Adapters for Altering LLM Vocabularies: What Languages Benefit the Most?

Contact: HyoJung Han
Ph.D. Candidate
h-j-han.github.io
hjhan@cs.umd.edu

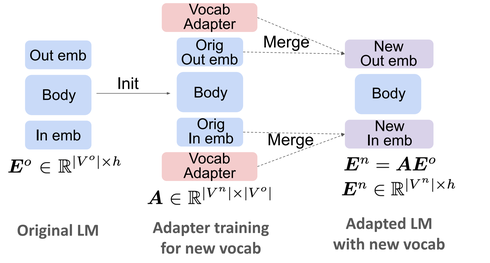
HyoJung Han, Akiko I. Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, Huda Khayrallah

Motivation

- **Vocabulary Adaptation** advantages: Introduction of new languages, Improving downstream performance in target lang, **Mitigating over-fragmentation** [1]
- Existing Works: Heuristics based initialization, Dependency on external embeddings, Language-specific approaches or restrictions
- **Understudied** Impact of Vocabulary Adaptation Across Diverse Linguistic and Task Settings: few languages, few on MT task

VocADT: Multilingual Vocabulary Adaptation with Adapters

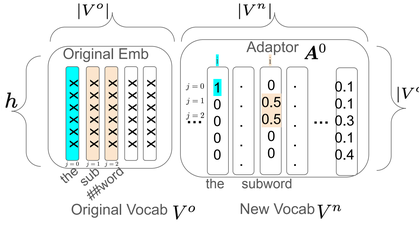
Adapter modules to learn the best combination of the original embeddings without relying on heuristics, external resources. **Adaptability** & **flexibility** with learned approach.



: updated weights
: fixed weights
V^0: original vocab
V^n: new vocab
h : embedding dim

At the end of training, the adapter is **merged** with the original embeddings to create a new embedding matrix.

Initialization Scheme for the Vocabulary Adapter



1. Copying the original embeddings of overlapping tokens

$$A^0_{i,\mathcal{I}^o(w)} = 1, \quad A_{i,j} = 0 \quad \forall j \neq \mathcal{I}^o(w)$$

2. Initializing the row of a token whose partitioned tokens by the original tokenizer are subset of the original vocabulary

$$\mathcal{T}^o(w) = \{t_1, \ldots, t_m\}^{\text{subword}}_{\text{sub}}{}^{\text{##word}} \subset V^o, m > 1$$

$$A^0_{i,j} = \begin{cases} \frac{1}{m} & \text{if } j \in \{\mathcal{I}^o_{j=1}(t_1), \ldots, \mathcal{I}^o_{j=2}(t_m)\} \\ 0 & \text{otherwise} \end{cases}$$

$i = \mathcal{I}^n(w)$ be the index of a token w in V^n
$\mathcal{T}^x : w \rightarrow (t_1, t_2, \ldots, t_k)$
a tokenizer associated with a vocabulary V^x

Experiment Design

	idx	Full Name	Short	Script	FLORES	XNLI	XCOPA	Belebele
					Evaluation			
Latin group	1	English	en	Latin	✓	✓		✓
	2	Swahili	sw	Latin	✓	✓	✓	✓
	3	Indonesian	id	Latin	✓		✓	✓
	4	Estonian	et	Latin	✓		✓	✓
	5	Haitian Creole	ht	Latin	✓		✓	✓
Mixed group	6	Korean	ko	Hangul	✓			✓
	7	Greek	el	Greek	✓	✓		✓
	8	Russian	ru	Cyrillic	✓	✓		✓
Cyrillic group	9	Bulgarian	bg	Cyrillic	✓	✓		✓
	10	Ukrainian	uk	Cyrillic	✓			✓
	11	Kazakh	kk	Cyrillic	✓			✓

Experimental Details:
$|V^0|$=32k → $|V^n|$=50k
V^n: non-Eng group + Eng
Basemodel: Mistral-7B [2]

Adapter Training:
0.5B tokens per langs
(2.5B per group)

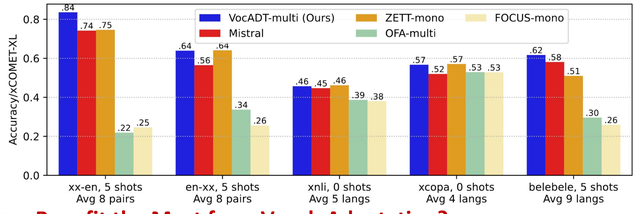
Baselines:
ZeTT[3], OFA[4], FOCUS[5]

Result of Vocabulary Adaptation (only embeddings replaced, no full-weight updates)

Overall Task Performance

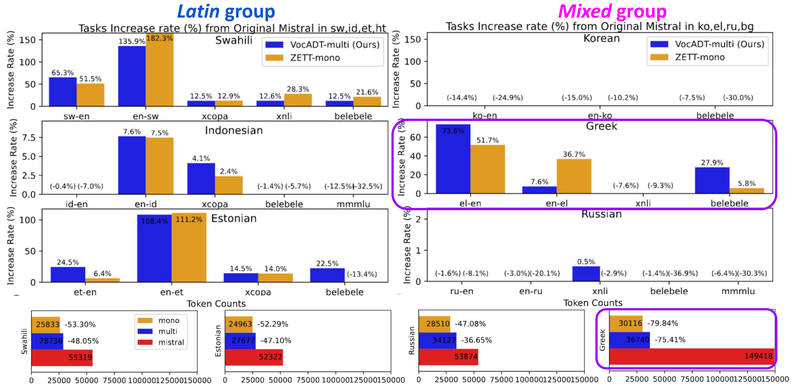
VocADT outperforms the original Mistral model, and either surpasses or performs on par with a strong baseline.

Average Performance of multilingual MT/Reasoning/QA tasks



Which Languages Benefit the Most from Vocab Adaptation?

Languages with *Latin* Scripts or Severe Fragmentation Benefit the Most.



Downstream Fine-tuning (full-weight update fine-tuning)

To understand the **impact of vocabulary adaptation after task-specific fine-tuning**, we follow the full **ALMA** [6] training on all model parameters for the cross-lingual generation task of machine translation after our VocADT on just the embeddings.

MT Fine-tuning Results : All vocabulary adaptation approaches are effective compared to Mistral except for en-sw, and among those, our approach (VocADT) achieves the highest average score in both en-xx and xx-en directions.

Table: Machine Translation performance after full-weight fine-tuning the vocabulary-adapted model.

Direction	Lang (*group*) ↓ xCOMET-XL →	VocADT	Mistral	ZeTT	OFA	FOCUS
xx-en	Avg 8 pairs (*Latin, Mixed*)	**0.899**	0.875	**0.899**	0.895	0.895
	Avg 10 pairs (*Latin, Mixed, Cyrillic*)	**0.902**	0.881	-	0.897	0.897
en-xx	Avg 8 pairs (*Latin, Mixed*)	**0.779**	0.757	0.778	0.774	0.778
	Avg 10 pairs (*Latin, Mixed, Cyrillic*)	**0.792**	0.772	-	0.785	0.786
	# of Models →	**3**	**3**	8	**3**	10

Summary

- We propose **VocADT**, a simple and effective solution for **vocabulary adaptation** using **adapters**, that addresses key limitations in prior work such as reliance on external embedding or language constraints.
- Overall, adapting the vocabulary using VocADT generally leads to better performance compared to the original Mistral model, and either surpasses or performs on par with a competitive baseline.
- We conduct experiments that cover a wide range of languages and scripts, finding that languages with **Latin scripts** or **severe fragmentation** **benefit the most** and that having a **consistent grouping** of scripts for multilingual vocabulary is helpful.
- We confirm that vocabulary adaptation remains effective even after **full-weight fine-tuning**, and VocADT is the most effective approach with a focus on machine translation

Reference

[1] Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., EMNLP 2023)
[2] https://huggingface.co/mistralai/Mistral-7B-v0.1
[3]Zero-Shot Tokenizer Transfer (Minixhofer et al., arXiv 2024)
[4]OFA: A Framework of Initializing Unseen Subword Embeddings for Efficient Large-scale Multilingual Continued Pretraining (Liu et al., Findings 2024)
[5]FOCUS: Effective Embedding Initialization for Monolingual Specialization of Multilingual Models (Dobler & de Melo, EMNLP 2023)
[6] A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models (Xu et al., ICLR 2024)