You Can Have Your Data and Balance It Too: Towards Balanced and Efficient Multilingual Models

Dan Malkin*, Tomasz Limisiewicz*, Gabriel Stanovsky (* equal contribution)



Multilingual Language Models are go to solution for cross-lingual transfer. However, the performance on some languages of interest is hindered by their underrepresentation in training data.

“

Motivation!

Multilingual Language Models are go to solution for cross-lingual transfer. However, the performance on some languages of interest is hindered by their underrepresentation in training data.

How do we improve performance on low-resource, while preserving the good performance on high-resource?

“ **Motivation!**

Approach

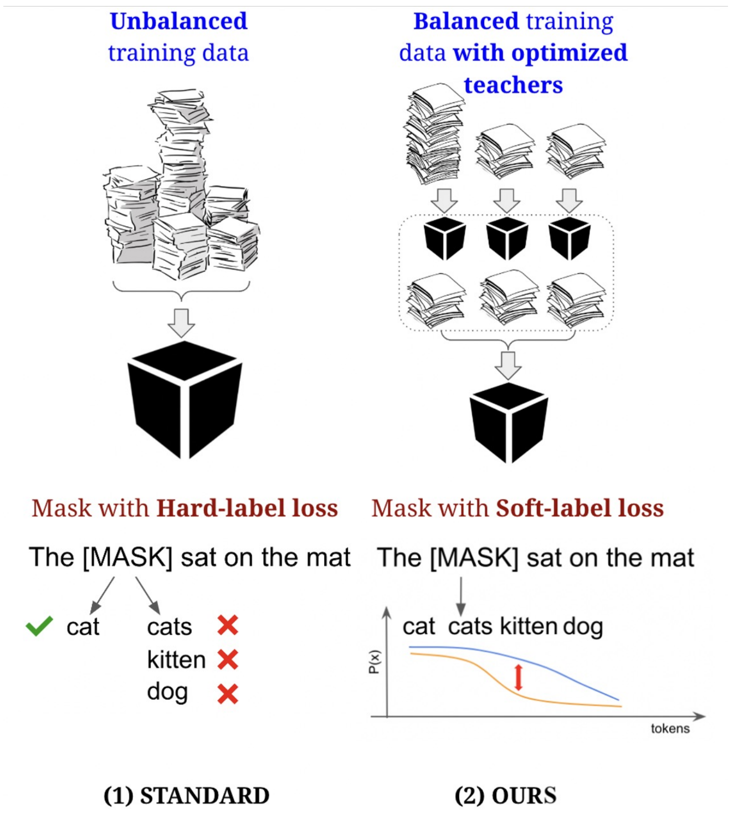
- We train multilingual language model performing well on languages with small digital footprint (low-resource)
- Preserve good performance on high-resource languages

(1) STANDARD

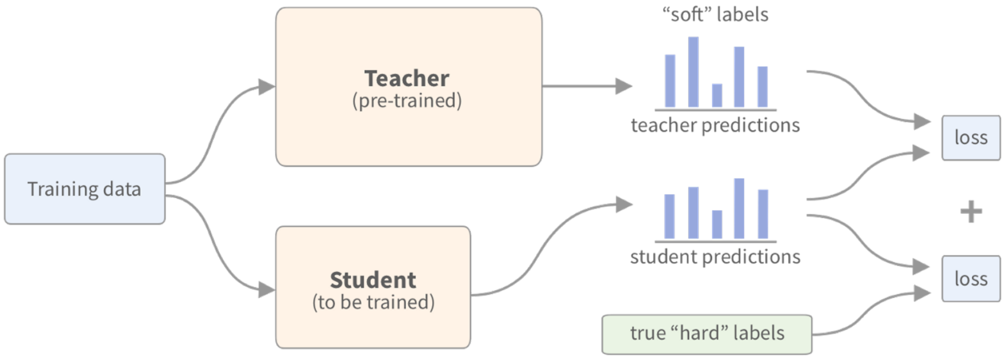
Approach

- We train multilingual language model performing well on languages with small digital footprint (low-resource)

- Preserve good performance on high-resource languages

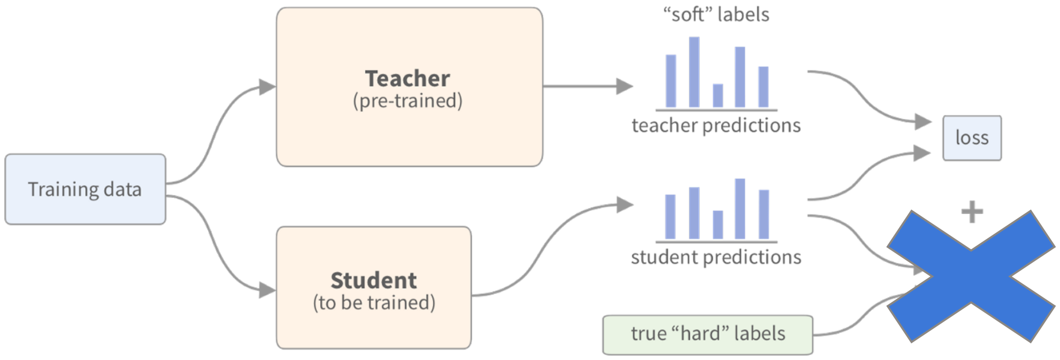


Background: Soft Label Knowledge Distillation



Soft-target distillation used e.g. in DistilledBERT Sanh et al. 2019

Background: Soft Label Knowledge Distillation

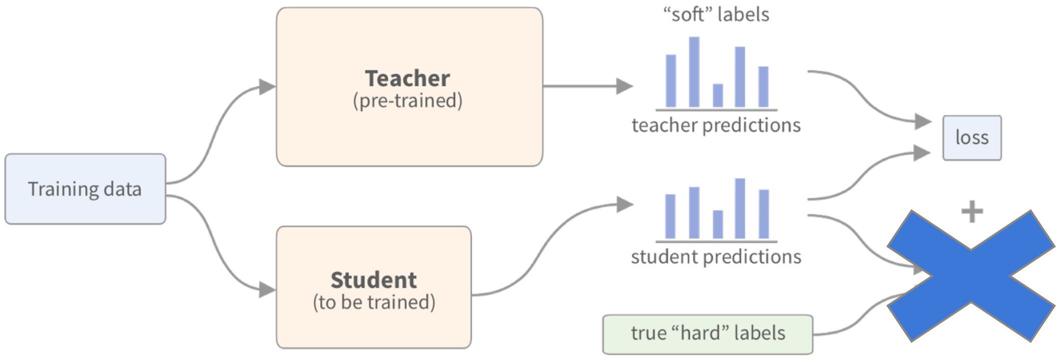


Soft-target distillation used e.g. in DistilledBERT Sanh et al. 2019

Our Approach:

- Use only "soft" labels for student

Background: Soft Label Knowledge Distillation

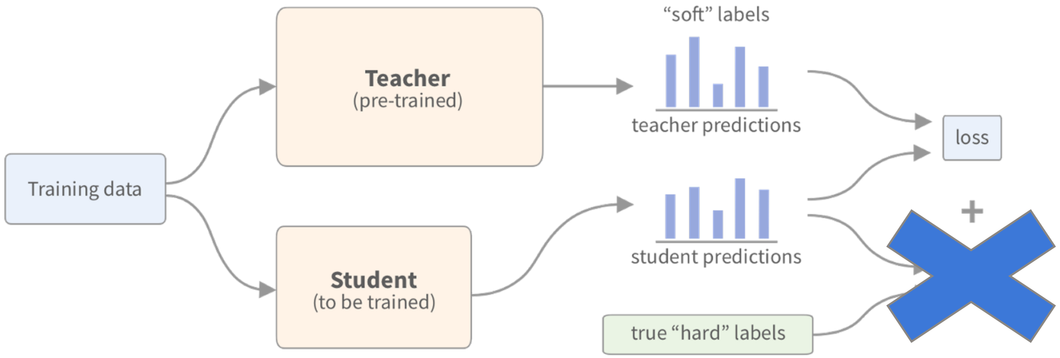


Soft-target distillation used e.g. in DistilledBERT Sanh et al. 2019

Our Approach:

- Use only "soft" labels for student
- Use many monolingual teachers

Background: Soft Label Knowledge Distillation



Soft-target distillation used e.g. in DistilledBERT Sanh et al. 2019

Our Approach:

- Use only "soft" labels for student
- Use many monolingual teachers
- Scale down data not model size

Experiments

Baselines

Hard-Labels: all data

Hard-Labels: balanced

Ours: Soft-Labels
all data ➡ balanced

Languages

Shared scrip (Latin):
eu, tr, vi, hu, es, **de**, en

Diverse script:
te, ur, hi, el, ko, ru, **de**

Evaluation

Language modeling

Part of Speech

Named Entity Recognition

Experiments

Baselines	Languages	Evaluation
Hard-Labels: all data	Shared scrip (Latin): eu, tr, vi, hu, es, **de**, en	Language modeling
Hard-Labels: balanced	Diverse script: te, ur, hi, el, ko, ru, **de**	Part of Speech
Ours: **Soft-Labels** all data ➡ balanced		Named Entity Recognition

Experiments

Baselines

Hard-Labels: all data

Hard-Labels: balanced

Ours: Soft-Labels
all data ➡ balanced

Languages

Shared scrip (Latin):
eu, tr, vi, hu, es, **de**, en

Diverse script:
te, ur, hi, el, ko, ru, **de**

Evaluation

Language modeling

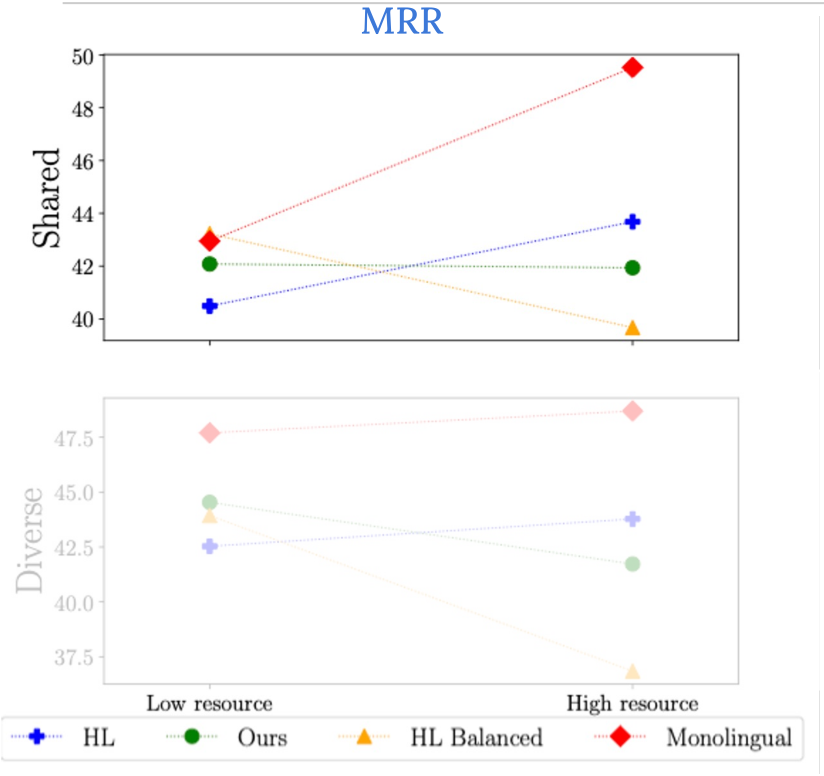
Part of Speech

Named Entity Recognition

Results: Language Modeling

- Our approach achieves good balance of scores between low- and high-resource languages.

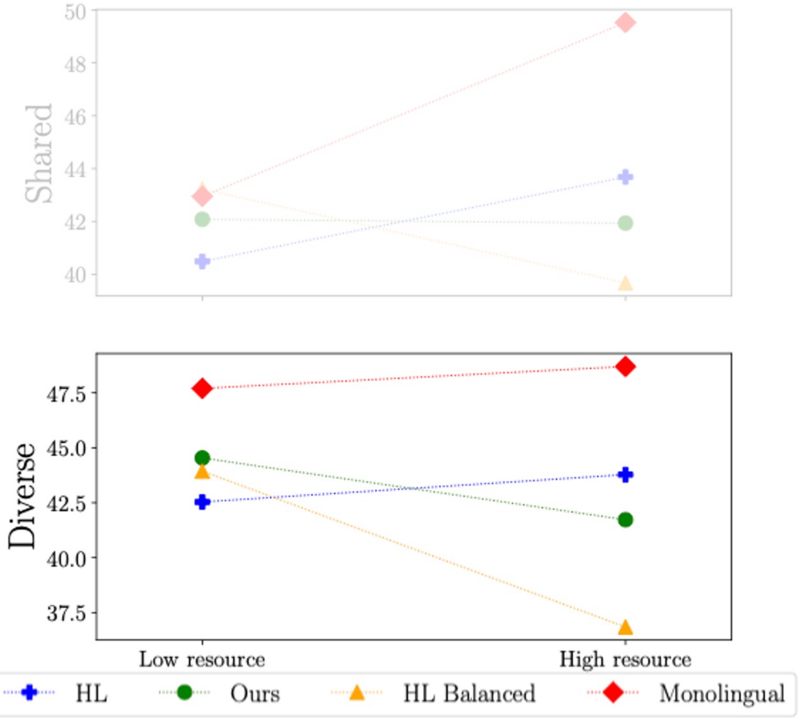
- For low-resource naive balancing significantly improves results but it hinders performance for high-resource languages.



Results: Language Modeling

- Our approach achieves good balance of scores between low- and high-resource languages.
- For low-resource naive balancing significantly improves results but it hinders performance for high-resource languages.

MRR



Results: Zero Shot

- Our approach performs best in POS

POS

	Lang. set	HL		HL balanced		Ours	
		I-L	Z-S	I-L	Z-S	I-L	Z-S
Shared	Low-Res	35.2	33.4	35.5	34.3	36.6	**34.5**
	High-Res	83.3	33.7	81.2	32.4	84.3	**33.8**
	{de}	**87.1**	32.3	84.1	32.2	86.8	**33.0**
	All	55.8	33.5	55.1	33.5	**57.0**	**34.2**
Diverse	Low-Res	53.1	35.8	54.6	34.9	**55.7**	**35.9**
	High-Res	76.8	36.2	73.4	34.7	**77.3**	**36.8**
	{de}	**87.7**	36.8	83.3	35.3	87.4	**38.1**
	All	63.3	36.0	62.7	34.8	**64.9**	**36.3**

Results: Zero Shot

- Our approach performs best in POS and the most cases of NER zero-shot transfers.

POS

	Lang. set	HL		HL balanced		Ours	
		I-L	Z-S	I-L	Z-S	I-L	Z-S
Shared	Low-Res	35.2	33.4	35.5	34.3	36.6	**34.5**
	High-Res	83.3	33.7	81.2	32.4	84.3	**33.8**
	{de}	**87.1**	32.3	84.1	32.2	86.8	**33.0**
	All	55.8	33.5	55.1	33.5	57.0	**34.2**
Diverse	Low-Res	53.1	35.8	54.6	34.9	55.7	**35.9**
	High-Res	76.8	36.2	73.4	34.7	77.3	**36.8**
	{de}	**87.7**	36.8	83.3	35.3	87.4	**38.1**
	All	63.3	36.0	62.7	34.8	64.9	**36.3**

NER

	Lang. set	HL		HL balanced		Ours	
		I-L	Z-S	I-L	Z-S	I-L	Z-S
Shared	Low-Res	26.5	23.7	27.9	**24.3**	29.8	23.9
	High-Res	34.2	24.9	34.7	24.7	37.6	**26.0**
	{de}	31.4	**27.4**	32.1	25.7	32.0	23.9
	All	29.8	24.2	30.8	24.5	33.1	**24.8**
Diverse	Low-Res	25.7	12.8	28.0	**13.8**	29.9	12.9
	High-Res	32.8	14.9	29.9	15.1	37.2	**17.1**
	{de}	32.5	14.8	31.5	15.7	35.3	**17.2**
	All	28.7	13.7	28.8	14.4	33.0	**14.7**

Results: In Language

- Our approach performs best in POS and the most cases of NER zero-shot transfers.
- We also overperform in in-language probing results.

POS

	Lang. set	HL		HL balanced		Ours	
		I-L	Z-S	I-L	Z-S	I-L	Z-S
Shared	Low-Res	35.2	33.4	35.5	34.3	**36.6**	34.5
	High-Res	83.3	33.7	81.2	32.4	**84.3**	33.8
	{de}	**87.1**	32.3	84.1	32.2	86.8	33.0
	All	55.8	33.5	55.1	33.5	**57.0**	34.2
Diverse	Low-Res	53.1	35.8	54.6	34.9	**55.7**	35.9
	High-Res	76.8	36.2	73.4	34.7	**77.3**	36.8
	{de}	**87.7**	36.8	83.3	35.3	87.4	38.1
	All	63.3	36.0	62.7	34.8	**64.9**	36.3

NER

	Lang. set	HL		HL balanced		Ours	
		I-L	Z-S	I-L	Z-S	I-L	Z-S
Shared	Low-Res	26.5	23.7	27.9	24.3	**29.8**	23.9
	High-Res	34.2	24.9	34.7	24.7	**37.6**	26.0
	{de}	31.4	27.4	**32.1**	25.7	32.0	23.9
	All	29.8	24.2	30.8	24.5	**33.1**	24.8
Diverse	Low-Res	25.7	12.8	28.0	13.8	**29.9**	12.9
	High-Res	32.8	14.9	29.9	15.1	**37.2**	17.1
	{de}	32.5	14.8	31.5	15.7	**35.3**	17.2
	All	28.7	13.7	28.8	14.4	**33.0**	14.7

Results: Script Groups

- Our approach performs best in POS and the most cases of NER zero-shot transfers.
- We also overperform in in-language probing results.
- Interestingly, in the set of languages with diverse scripts transfer is worse for NER and better for POS, in comparison to same script set.

POS

	Lang. set	HL I-L	HL Z-S	HL balanced I-L	HL balanced Z-S	Ours I-L	Ours Z-S
Shared	Low-Res	35.2	33.4	35.5	34.3	**36.6**	**34.5**
	High-Res	83.3	33.7	81.2	32.4	**84.3**	**33.8**
	{de}	**87.1**	32.3	84.1	32.2	86.8	**33.0**
	All	55.8	33.5	55.1	33.5	**57.0**	**34.2**
Diverse	Low-Res	53.1	35.8	54.6	34.9	**55.7**	**35.9**
	High-Res	76.8	36.2	73.4	34.7	**77.3**	**36.8**
	{de}	**87.7**	36.8	83.3	35.3	87.4	**38.1**
	All	63.3	36.0	62.7	34.8	**64.9**	**36.3**

NER

	Lang. set	HL I-L	HL Z-S	HL balanced I-L	HL balanced Z-S	Ours I-L	Ours Z-S
Shared	Low-Res	26.5	23.7	27.9	**24.3**	**29.8**	23.9
	High-Res	34.2	24.9	34.7	24.7	**37.6**	**26.0**
	{de}	31.4	**27.4**	**32.1**	25.7	32.0	23.9
	All	29.8	24.2	30.8	24.5	**33.1**	**24.8**
Diverse	Low-Res	25.7	12.8	28.0	**13.8**	**29.9**	12.9
	High-Res	32.8	14.9	29.9	15.1	**37.2**	**17.1**
	{de}	32.5	14.8	31.5	15.7	**35.3**	**17.2**
	All	28.7	13.7	28.8	14.4	**33.0**	**14.7**

Summary:

1. Better low-res performance, while high-resource results are preserved.
2. Diverse script can be beneficial for cross-lingual transfer.
3. Future work needed to validate the method on larger models.

“

Thank You
For your Attention!