Internal Language Model Estimation based Adaptive Language Model Fusion for Domain Adaptation

Rao Ma, Xiaobo Wu, Jin Qiu, Yanan Qin, Haihua Xu, Peihao Wu, Zejun Ma

ByteDance

Abstract

- We propose a simple yet effective language model fusion approach to adapt ASR models to unseen domains.
- Extensive experiments are conducted with NNLM and N-gram LMs, under both RNN-T and LAS ASR frameworks.
- Compared to strong shallow fusion and ILME baselines, we can achieve significantly better results on the general domain while maintaining good performance on the target domain.

Introduction and Motivation

- **Text-based Adaptation**:
  - Text data from the target domain is easy to collect.
  - The ASR model remains unchanged and can be shared for different domain adaptation settings.

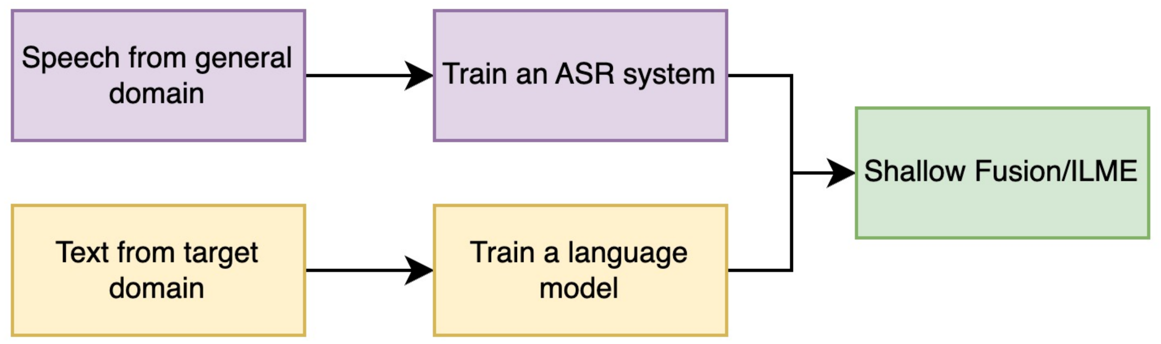

Figure 1: ASR system adaptation with text from the target domain.

- **Language Model Fusion**:
  - E2E ASR Model:

$$\hat{W} = \arg\max_{W} \log p_\theta(W|X)$$

  - Shallow Fusion:

$$\hat{W} = \arg\max_{W} [\log p_\theta(X|W) + \lambda \log p_{\text{LM}}(W)]$$

  - Internal Language Model Estimation (ILME) [1]:

$$\hat{W} = \arg\max_{W} [\log p_\theta(W|X) - \lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) + \lambda \log p_{\text{LM}}(W)]$$

- **Motivation**:
  - In real-life scenarios, the incoming speech can be switched across different domains during a session.
  - With SF/ILME, the ASR model performs well on the target domain while degrading performance on the general domain.
  - We hope to achieve target domain adaptation without sacrificing the model performance on the general domain.

ILME-based Adaptive Domain Adaptation

$$\hat{W} = \arg\max_{W} [\log p_\theta(W|X) - \lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) + \max(\lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W), \lambda \log p_{\text{LM}}(W))]$$

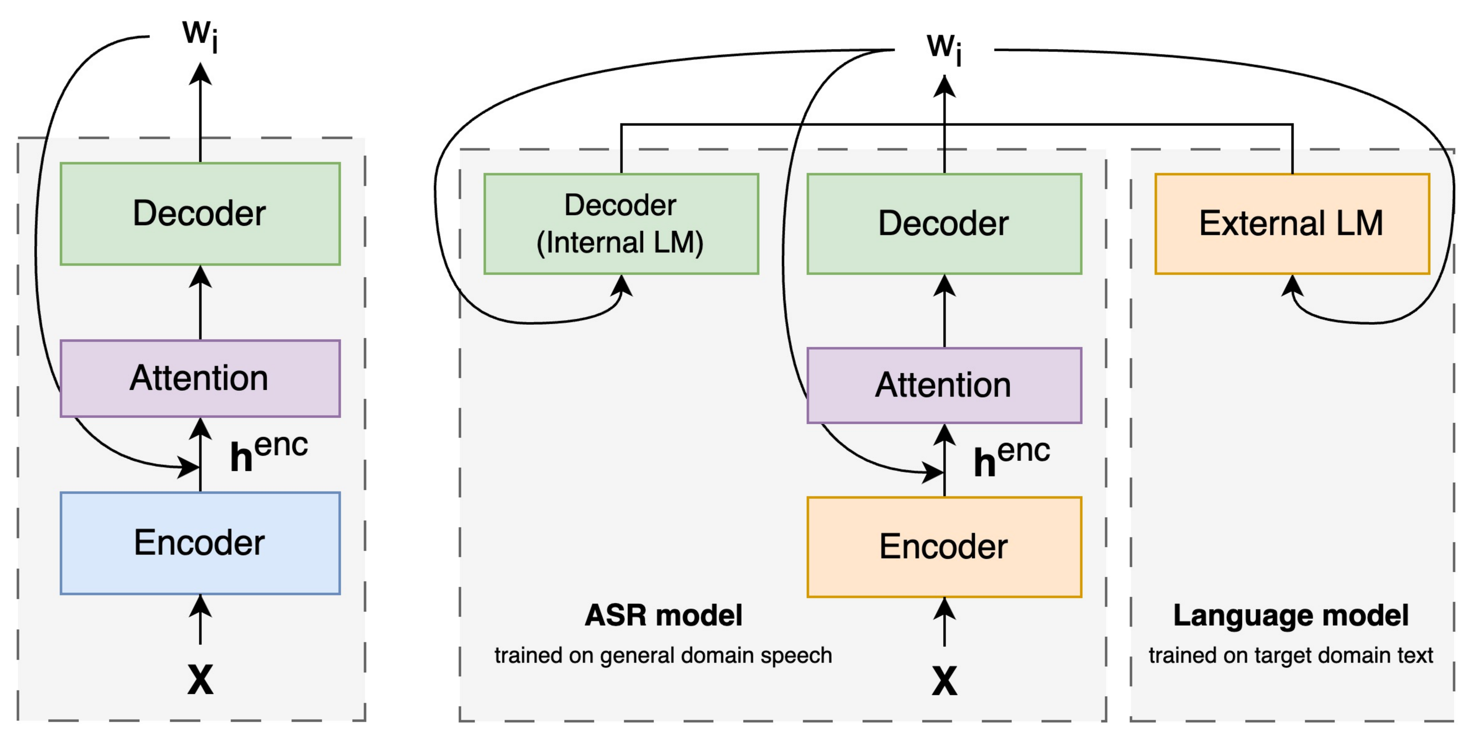

Figure 2: For the proposed ILME-ADA, the ASR model uses different calculations depending on the comparison results of the external LM score and internal LM (ILM) score at each timestep.

- Different fusion methods are chosen automatically in decoding.
- When $\lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) >= \lambda \log p_{\text{LM}}(W)$, only the ASR score is added to the decoding hypothesis.
- When $\lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) < \lambda \log p_{\text{LM}}(W)$, we assume the speech is from the target domain.

Experiment Setup

- For the baseline ASR systems, we train LAS and RNN-T models on over 100k hours of speech data covering a wide range.

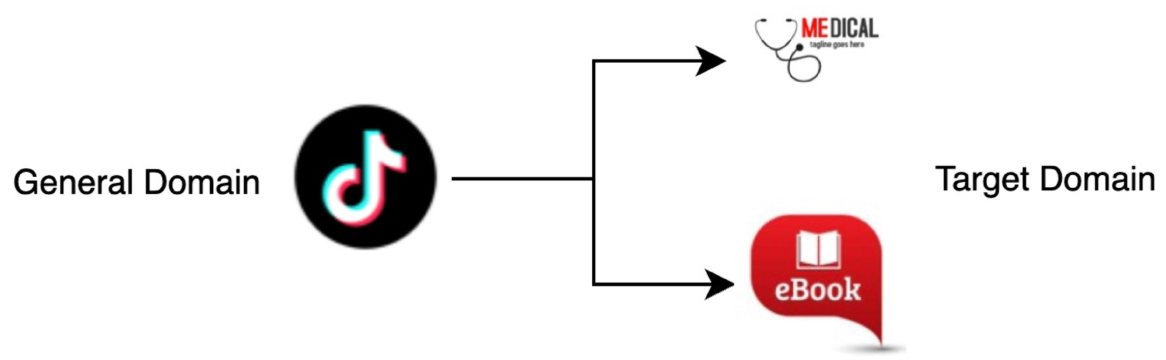

Figure 3: ASR system adaptation with text from the target domain.

Target Domain	Train	Test	
	characters (M)	character (K)	duration (h)
Search	412	33.03	3.25
Medical	519	287.05	22.83

Table 1: Statistics for datasets utilised in the domain adaptation experiments.

Perplexity (PPL) Results

Model		PPL (Search)		PPL (Medical)	
		Target	General	Target	General
LM	NNLM	12.15	169.05	16.58	100.32
	n-gram LM	34.47	239.86	24.61	148.13
ILM	RNN-T	460.76	132.86	186.31	132.86

Table 2: Perplexity results on the general and target domain test sets calculated by the external LMs and ILMs.

- External language model yields lower PPL on the target domain while ILM shows lower PPL on the general domain.

Comparison of Different Adaptation Methods

Model		Target: Search		Target: Medical	
		Target	General	Target	General
Baseline (no fusion)		21.88	13.89	4.47	13.89
NNLM	SF	14.89	28.55	3.43	14.56
	ILME	10.47	20.36	3.53	14.56
	ILME-ADA	13.35	14.81	3.53	14.25
N-gram	SF	15.08	17.49	3.62	15.56
	ILME	12.69	19.97	3.47	15.82
	ILME-ADA	11.72	15.35	3.44	14.03

Table 3: CERs (%) of adapted RNN-T models on the eBook search and medical domain test sets with the proposed ILME-ADA method. For each domain, an external NNLM and n-gram LM are trained with **ONLY** target domain text data.

- ILME-ADA largely improves ASR performance on the target domain while minimally influencing general domain performance.

Analysis

- ***Cond.A*** refers to $\lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) < \lambda \log p_{\text{LM}}(W)$.
- ***Cond.B*** refers to $\lambda^{\text{ILM}} \log p_\theta^{\text{ILM}}(W) >= \lambda \log p_{\text{LM}}(W)$.

Dataset	Target: Medical		Target: Search	
	Cond.A	Cond.B	Cond.A	Cond.B
Target	81.5%	18.5%	88.4%	11.6%
General	50.5%	49.5%	33.1%	66.9%

Table 4: Percentage of tokens satisfying different conditions in ILME-ADA decoding results on RNN-T with NNLM fusion.

References

[1] Meng, Zhong, et al. "Internal language model estimation for domain-adaptive end-to-end speech recognition", *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

E-Mail: rm2114@cam.ac.uk {xiaobo.wu, qiujin, haihua.xu, wupeihao, mazejun}@bytedance.com yanan.qin@xjsd.com