Experiment---Adapt domain

Target: Adapt the system to out-of-domain test data.

Specific:

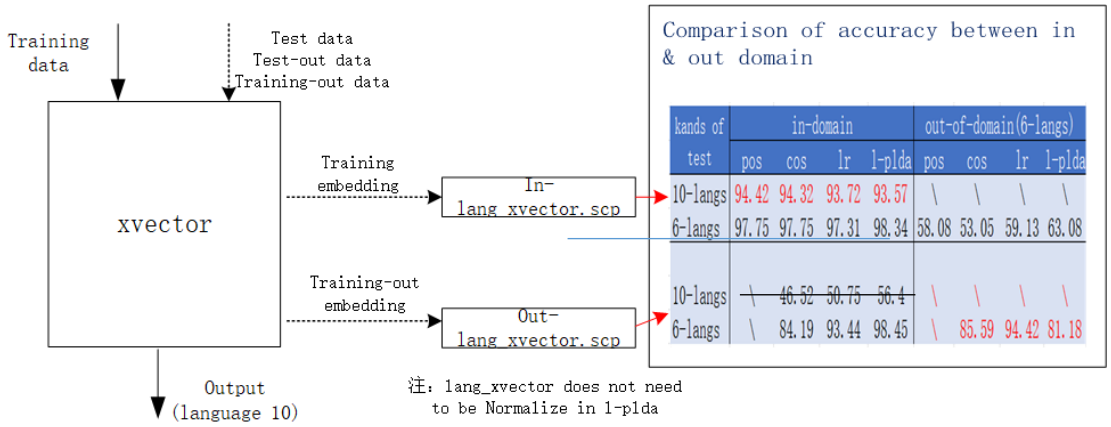
it use training data to train xvector system and training-out data to extract the language.vector about out-of-domain.(training-out data is taken from out-of-domain test data.)

training data:

Training_data = train_25h

Training-out_data dur = 1h (939 utts)

Structure diagram:



Result:

Training_data = train_25h (10 languages)

Training-out_data dur = 1h(939utts) and different lengths of training-out data 0.2h(188 utts).

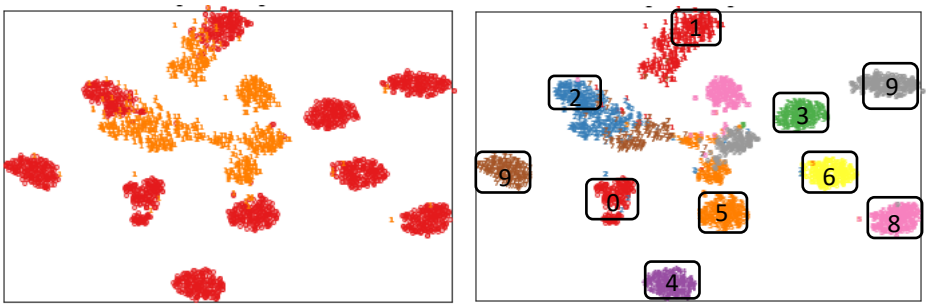
N.utts	kinds of test	in-domain				out-of-domain(6-langs)			
		pos	cos	lr	l-plda	pos	cos	lr	l-plda
original	10-langs	94.42	94.32	93.72	93.57	\	\	\	\
	6-langs	97.75	97.75	97.31	98.34	58.08	53.05	59.13	63.08
939 utts	10-langs	\	\	\	\	\	\	\	\
	6-langs	\	84.19	93.44	98.00	\	85.59	94.42	81.18
188 utts	10-langs	\	\	\	\	\	\	\	\
	6-langs	\	88.59	94.33	98.12	\	88.89	92.61	79.97

Conclusion:

1. Original : the accuracy of the 6 languages is better than 10. Because the other 4 are easy to mix with others
2. Adapt: training-out data has only 6 languages, and only 6 languages containing training-out data are tested. The above results show that using the adapt method, the accuarcy will increase significantly in out-of-domain test, and the accuracy will drop slightly in in-domain.

Analysis:

It utilize t-sne to analyze the distribution of in-domain and out-of-domain data.



(a) (b)

(a) in-domain & out-domain, It can be seen that the data distribution of out-of-domain is more compact.

(b) Specific distribution, where the frame represents the in-domain training.

0-Kazak, 1-Tibet, 2-Uyghu, 3-ct-cn, 4-id-id, 5-ja-jp, 6-ko-kr, 7-ru-ru, 8-vi-vn, 9-zh-cn

It can be seen that if there are no four languages, 0, 3, 4, and 6, the out-of-domain data is equivalent to the convergence of the in-domain data to the center. I did an experiment with only 6 training data, and the results were indeed better than the 10 languages.

Experiment---Compare the accuracy different types of training datas (no adapt)

Training data	kinds of test	in-domain				out-of-domain(6-langs)			
		pos	cos	lr	l-plda	pos	cos	lr	l-plda
Original	10-langs	94.42	94.32	93.72	93.57	\	\	\	\
	6-langs	97.75	97.75	97.31	98.34	58.08	53.05	59.13	63.08
6-langs	6-langs	\	99.04	99.05	98.89	\	65.69	68.13	68.79

Comparative Experiment:

The new training data consists of the original training data and a small amount of training-out data that is taken from out-of-domain test data. The experimental results are as follows:

Training data	kinds of test	in-domain				out-of-domain(6-langs)			
		pos	cos	lr	l-plda	pos	cos	lr	l-plda
Original	10-langs	94.42	94.32	93.72	93.57	\	\	\	\
	6-langs	97.75	97.75	97.31	98.34	58.08	53.05	59.13	63.08
Ori+939utts	10-langs	93.64	94.30	93.50	86.48	\	\	\	\
	6-langs	97.58	97.77	98.31	98.28	93.22	81.37	93.50	83.32
Ori+188utts	10-langs	93.19	93.27	92.93	85.87	\	\	\	\
	6-langs	97.68	97.78	98.90	98.54	78.72	63.34	79.85	79.17

Conclusion:

1. Adapt method is better than combine(original, training-out) as new training data when training-out is low-resource. And it doen't require retraining the xvector system, when identifying data for different channels.