Improved Training of Neural Trans-dimensional Random field Language Models with Dynamic 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

Open-source toolkit: https://github.com/wbengine/TRF-NN-Tensorflow

Introduction

Trans-dimensional random field (TRF) LMs

Whole-sentence modeling: directly fit the joint probability $p(x_1, \dots, x_l)$;

☺ Avoid local normalization;

☺ Flexible: no acyclic and local normalization constraint.

Propose the dynamic noise-contrastive estimation (DNCE) to solve the two problems of NCE:

1 Cut down the noise sample number (20 -> 4);

2 Alleviate the overfitting problem.

Humans employ context for reading and writing.

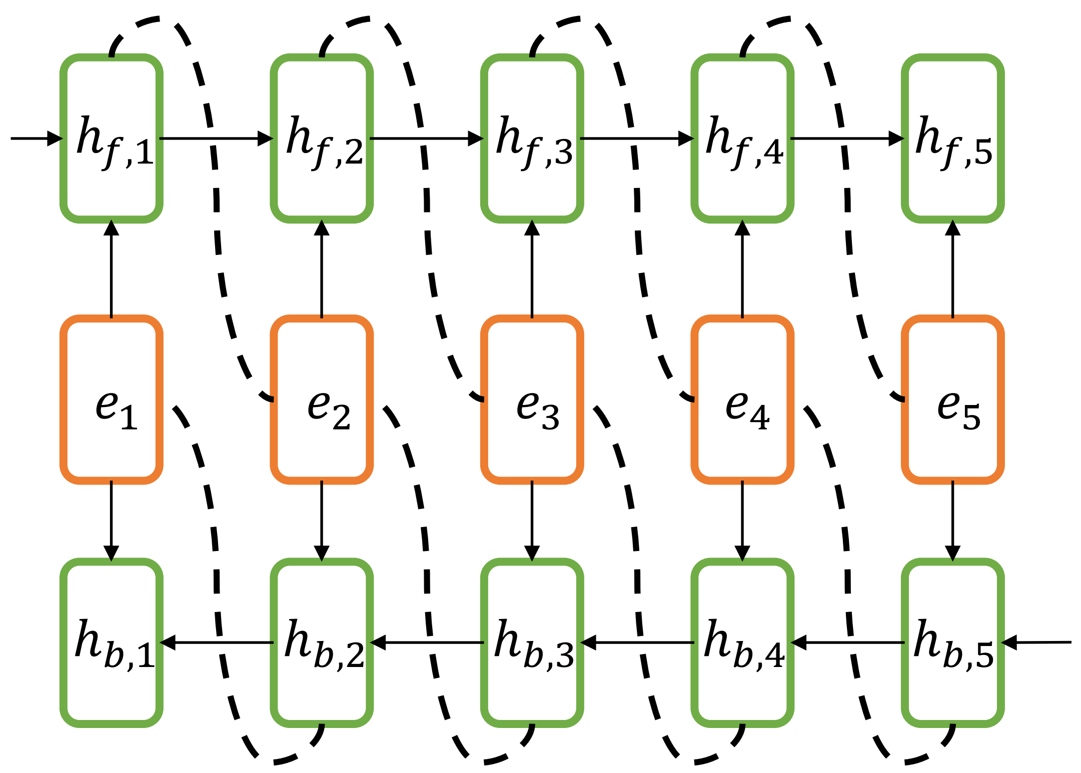
The cat is on the table.

The cat is in the house.

Model Definition

$$p_m(s;\theta,Z) = \pi_l e^{\phi(s;\theta) - \log Z_l}$$

$s = (x_1, \dots, x_l)$	A word sequence of length l
π_l	The prior length probability
Z_l	The normalization constant of length l (to be estimated)
$\phi(s;\theta)$	Potential function with parameter θ $\phi(s;\theta) = \sum_{i=1}^{l-1} h_{f,i}^T e_{i+1} + \sum_{i=2}^{l} h_{b,i}^T e_{i-1}$



Model Training

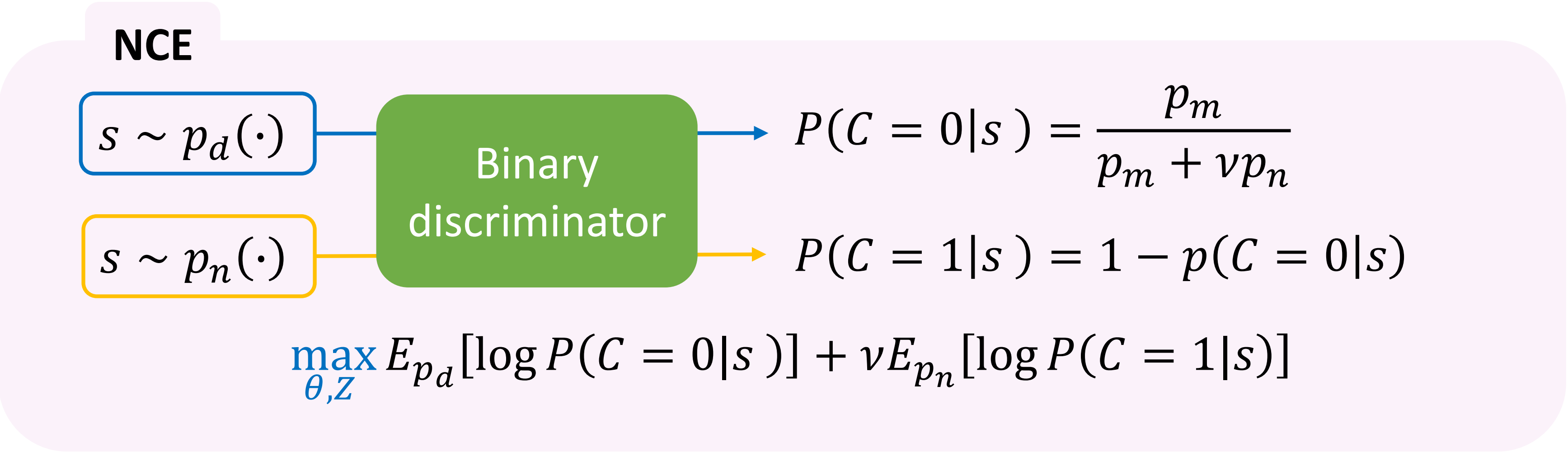
$$\nabla_\theta \mathrm{LogLikelihood} = E_{p_d(s)}[\nabla_\theta \phi(s;\theta)] - E_{p_m(s;\theta)}[\nabla_\theta \phi(s;\theta)]$$

Expectation under empirical distribution

Expectation under model distribution

AugSA (ACL 2015, TPAMI 2018), AugSA plus JSA (ASRU 2017), NCE (ICASSP 2018)

Dynamic Noise-contrastive Estimation (DNCE)



NCE

$s \sim p_d(\cdot)$

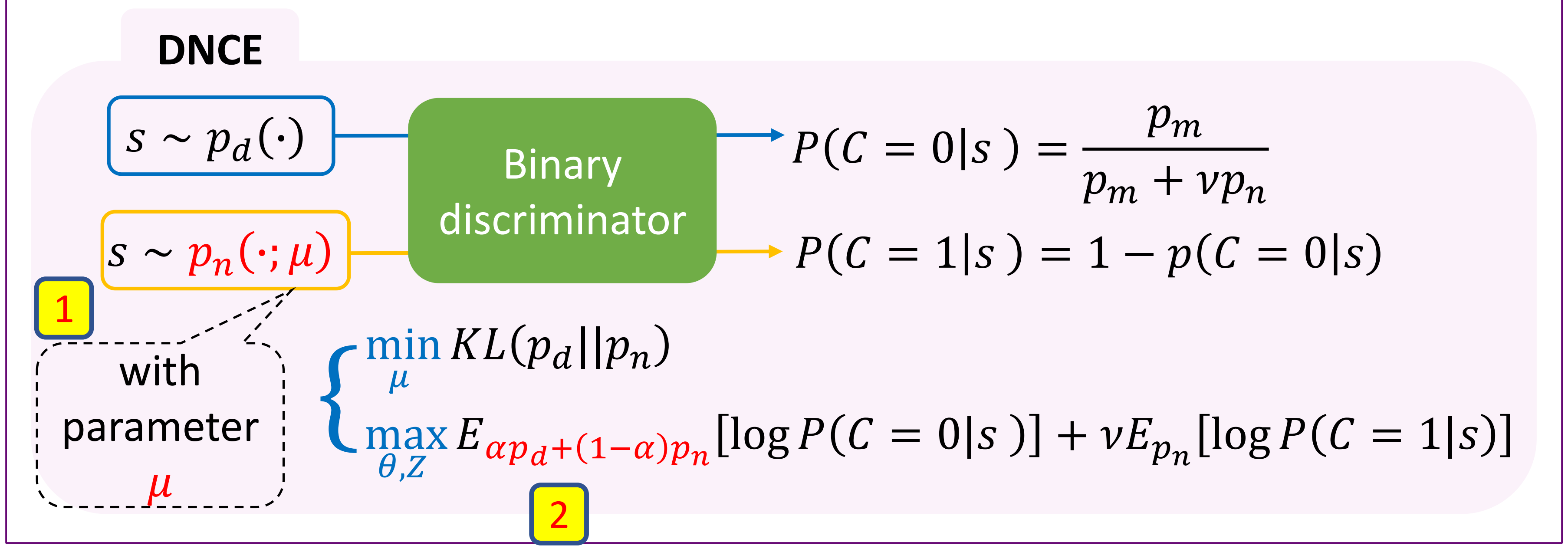
$s \sim p_n(\cdot)$

Binary discriminator

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

$$P(C=1|s) = 1 - p(C=0|s)$$

$$\max_{\theta,Z} E_{p_d}[\log P(C=0|s)] + \nu E_{p_n}[\log P(C=1|s)]$$



DNCE

$s \sim p_d(\cdot)$

$s \sim p_n(\cdot;\mu)$

Binary discriminator

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

$$P(C=1|s) = 1 - p(C=0|s)$$

1 with parameter μ

$$\begin{cases} \min_{\mu} KL(p_d || p_n) \\ \max_{\theta,Z} E_{\alpha p_d + (1-\alpha) p_n}[\log P(C=0|s)] + \nu E_{p_n}[\log P(C=1|s)] \end{cases}$$

2

Experiments

Models	PTB			HKUST			Google one-billion			Device
	WER	#Param (M)	Infer. (s)	WER	#Param (M)	Infer. (s)	WER	#Param (M)	Infer. (s)	
KN5	8.78	2.3	0.06	28.48	3.5	0.004	6.13	133	0.49	CPU
LSTM	7.36	66.0	9.09	27.60	2.2	0.048	5.55	191	0.91	GPU
TRF	7.40	2.6	0.08	27.72	1.4	0.009	5.47	114	0.02	GPU
TRF+KN5+LSTM	-		-	**26.87**		-	**5.06**		-	GPU

TRFs perform as good as LSTMs with less parameters and being 5x ~ 114x faster in inference.