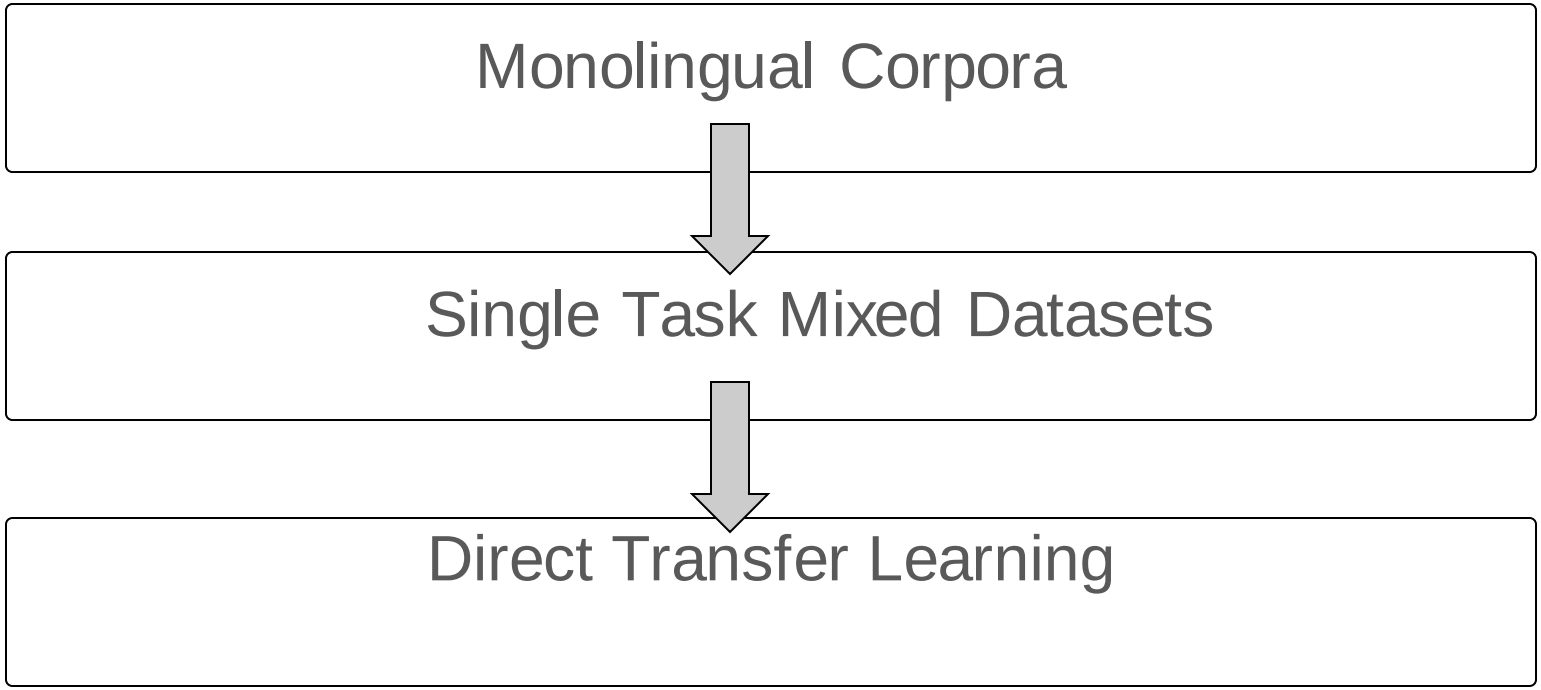
Meta-Learning for Low Resource NMT

Introduction

- Historically Statistical Translation
- Neural Machine Translation recently outperforms
- Statistical Models outperformed translations on low resource language pairs

NMT Previous Work



Meta Learning in NMT

Idea:

Improve on direct transfer learning by better fine-tuning

MAML for NMT

17 High-Resource Languages

Danish Greek
Spanish Italian
French
Portuguese
Greek Polish

4 Low-resource Languages

Turkish Finnish
Romanian Latvian

17 High-Resource Languages

Danish Greek
Spanish Italian
French
Portuguese
Greek Polish

Meta-train on these!

e.g. Spanish → English

4 Low-resource Languages

Turkish Finnish
Romanian Latvian

Meta-test on these!

e.g. Turkish → English

Note: they simulate low-resource by sub-sampling

Gradient Update

$$\boxed{\theta'} = \theta - \eta \nabla_\theta \mathcal{L}^{D^{train}}(\theta)$$

Meta-Gradient Update

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{D^{test}}(\boxed{\theta'})$$

Gradient Update

$$\boxed{\theta'} = \theta - \eta \nabla_\theta \mathcal{L}^{D^{train}}(\theta)$$

1st-order Approximate Meta-Gradient Update

$$\theta \leftarrow \theta - \eta \nabla_{\boxed{\theta'}} \mathcal{L}^{D^{test}}(\boxed{\theta'})$$

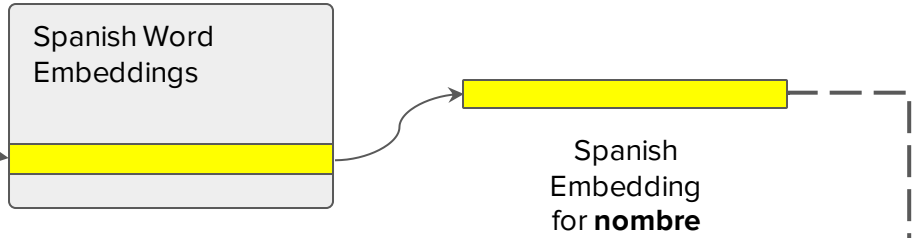
~~Meta-Gradient Update~~

$$\cancel{\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{D^{test}}(\theta')}$$

Issue: Meta-train and meta-test input spaces should match!

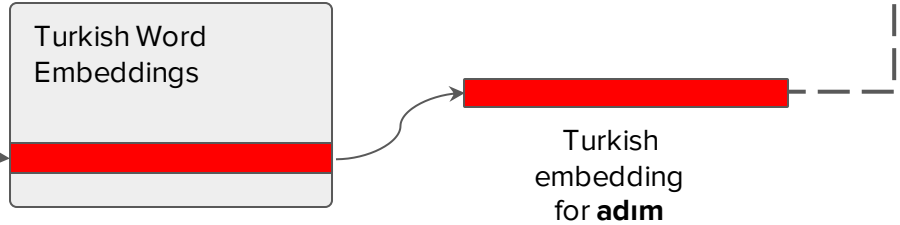
Meta-train

En un lugar de la mancha, de cuyo nombre no puedo... → In some place in the Mancha, whose name...



Meta-test

Benim adım kırmızı... → My name is Red...



Trained independently

Universal Lexical Representation

Word embeddings **trained independently** on monolingual corpora

English Word Embeddings

$$\epsilon^0 \in \mathbb{R}^{|V_0| \times d}$$

Spanish Word Embeddings

$$\epsilon^1 \in \mathbb{R}^{|V_1| \times d}$$

French Word Embeddings

$$\epsilon^2 \in \mathbb{R}^{|V_2| \times d}$$

Turkish Word Embeddings

$$\epsilon^3 \in \mathbb{R}^{|V_3| \times d}$$

Universal Lexical Representation

Word embeddings **trained independently** on monolingual corpora

English Word Embeddings

$\epsilon^0 \in \mathbb{R}^{|V_0| \times d}$

Spanish Word Embeddings

$\epsilon^1 \in \mathbb{R}^{|V_1| \times d}$

French Word Embeddings

$\epsilon^2 \in \mathbb{R}^{|V_2| \times d}$

Turkish Word Embeddings

$\epsilon^3 \in \mathbb{R}^{|V_3| \times d}$

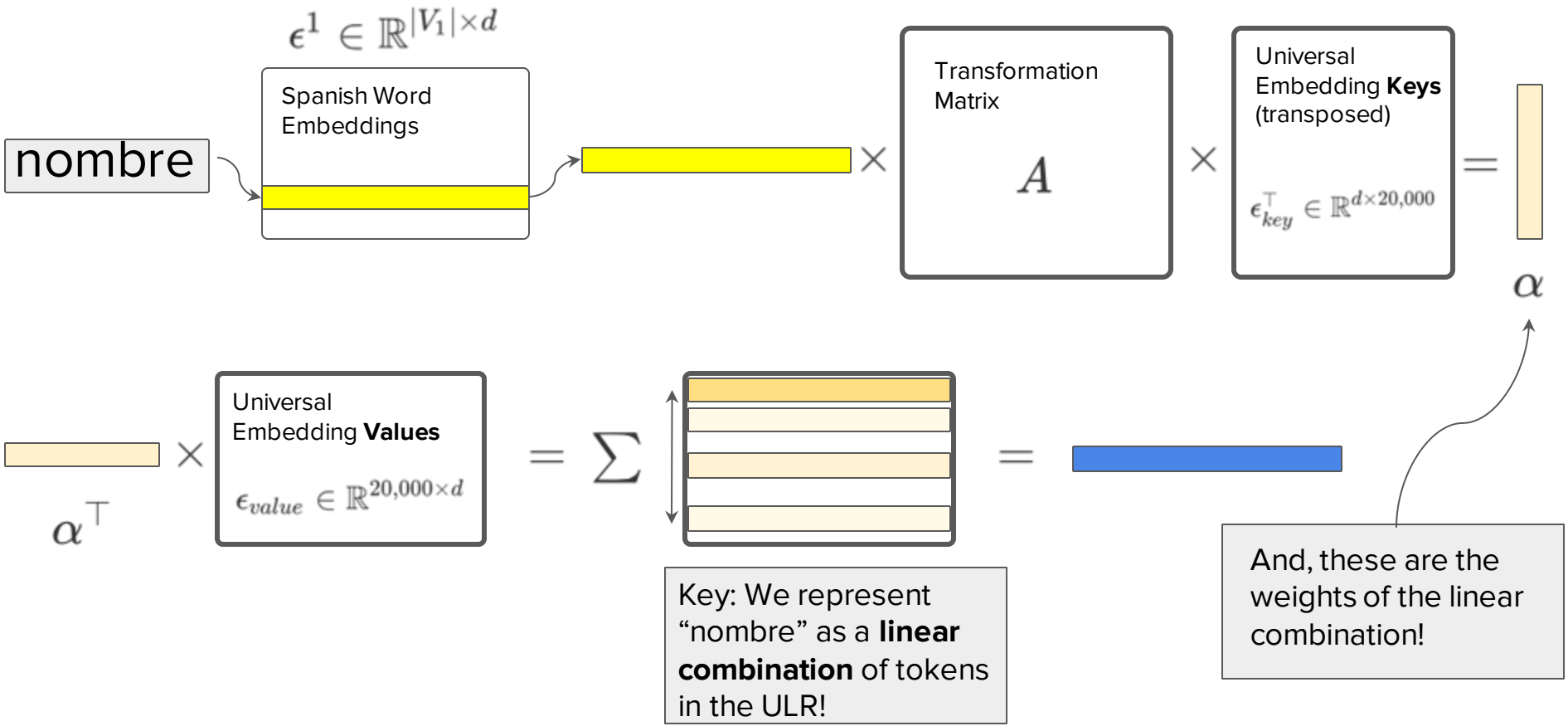
Universal Embedding **Values**

$\epsilon_{value} \in \mathbb{R}^{20{,}000 \times d}$

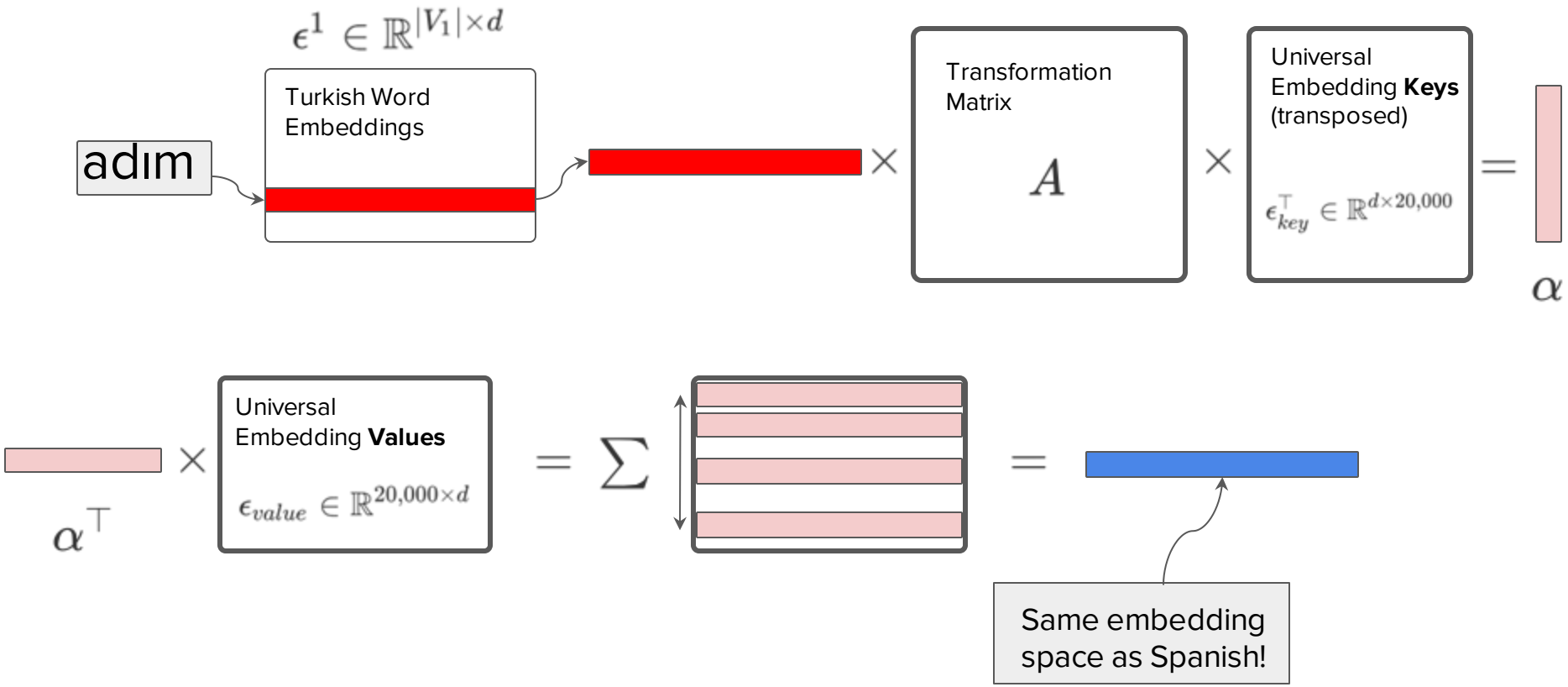
Universal Embedding **Keys**

$\epsilon_{key} \in \mathbb{R}^{20{,}000 \times d}$

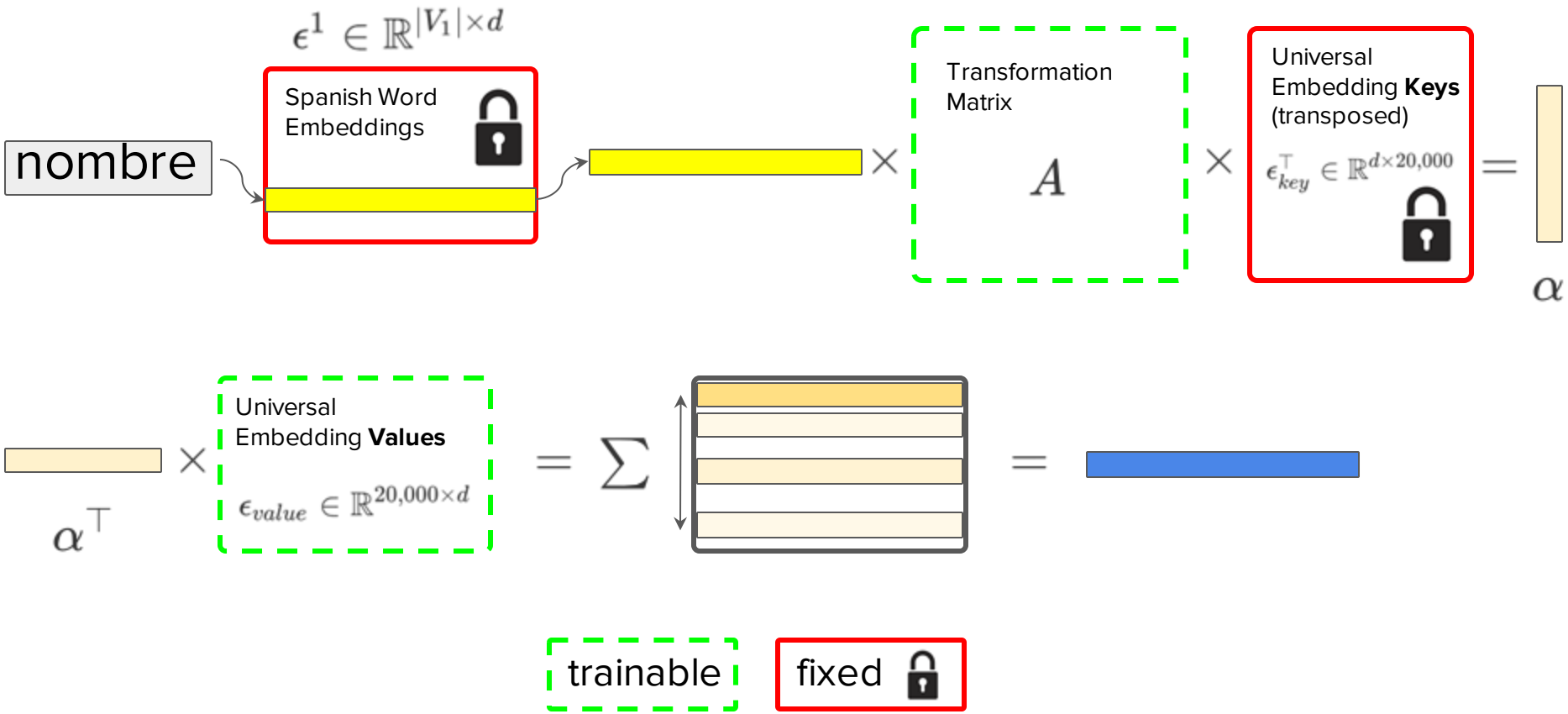
Universal Lexical Representation



Universal Lexical Representation



Training


$\epsilon^1 \in \mathbb{R}^{|V_1| \times d}$
nombre
Spanish Word Embeddings
Transformation Matrix
A
Universal Embedding **Keys** (transposed)
$\epsilon_{key}^\top \in \mathbb{R}^{d \times 20{,}000}$
α
$\alpha^\top$
Universal Embedding **Values**
$\epsilon_{value} \in \mathbb{R}^{20{,}000 \times d}$
trainable
fixed

Experiments

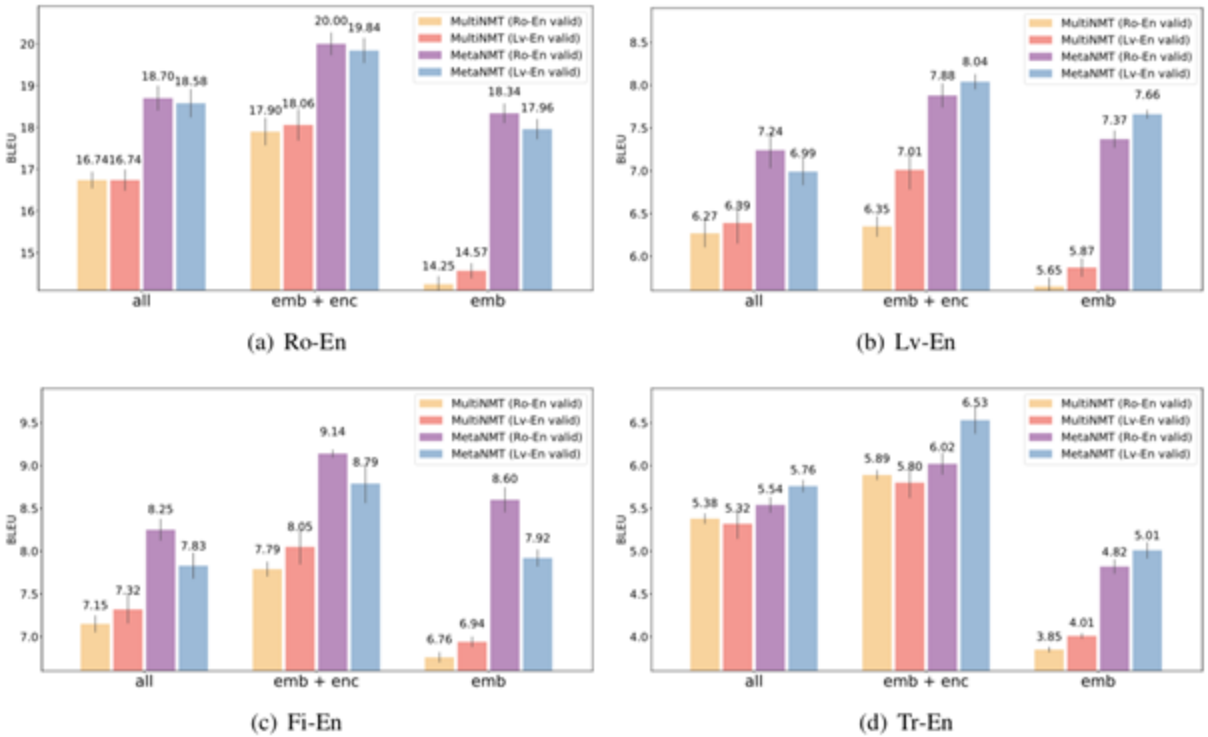
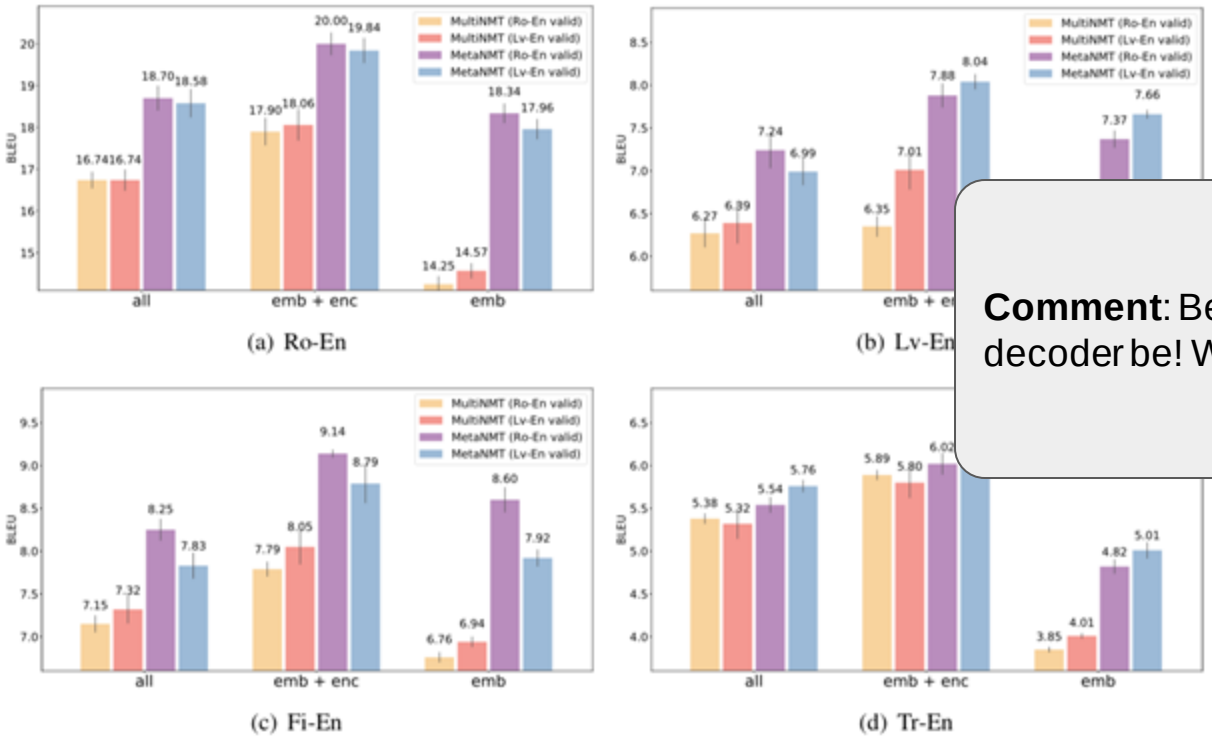


Figure 3: BLEU scores reported on test sets for {Ro, Lv, Fi, Tr} to En, where each model is first learned from 6 source tasks (Es, Fr, It, Pt, De, Ru) and then fine-tuned on randomly sampled training sets with around 16,000 English tokens per run. The error bars show the standard deviation calculated from 5 runs.

Experiments



(a) Ro-En (b) Lv-En (c) Fi-En (d) Tr-En

Figure 3: BLEU scores reported on test sets for {Ro, Lv, Fi, Tr} to En, where each model is first learned from 6 source tasks (Es, Fr, It, Pt, De, Ru) and then fine-tuned on randomly sampled training sets with around 16,000 English tokens per run. The error bars show the standard deviation calculated from 5 runs.

Comment: Best to leave the decoder be! Why?

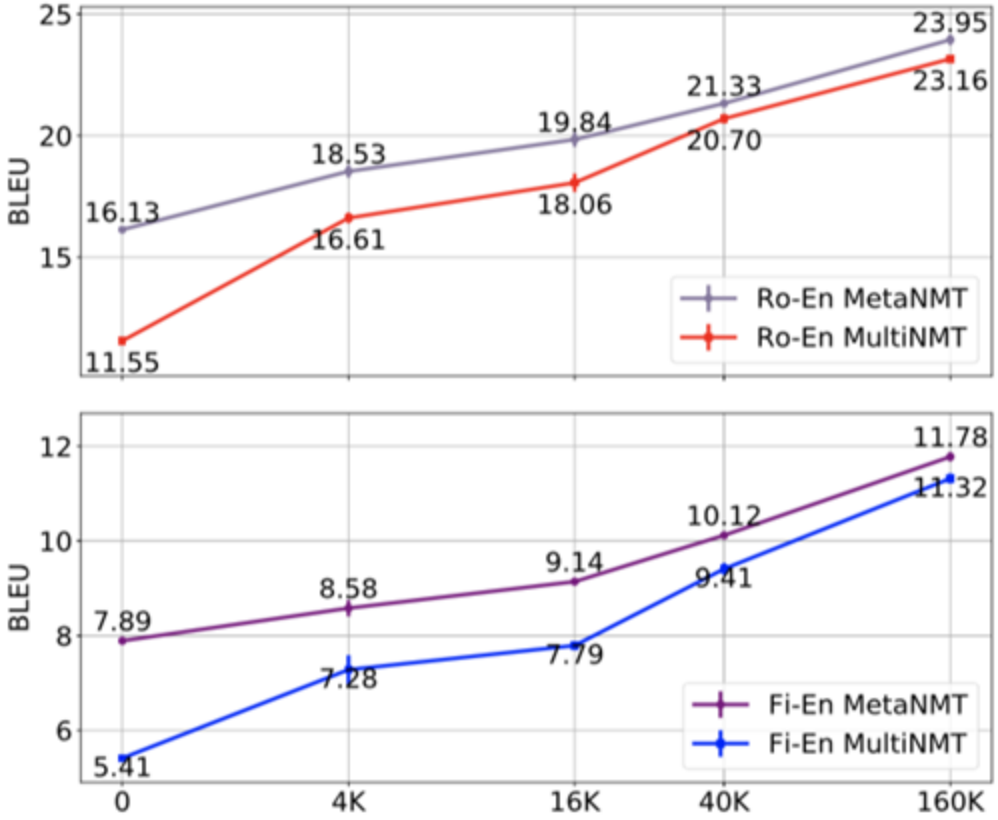


Figure 4: BLEU Scores w.r.t. the size of the target task's training set.

Comment: Gap narrows as more training examples are included

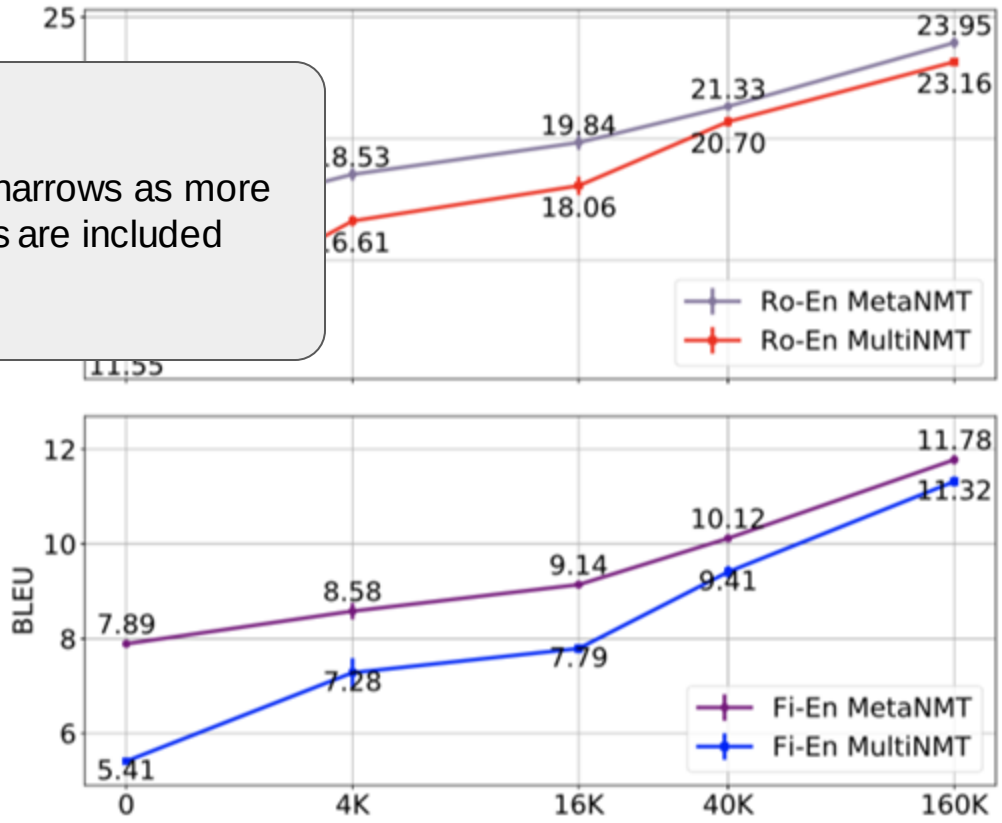


Figure 4: BLEU Scores w.r.t. the size of the target task's training set.

Critique: Don't evaluate on any real low-resource languages!

Meta-Train	Ro-En		Lv-En		Fi-En		Tr-En		Ko-En	
	zero	finetune	zero	finetune	zero	finetune	zero	finetune	zero	finetune
–		00.00 ± .00		0.00 ± .00		0.00 ± .00		0.00 ± .00		0.00 ± .00
Es	9.20	15.71 ± .22	2.23	4.65 ± .12	2.73	5.55 ± .08	1.56	4.14 ± .03	0.63	1.40 ± .09
Es Fr	12.35	17.46 ± .41	2.86	5.05 ± .04	3.71	6.08 ± .01	2.17	4.56 ± .20	0.61	1.70 ± .14
Es Fr It Pt	13.88	18.54 ± .19	3.88	5.63 ± .11	4.93	6.80 ± .04	2.49	4.82 ± .10	0.82	1.90 ± .07
De Ru	10.60	16.05 ± .31	5.15	7.19 ± .17	6.62	7.98 ± .22	3.20	6.02 ± .11	1.19	2.16 ± .09
Es Fr It Pt De Ru	15.93	20.00 ± .27	6.33	7.88 ± .14	7.89	9.14 ± .05	3.72	6.02 ± .13	1.28	2.44 ± .11
All	18.12	**22.04 ± .23**	9.58	**10.44 ± .17**	11.39	**12.63 ± .22**	5.34	**8.97 ± .08**	1.96	**3.97 ± .10**
Full Supervised	31.76		15.15		20.20		13.74		5.97	

Table 2: BLEU Scores w.r.t. the source task set for all five target tasks.

Critique: Don't know how many training examples per task? k-shot, but what is k?