Towards a World-English Language Model for On-Device Virtual Assistants

Richa Jalota[2*], Lyan Verwimp[1], Markus Nussbaum-Thom[1], Amr Mousa[1], Arturo Argueta[1], Youssef Oualil[1]
Apple[1], AppTek GmbH[2]

Abstract

- **Neural Network Language Models** (NNLMs) for Virtual Assistants are generally language-, region- or device-dependent. Combining NNLMs for one or more categories is one way to **improve scalability**.
- This study focusses on developing a **World-English NNLM** that meets the accuracy, latency and memory constraints of single-dialect models.
  - Given **three high-resourced dialects**: American (US), British (UK), and Indian (IN) English
- Results indicate that **adapter modules** are more effective in modeling dialects than specialised sub-networks.

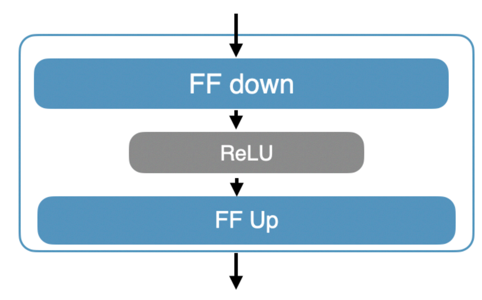
Background

FOFE-based FeedForward NNs

Fixed-size Ordinally-Forgetting Encoding (FOFE) method [1] uniquely encodes variable-length sequences into fixed-size representations, serving as an alternative to RNNs for sequence modeling tasks.

Adapters

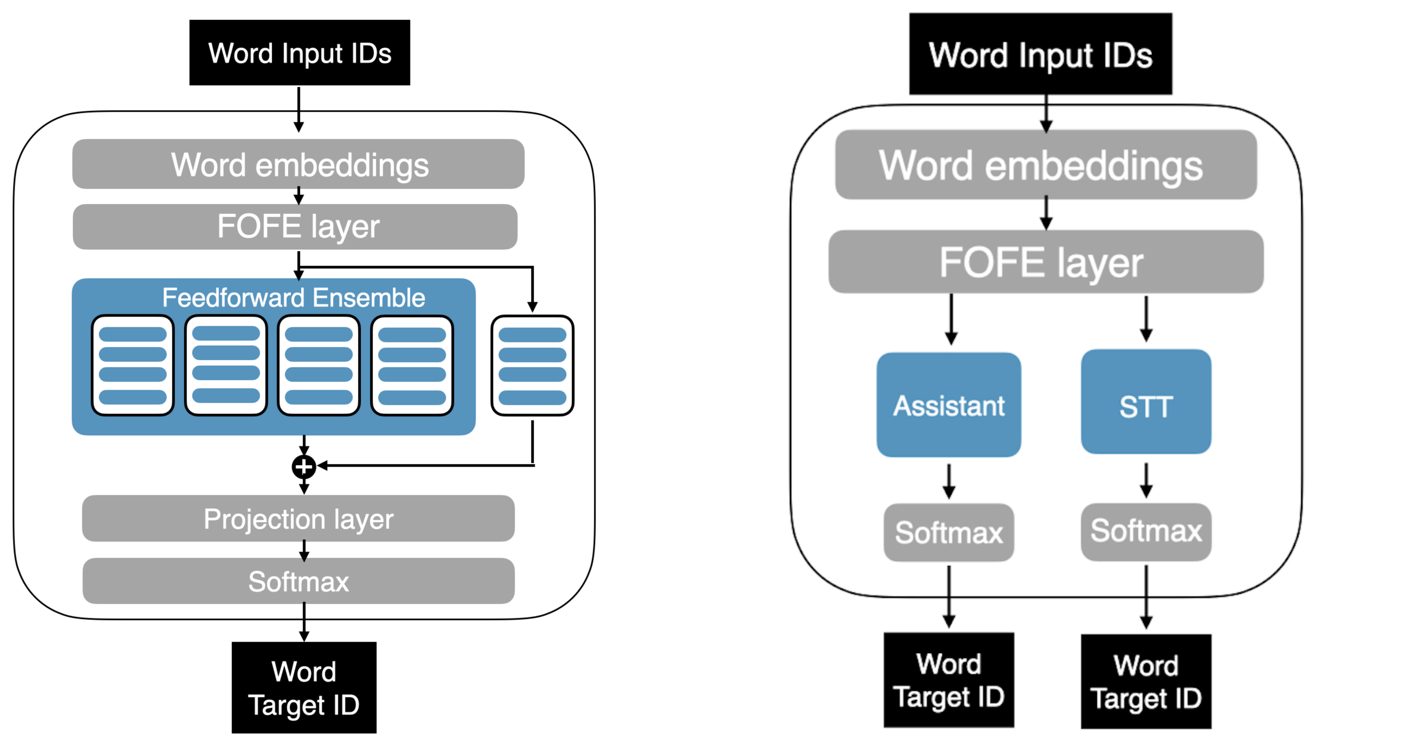
Parameter-efficient modules for adapting pre-trained models to new tasks [2]



[1]Zhang, Shiliang et al. "The Fixed-Size Ordinally-Forgetting Encoding Method for Neural Network Language Models." Annual Meeting of the Association for Computational Linguistics (2015).
[2] Adapters: Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." *International Conference on Machine Learning.* PMLR, 2019.

Model Architecture and Experimental Setup

Base Models: FOFE-based NNLMs



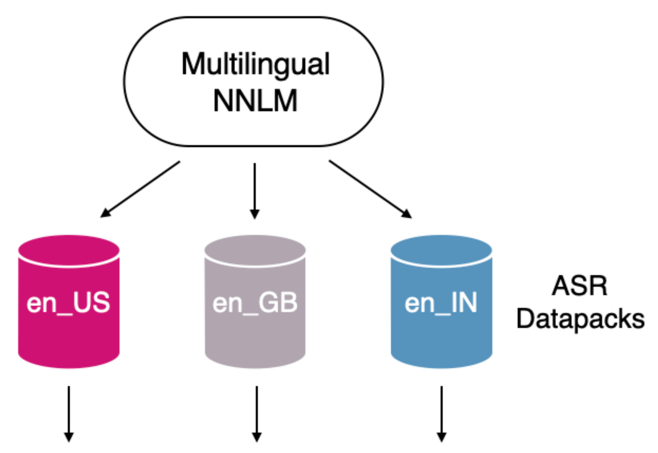
Mixture
(Higher Accuracy)

Application Dependent (AD)
(Lower Latency)

Experimental Setup

Data: Anonymised randomly sampled user requests from multiple domains and applications. Equal amounts of data sampled for each dialect for training.

Evaluation:



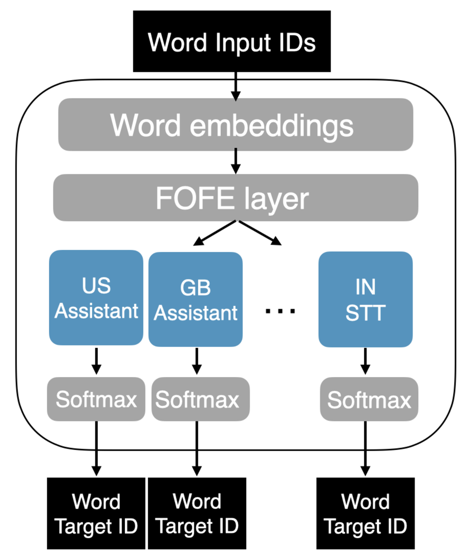
First-pass Decoding

Dialect	Ast.	STT	T.E.
US	226K	292K	454K
GB	155K	114K	232K
IN	153K	54K	239K

Number of words in test sets

World-English NNLMs

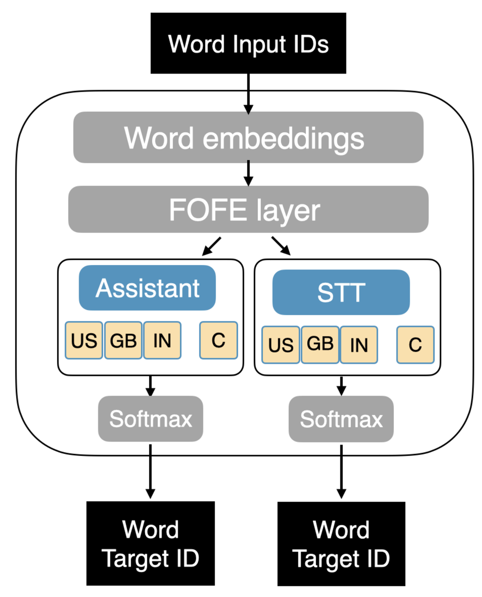
Baseline: Train Mixture FOFE and AD FOFE with multi-dialect data.



Multi-dialect AD

Extension with Adapters

1. Placement
2. Training Strategy
   i. RI-A: Add a randomly-initialised adapter to pre-trained multi-dialect model
   ii. PT-A: Train together with the base model with multi-dialect data (Mix+A)
   iii. FT-A: Fine-tune PT-A
3. Dual-Adapter (DA) Variant



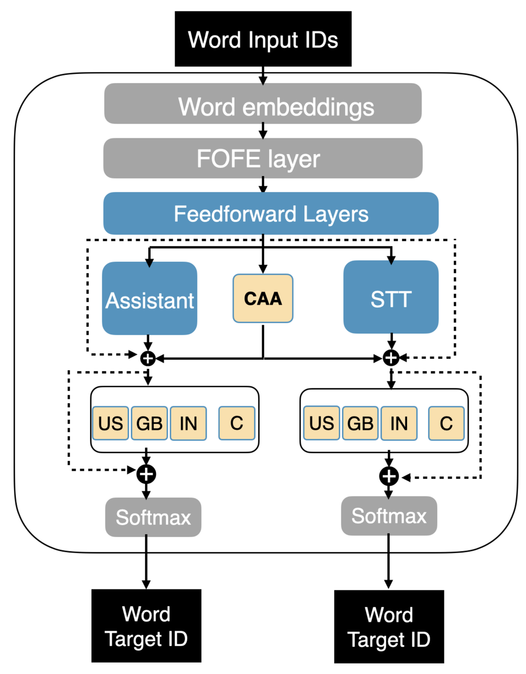
AD + DA

Proposed Architecture

Motivation:

Improve the accuracy of AD FOFE while maintaining its lower latency.

As in Mixture FOFE, add a shared representation for Applications!



AD + CAA + DA

Results

First-pass Decoding Results (WERs)

Model	Model Size	en_US Ast.	en_US STT	en_US T.E.	en_GB Ast.	en_GB STT	en_GB T.E.	en_IN Ast.	en_IN STT	en_IN T.E.
Mono	111M	3.97	3.47	18.24	**5.26**	**6.16**	16.3	6.92	9.62	26.14
Mix	89M	3.97	3.41	16.84	5.33	6.17	16.29	**6.69**	9.46	24.01
Mix+A	89M	3.95	3.41	**16.83**	5.33	6.18	**16.27**	**6.69**	**9.18**	23.99
AD	54M	4.01	3.43	17.52	5.34	6.28	16.69	7.16	9.57	24.67
AD+A	55M	3.99	3.41	21.94	5.38	6.33	21.88	7.24	9.64	**21.80**
AD+DA	45M	3.97	3.42	17.32	5.36	6.21	16.53	6.90	9.54	24.34
AD+CAA+DA	49M	**3.93**	**3.39**	17.32	5.35	6.25	16.44	6.90	9.42	24.32

Latency Results (in milliseconds)

Model	Ast. Avg.	Ast. P95	STT Avg.	STT P95
Mono_150k	334	425	50	185
Mix+A	421	785	74	230
AD+CAA+DA	359	474	54	182

Conclusions

- We build a World-English NNLM for an on-device ASR system for three high-resourced English dialects.
- After examining the application of adapters in FOFE-based models, **we introduce an architecture that bridges the accuracy and latency gap between the baseline multi-dialect models.**
- The proposed model relatively improves the accuracy of single-dialect baselines by **an average of 1.63% on head-heavy test sets** and **3.72% on tail entities across dialects**. Moreover, it matches the latency and memory constraints of on-device VAs.

*Work done while the author was an intern at Apple.