LEARNING NEURAL TRANS-DIMENSIONAL RANDOM FIELD LANGUAGE MODELS WITH NOISE-CONTRASTIVE ESTIMATION

Bin Wang, Zhijian Ou
Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, Beijing, China.
wangbin12@mails.Tsinghua.edu.cn, ozj@Tsinghua.edu.cn

清華大學
Tsinghua University
Department of Electronic Engineering

Introduction

Trans-dimensional random field (TRF) LMs

- ◆ To fit the joint probability $p(x_1, \dots, x_l)$ directly
- ◆ Support both discrete features and neural network features
- ◆ **Inference is fast** but **training is slow**

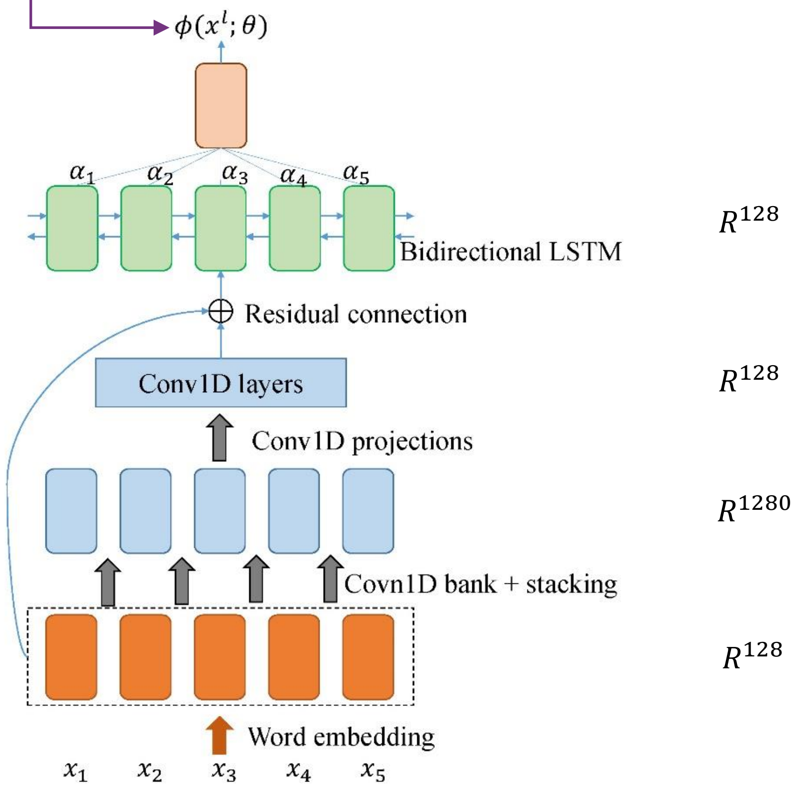
To improve the **training efficiency** and the **performance** of neural TRF LMs:

- ✓ Define the TRF in the form of exponential tilting of a reference distribution
- ✓ Introduce the noise-contrastive estimation (NCE) to train TRF LM.
- ✓ Marry the deep CNN and the bi-directional LSTM

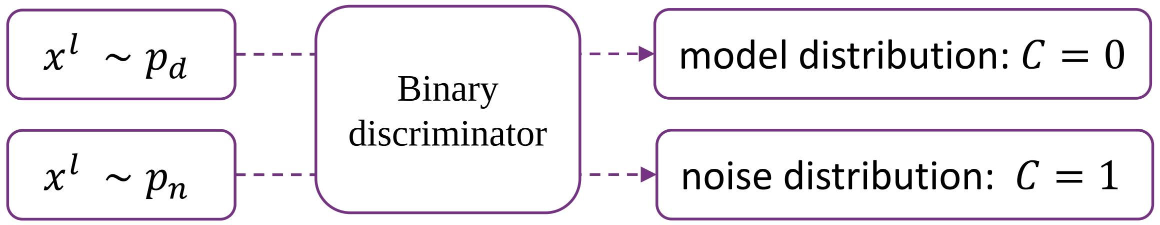
Model Definition

$$p_m(x^l;\theta,\zeta) = \pi_l q(x^l) e^{\phi(x^l;\theta)-\zeta_l}$$

$x^l = (x_1, \dots, x_l)$	a word sequence of length l
π_l	the prior length probability
$q(x^l)$	a LSTM language model
ζ_l	the normalization constant of length l need to be estimated
$\phi(x^l;\theta)$	potential function with parameter θ



Noise-contrastive Estimation (NCE)



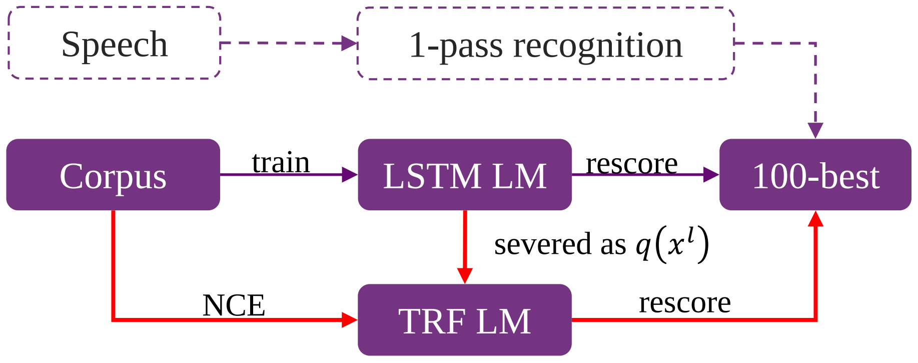
$$P(C=0|x^l) = \frac{p_m}{p_m + \nu p_n}$$

$$P(C=1|x^l) = 1 - p(C=0|x^l)$$

$$\max_{\theta,\zeta} \frac{1}{|D|}\sum_{x^l\in D} \log P(C=0|x^l) + \frac{\nu}{|B|}\sum_{x^l \in B} \log P(C=1|x^l)$$

Experiments

Speech recognition WERs on CHiME-4 Challenge data.



model	Dev		Test	
	real	simu	real	simu
KN5	5.03	4.79	7.38	5.78
LSTM (i.e. $q(x^l)$)	3.63	3.24	5.70	4.53
TRF	3.53	3.20	5.68	4.36
KN5+LSTM	3.56	3.29	**5.71**	4.18
KN5+TRF	3.53	3.22	5.54	4.20
KN5+LSTM+TRF	3.42	3.10	**5.44**	4.13

- ◆ KN5: 5gram LM with modified Kneser-Ney smoothing
- ◆ LSTM: 2 hidden layers, 512 hidden units per layer
- ◆ "Dev" denotes the development set and "Test" denotes the test set.
- ◆ "+" denotes the log-linear interpolation with equal weights

Conclusion:

- ✓ On a 40x larger training set use only **1/3** training time
- ✓ Achieve a **4.7%** relative WER reduction on the top of a strong LSTM LM baseline.