Language modeling with neural trans-dimensional random fields

Bin Wang, Zhijian Ou

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

Introduction

Directed graphical language models:

- p(x_t|x_{t-1}, ..., x_{t-i})

Directed graphical language models:

- p(x_t|x_{t-1}, ..., x_{t-i})

- Being able to flexibly integrate rich features – discrete features and neural network features

- Computationally more efficient in inference than LSTM LMs.

Training objectives

p(l, x_t; θ, ζ) an TRF LM with parameters θ, ζ
q(l, x_t; μ) an auxiliary LM with parameter μ

1. For θ. Maximize the likelihood.

\[ E_D \frac{\partial \phi}{\partial \theta} = E_{p(l, x_t; \theta, \zeta)} \frac{\partial \phi}{\partial \theta} = 0 \]

The expectation on the training set D

2. For ζ. Optimize the length distribution

\[ \sum_l p(l, x_t; \theta, \zeta) = \pi_t \]

The marginal length probability

3. For μ. Minimize the KL divergence between p and q

\[ \frac{\partial}{\partial \mu} KL(p(l, x_t; \theta, \zeta) \| q(l, x_t; \mu)) = 0 \]

Three objectives induce the following update operations

\[ \theta(t) = \theta(t-1) + \gamma_{\theta, \text{Adam}} \left\{ E_D \frac{\partial \phi}{\partial \theta} - \frac{1}{K_B} \sum_{(l,x_t) \in B(t)} \frac{\partial \phi}{\partial \theta} \right\} \]

\[ \zeta(t) = \zeta(t-1) + \gamma_{\zeta, 1} \frac{1}{\pi_t} K_B \sum_{(j,x_t) \in B(t)} 1(j == l) \quad \text{for } l = 1, ..., m \]

\[ \zeta(t) = \zeta(t-1) - \zeta(t) \]

\[ \mu(t) = \mu(t-1) + \gamma_{\mu, \text{Adam}} \sum_{(l,x_t) \in B(t)} \frac{\partial}{\partial \mu} \log q(l, x_t; \mu) \]

where:

- D(t) is the mini-batch of training data at iteration t
- B(t) is the sample set at iteration t, and \( K_B = |B(t)| \)
- \( \gamma_{\theta, \text{Adam}}, \gamma_{\zeta}, \gamma_{\mu, \text{Adam}} \) are the learning rates for θ, ζ, μ respectively.

- Adam is the Adam method

Model Definition

- p(l, x_t; θ, ζ)
- x^t is the a word sequence of length l, ranging from 1 to m
- Variables need to be estimated:
  - \( \theta \): the model parameters.
  - ζ = (ζ_1, ..., ζ_m): normalization constants.

\[ p(l, x_t; \theta, \zeta) = \pi_l \cdot \frac{1}{Z_0} \exp \left( \phi(x_t; \theta) - \zeta_l \right) \]

The joint probability of sequence x^t and l

The empirical length probability

Model Training

JSA:

AugSA:

- p(l, x_t; θ, ζ)
- Our target model - A TRF with parameters θ and ζ

- TransMS
- Trans-dimensional mixture sampling with an auxiliary distribution

Collection

Get a set of sequences satisfying p(l, x_t; θ, ζ)

Model Evaluation

LMs trained on Penn Treebank (PTB) training set are applied to rescore the 1000-best lists from recognizing WSJ 92 test data (330 utterances).

Trans-dimensional mixture sampling

Step I: local jump

\[ (l^{(t)}, x_t^{(t)}) \]

propose a new length j

\[ (l^{(t)}, x_t^{(t)}) \]

or

\[ j < l^{(t)} \]

MH accept or reject

Step II: Markov move

\[ (l^{(t)}, x_t^{(t)}) \]

propose words

\[ q(l, x_t; \mu) \]

MH accept

\[ (l^{(t)}, x_t^{(t)}) \]

or

\[ j > l^{(t)} \]

MH accept or reject

\[ (l^{(t)}, x_t^{(t)}) \]

or

\[ j > l^{(t)} \]

MH reject

Fig. 1. The deep CNN architecture used to define the potential function \( \phi(x_t; \theta) \). Shadow areas denote the padded zeros.

Fig. 3. The KL-divergence KL(p||q)

Fig. 4. The negative log-likelihood on PTB test set