
		Multiple	❌



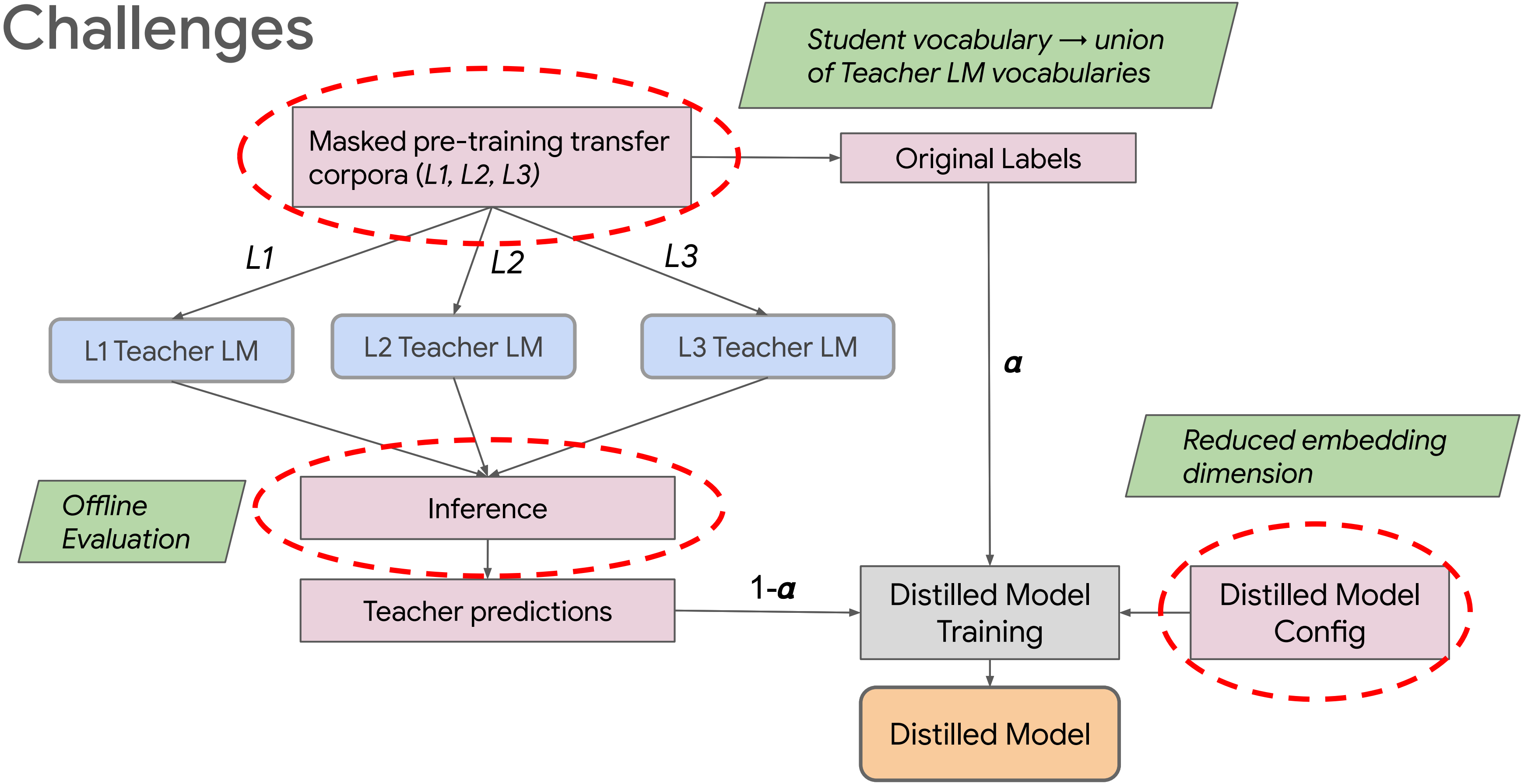


Proposal



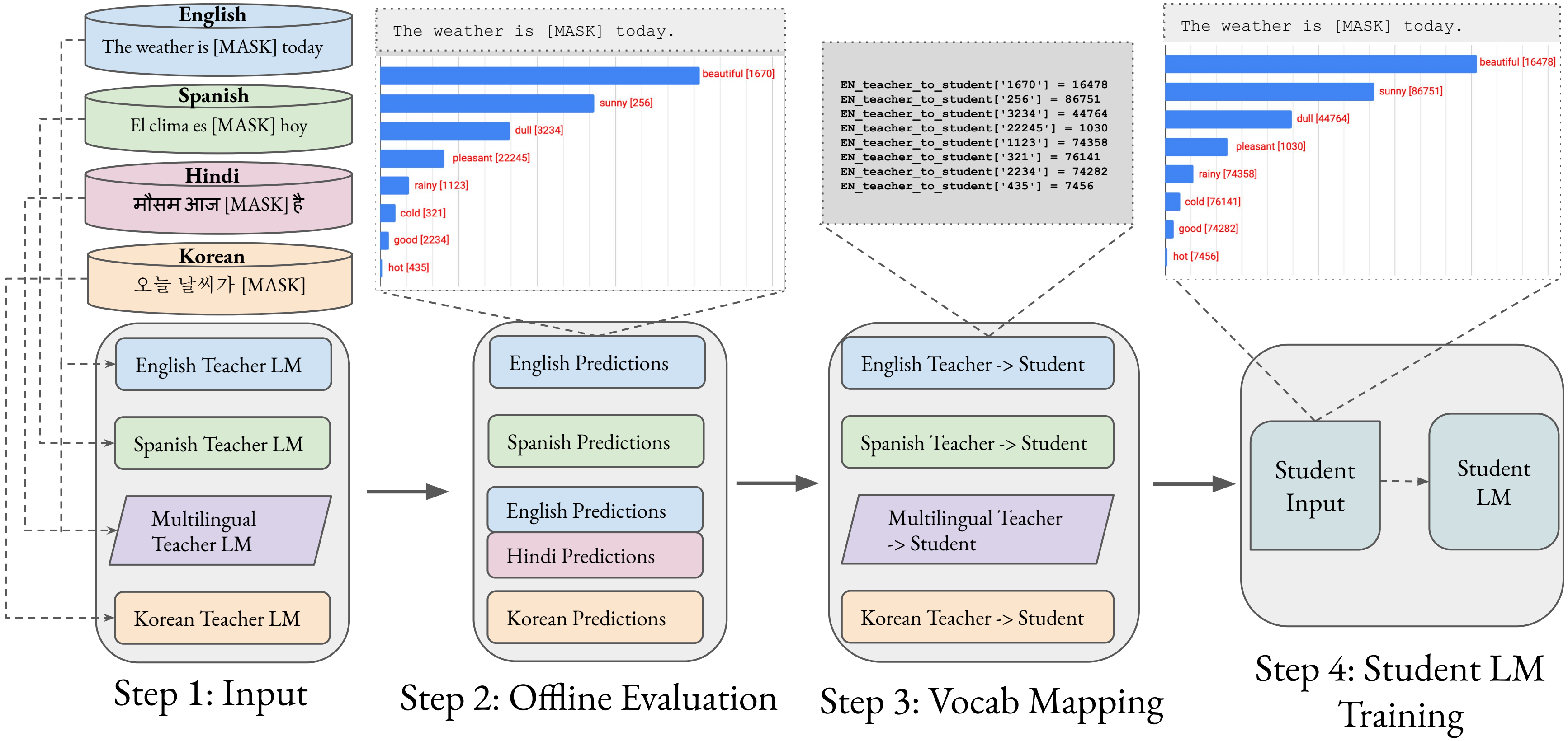
Source

Challenges

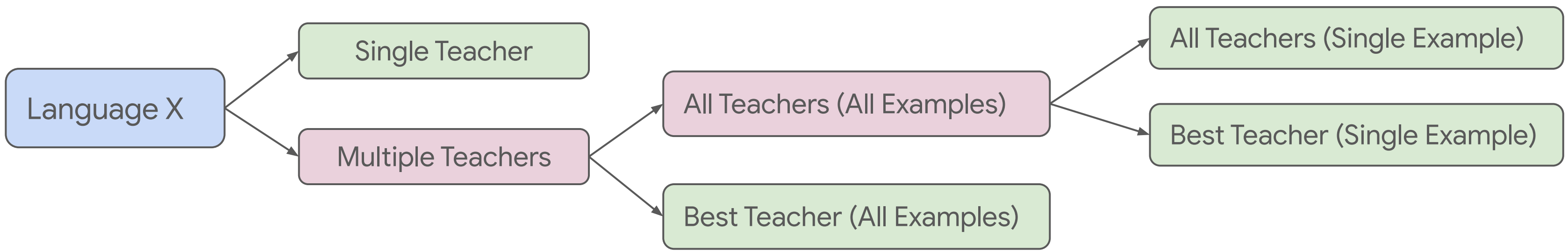


Source

MergeDistill Framework



Experiments

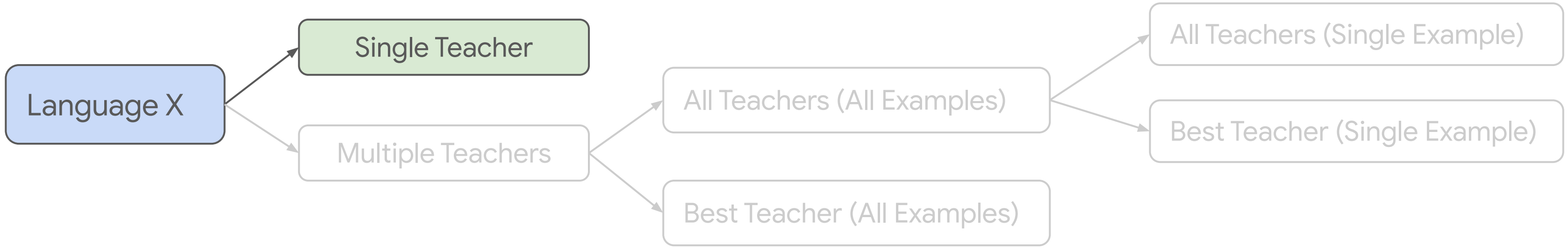


Experiments

Setup

- Pre-training Transfer Corpora : Wikipedia
- Model Size : ~mBERT model size (178M parameters)
- Distillation Parameters :
  - k value in top-k logits is set to 8
  - Teacher Annealing

Experiments



Experiments

Q1) How effective is MergeDistill in combining **disjoint language set** teacher LMs, to train a **multilingual** student LM that leverages the benefits of multilinguality while performing competitively with individual teacher LMs?

Student	Language	Language Family	Model
$\text{Student}_{\text{similar}}$	English	Indo-European	BERT(Devlin et al., 2019)
	German	Indo-European	DeepSet(Chan et al., 2020)
	Italian	Indo-European	ItalianBERT(Schweter, 2020b)
	Spanish	Indo-European	BETO(Cañete et al., 2020)
$\text{Student}_{\text{dissimilar}}$	Arabic	Afroasiatic	AraBERT(Antoun et al., 2020)
	English	Indo-European	BERT(Devlin et al., 2019)
	Finnish	Uralic	FinBERT(Virtanen et al., 2019)
	Turkish	Turkic	BERTurk(Schweter, 2020a)
	Chinese	Sino-Tibetan	ChineseBERT(Devlin et al., 2019)

Experiments

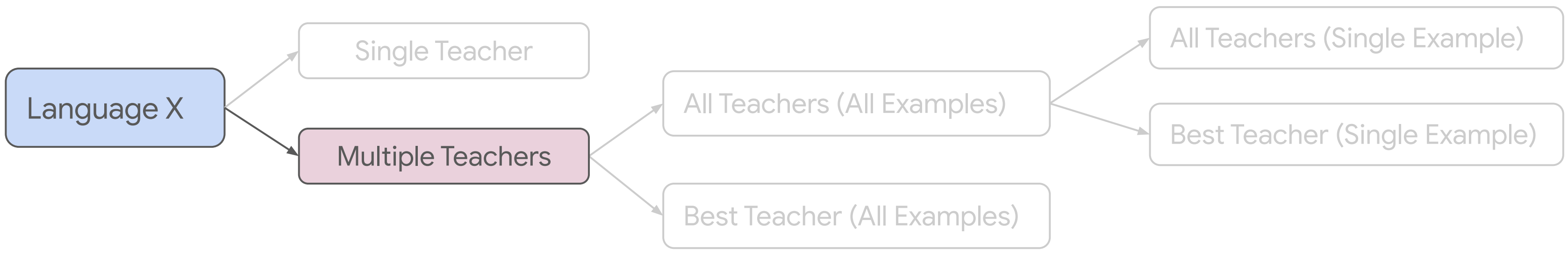
Student	Language	Teacher LM Tokens	Student LM Tokens	% of Data
$\text{Student}_{\text{similar}}$	English	3300M	2285M	69.25%
	German	23723M	847M	3.57%
	Italian	13139M	506M	3.85%
	Spanish	3000M	639M	21.31%
	Total	**43162M**	**4277M**	**9.9%**
$\text{Student}_{\text{dissimilar}}$	Arabic	8600M	135M	1.58%
	English	3300M	2285M	69.25%
	Finnish	3000M	83M	2.77%
	Turkish	4405M	60M	1.36%
	Chinese	71M	71M	100.00%
	Total	**19376M**	**2634M**	**13.6%**

MergeDistill can train multilingual Student LMs competitive with their monolingual counterparts using ~10% of pre-training data!

Language	Model	NER F1	UDPOS F1	QA F1/EM
English	BERT	89.5	96.6	87.1/78.6
	$\text{Student}_{\text{similar}}$	89.8	96.3	89.8/82.1
German	DeepsetBERT	93.0	98.3	-
	$\text{Student}_{\text{similar}}$	93.9	98.3	-
Italian	ItalianBERT	94.5	98.6	73.5/61.6
	$\text{Student}_{\text{similar}}$	95.2	98.6	75.8/63.8
Spanish	BETO	94.2	99.0	74.9/56.6
	$\text{Student}_{\text{similar}}$	94.7	98.9	76.5/58.4
	RDT(%)	**+0.6**	**-0.1**	**+2.8/+3.7**
Arabic	AraBERT	94.3	96.3	83.1/68.6
	$\text{Student}_{\text{dissimilar}}$	93.7	96.4	81.3/66.6
Chinese	ChineseBERT	83.0	96.9	81.8/81.8
	$\text{Student}_{\text{dissimilar}}$	82.6	96.8	80.8/80.8
English	BERT	89.5	96.6	87.1/78.6
	$\text{Student}_{\text{dissimilar}}$	89.5	96.3	88.6/80.7
Finnish	FinBERT	94.4	97.9	81.0/68.8
	$\text{Student}_{\text{dissimilar}}$	94.4	95.5	77.7/65.9
Turkish	BERTurk	95.2	95.6	76.7/59.8
	$\text{Student}_{\text{dissimilar}}$	95.4	92.9	76.2/59.1
	RDT(%)	**-0.2**	**-1.1**	**-1.3/-1.4**

$$\mathrm{RDT}(\mathrm{S}, \{\mathrm{T}_1, \ldots, \mathrm{T}_n\}) = \frac{100}{n} \sum_{i=1}^{n} \frac{(\mathrm{P}_{\mathrm{T}_i} - \mathrm{P}_{\mathrm{S}})}{\mathrm{P}_{\mathrm{T}_i}}$$

Experiments



Experiments

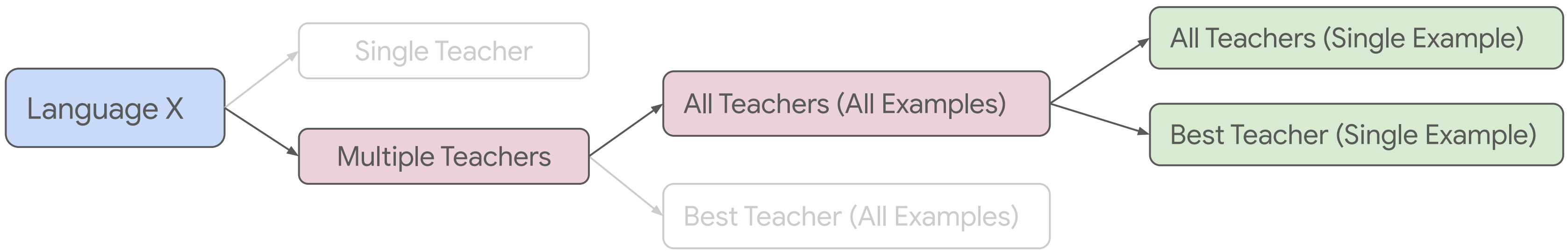
Q2) How effective is MergeDistill in combining **multilingual** teacher LMs, trained on an **overlapping set** of languages, such that each language can benefit from *multiple* teachers?

Setup :
Combine mBERT and MuRIL using Wikipedia text as our pre-training transfer corpora.

Languages	Teacher LMs
Non MuRIL	mBERT
MuRIL	mBERT, MuRIL

Teacher	Language	Teacher LM Tokens	Student LM Tokens	% of Data
MuRIL	Bengali	1181M	27M	2.30%
	English	6986M	2816M	40.30%
	Gujarati	173M	7M	3.90%
	Hindi	2368M	38M	1.61%
	Kannada	196M	15M	7.64%
	Malayalam	337M	14M	4.17%
	Marathi	274M	8M	3.02%
	Nepali	231M	5M	2.16%
	Punjabi	141M	9M	6.45%
	Tamil	769M	26M	3.34%
	Telugu	331M	30M	8.99%
	Urdu	722M	23M	3.21%
	Total	**13709M**	**3018M**	**22%**

Experiments

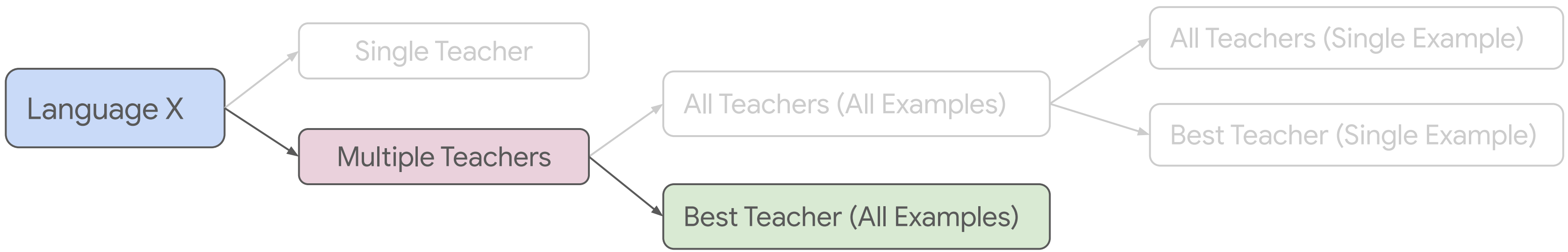


Experiments

Languages	Model	Teacher	PANX F1	UDPOS F1	PAWSX Acc.	XNLI Acc.	XQUAD F1/EM	MLQA F1/EM	TyDiQA F1/EM	Avg.
MuRIL Languages	mBERT	-	58.8	68.5	93.4	66.2	70.3/57.5	65.0/50.8	62.5/52.	69.2
	MuRIL	-	76.9	74.5	95.0	74.4	77.7/64.2	73.6/58.6	76.1/60.2	78.3
	$\text{Student}_{\text{MuRIL}}$	MuRIL	69.3	72.3	95.4	71.9	75.7/62.1	72.0/56.3	70.7/59.2	75.3
	$\text{Student}_{\text{mBERT}}$	mBERT	38.1	52.1	93.5	64.8	56.9/44.8	51.1/39.7	41.6/33.9	56.9
	$\text{Student}_{\text{Both\_all}}$	mBERT + MuRIL	67.9	72.3	94.5	71.1	76.1/62.9	70.4/55.5	70.8/55.3	74.7
	$\text{Student}_{\text{Both\_best}}$	mBERT + MuRIL	68.5	71.5	93.9	70.7	77.7/64.3	70.8/55.6	70.6/58.4	74.8
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{mBERT})$ (%)		**+17.9**	**+5.6**	**+2.1**	**+8.6**	**+7.7/+8**	**+10.8/+10.8**	**+13.1/+12.3**	**+8.8**
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{MuRIL})$ (%)		-9.9	**-3**	**+0.4**	**-3.4**	**-2.6/-3.3**	**-2.2/-3.9**	-7.1/**-1.7**	**-3.8**
Non MuRIL Languages	mBERT	-	63.5	71.1	80.2	65.9	62.2/47.1	59.7/41.4	60.4/46.1	66.1
	$\text{Student}_{\text{MuRIL}}$	mBERT	63.9	72.8	83.3	68.7	66.5/51.2	63.1/44.4	61.7/45.0	68.6
	$\text{Student}_{\text{mBERT}}$	mBERT	64.6	72.1	84.0	68.8	64.5/49.0	61.1/42.7	58.9/44.1	67.7
	$\text{Student}_{\text{Both\_all}}$	mBERT	64.1	72.6	83.9	68.1	61.3/47.1	60.5/42.2	59.7/44.0	67.2
	$\text{Student}_{\text{Both\_best}}$	mBERT	63.3	72.6	83.2	67.2	66.0/50.6	61.4/43.2	62.4/46.5	68.0
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{mBERT})$ (%)		**+0.6**	**+2.4**	**+3.9**	**+4.3**	**+6.9/+8.7**	**+5.7/+7.2**	**+2.2/-2.4**	**+3.8**

We don't observe a significant change in performance for Student_both variants.

Experiments



Experiments

Languages	Model	Teacher	PANX F1	UDPOS F1	PAWSX Acc.	XNLI Acc.	XQUAD F1/EM	MLQA F1/EM	TyDiQA F1/EM	Avg.
MuRIL Languages	mBERT	-	58.8	68.5	93.4	66.2	70.3/57.5	65.0/50.8	62.5/52.	69.2
	MuRIL	-	76.9	74.5	95.0	74.4	77.7/64.2	73.6/58.6	76.1/60.2	78.3
	$\text{Student}_{\text{MuRIL}}$	MuRIL	69.3	72.3	95.4	71.9	75.7/62.1	72.0/56.3	70.7/59.2	75.3
	$\text{Student}_{\text{mBERT}}$	mBERT	38.1	52.1	93.5	64.8	56.9/44.8	51.1/39.7	41.6/33.9	56.9
	$\text{Student}_{\text{Both\_all}}$	mBERT + MuRIL	67.9	72.3	94.5	71.1	76.1/62.9	70.4/55.5	70.8/55.3	74.7
	$\text{Student}_{\text{Both\_best}}$	mBERT + MuRIL	68.5	71.5	93.9	70.7	77.7/64.3	70.8/55.6	70.6/58.4	74.8
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{mBERT})$ (%)		**+17.9**	**+5.6**	**+2.1**	**+8.6**	**+7.7/+8**	**+10.8/+10.8**	**+13.1/+12.3**	**+8.8**
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{MuRIL})$ (%)		-9.9	**-3**	**+0.4**	**-3.4**	**-2.6/-3.3**	**-2.2/-3.9**	-7.1/**-1.7**	**-3.8**
Non MuRIL Languages	mBERT	-	63.5	71.1	80.2	65.9	62.2/47.1	59.7/41.4	60.4/46.1	66.1
	$\text{Student}_{\text{MuRIL}}$	mBERT	63.9	72.8	83.3	68.7	66.5/51.2	63.1/44.4	61.7/45.0	68.6
	$\text{Student}_{\text{mBERT}}$	mBERT	64.6	72.1	84.0	68.8	64.5/49.0	61.1/42.7	58.9/44.1	67.7
	$\text{Student}_{\text{Both\_all}}$	mBERT	64.1	72.6	83.9	68.1	61.3/47.1	60.5/42.2	59.7/44.0	67.2
	$\text{Student}_{\text{Both\_best}}$	mBERT	63.3	72.6	83.2	67.2	66.0/50.6	61.4/43.2	62.4/46.5	68.0
	$\text{RDT}(\text{Student}_{\text{MuRIL}}, \text{mBERT})$ (%)		**+0.6**	**+2.4**	**+3.9**	**+4.3**	**+6.9/+8.7**	**+5.7/+7.2**	**+2.2/-2.4**	**+3.8**

Student_MuRIL performs the best for all languages. It beats mBERT while remaining in a RDT of 5% with MuRIL.

Experiments

Q3) How important are the teacher LM vocabulary and predictions in MergeDistill?

Model	Vocabulary	Labels	PANX	UDPOS	PAWSX	XNLI	XQUAD	MLQA	TyDiQA	Avg.
SM1	mBERT	Gold	63.2	73.0	94.8	71.2	70.2/57.9	65.1/51.3	60.8/48.7	71.2
SM2	mBERT∪MuRIL	Gold	**69.3**	**73.9**	95.3	71.2	**76.2/63.1**	71.1/56.0	70.9/56.0	**75.4**
SM3	mBERT∪MuRIL	Gold+Teacher	**69.3**	72.3	**95.4**	**71.9**	75.7/62.1	**72.0/56.3**	**70.7/59.2**	75.3
SM2_100k	mBERT∪MuRIL	Gold	65.5	72.3	94.3	67.5	72.3/58.2	66.9/51.5	62.5/51.9	71.6
SM3_100k	mBERT∪MuRIL	Gold+Teacher	71.2	73.5	93.1	69.6	76.4/62.9	69.1/53.9	68.6/54.9	74.5

Competent tokenizers play an important role in MergeDistill to boost student LM performance.

Teacher LM predictions help speed-up model convergence time by ~5x!

Conclusion

- MergeDistill is a first attempt at combining pre-trained LMs using *task-agnostic* distillation.
- **Benefits** :
  - More maintainability (less models)
  - Compute efficient (offline predictions)
  - Exploits benefits of multilinguality and language-specific LMs
- **Results** :
  - Student LMs competitive with teacher LMs, despite being trained on much less data
  - Training time speed-up by almost 5x without loss in performance with teacher labels!
- **Future Work** :
  - Experimenting with extreme resource-lean scenarios (data and training steps) to test effectiveness of MergeDistill, with a potential higher impact.
  - Other methods of learning student vocabulary, rather than taking a union of teacher LM vocabularies