Large-scale Reordering Model for SMT using Dual Multinomial Logistic Regression

Abdullah Alrajeh[a,b] and Mahesan Niranjan[b]

University of Southampton

emnlp2014

مدينة الملك عبدالعزيز للعلوم والتقنية KACST

Abstract

Phrase reordering is a challenge for statistical machine translation (SMT) systems. Posing phrase movements as a prediction problem using contextual features modeled by maximum entropy-based classifier is superior to the commonly used lexicalized reordering model. However, training this discriminative model using large-scale parallel corpus might be computationally expensive. In this paper, we explore recent advancements in solving large-scale classification problems. Using the dual problem to multinomial logistic regression, we managed to shrink the training data while iterating and produce significant saving in computation and memory while preserving the accuracy.

Introduction

Foreign sentence (f) $f_1\ f_2$ (1) $f_3\ f_4\ f_5$ (2) f_6 (3).

English sentence (e) e_1 (1) $e_2\ e_3$ (3) $e_4\ e_5$ (2).

The best translation according to a combination of different models is:

$$\mathbf{e}_{\text{best}} = \arg\max_{\mathbf{e}} \sum_{i=1}^{n} \lambda_i h_i(\mathbf{f}, \mathbf{e})$$

Reordering model is important for phrase-based systems defined as:

$$h_{reo}(\mathbf{f}, \mathbf{e}) = \log p(\mathbf{o}|\mathbf{f}, \mathbf{e}) \approx \sum_i \log p(o_i|\bar{f}_i, \bar{e}_i)$$

Extracted phrase pairs $(\mathbf{f}, \bar{\mathbf{e}})$:

$\bar{f}_i$	$\bar{e}_i$	o_i	alignment	context
$f_1\ f_2$	e_1	mono	0-0 1-0	$+f_3$
$f_3\ f_4\ f_5$	$e_4\ e_5$	swap	0-1 2-0	$-f_2\ +f_6$
f_6	$e_2\ e_3$	other	0-0 0-1	$-f_5$

Bag-of-words representation for feature-based models:

features: $f_1\&e_1$ $f_2\&e_1$ $+f_3$ $f_3\&e_5$ $f_5\&e_4$ $-f_2$ $+f_6$ $f_6\&e_2$ $f_6\&e_3$ $-f_5$

$\phi(\bar{f}_1, \bar{e}_1) = 1110000000$ $\phi(\bar{f}_2, \bar{e}_2) = 0001111000$ $\phi(\bar{f}_3, \bar{e}_3) = 000000111$

Multinomial Logistic Regression (MLR)

MLR (also known as MaxEnt):

$$p(o_k|\bar{f}_i, \bar{e}_i) = \frac{\exp(\mathbf{w}_k^\top \phi(\bar{f}_i, \bar{e}_i))}{\sum_{k'} \exp(\mathbf{w}_{k'}^\top \phi(\bar{f}_i, \bar{e}_i))}$$

The primal problem:

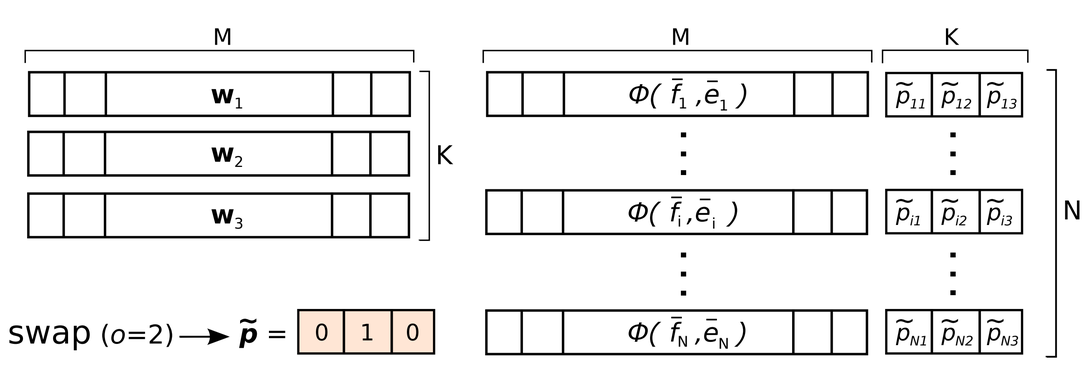
$$\min_{\mathbf{w}} \ \mathcal{P}(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 - \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{p}_{ik} \log p(o_k|\bar{f}_i, \bar{e}_i)$$

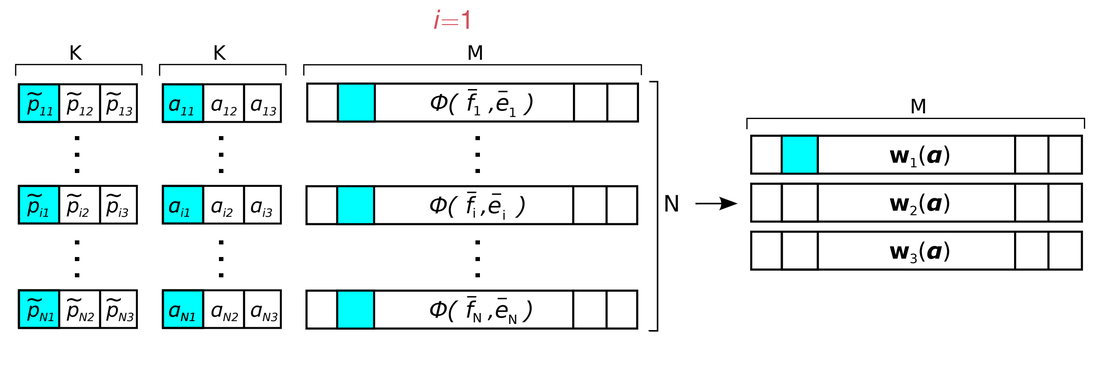
The dual problem:

$$\min_{\mathbf{w}} \quad \mathcal{D}(\boldsymbol{\alpha}) = \frac{1}{2\sigma^2} \sum_{k=1}^{K} \|\mathbf{w}_k(\boldsymbol{\alpha})\|^2 + \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_{ik} \log \alpha_{ik}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \alpha_{ik} = 1 \ \text{ and } \ \alpha_{\text{ik}} \geq 0\,, \forall \text{i}, \text{k}$$

$$\mathbf{w}_k(\boldsymbol{\alpha}) = \sigma^2 \sum_{i=1}^{N} (\tilde{p}_{ik} - \alpha_{ik}) \phi(\bar{f}_i, \bar{e}_i)$$



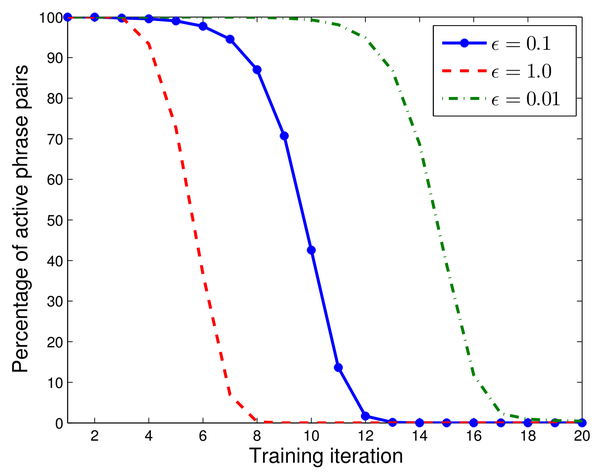


Shrinking Algorithm

Require: training set $S = \{\phi(\bar{f}_i, \bar{e}_i), o_i\}_{i=1}^{N}$

Given α and the corresponding $\mathbf{w}(\alpha)$
repeat
  Randomly pick i from S
  Claculate $\nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) = 1 + \log \alpha_{ik} + \mathbf{w}_y(\boldsymbol{\alpha})^\top \phi(\bar{f}_i, \bar{e}_i) - \mathbf{w}_k(\boldsymbol{\alpha})^\top \phi(\bar{f}_i, \bar{e}_i)$
  $v_i = \max_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha}) - \min_k \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha})$
  if $v_i \leq \epsilon$ **then**
    Remove i from S
    Continue from line 3
  end if
  $\eta = 0.5$
  for $t = 1$ **to** maxTrial **do**
    Calculate $\alpha'_{ik} = \alpha_{ik} \exp(-\eta \nabla_{ik}\mathcal{D}(\boldsymbol{\alpha})) / \sum_{k'} \alpha_{ik'} \exp(-\eta \nabla_{ik'}\mathcal{D}(\boldsymbol{\alpha}))$
    if $\mathcal{D}(\boldsymbol{\alpha}') - \mathcal{D}(\boldsymbol{\alpha}) \leq 0$ **then**
      Update $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ and $\mathbf{w}(\boldsymbol{\alpha}) = \mathbf{w}_k(\boldsymbol{\alpha}) - \sigma^2(\alpha'_{ik} - \alpha_{ik})\phi(\bar{f}_i, \bar{e}_i)$
      Break
    end if
    $\eta = 0.5\,\eta$
  end for
until $v_i \leq \epsilon \quad \forall i$

Classification & Translation Results (MultiUN Corpus)



Error rate based on held-out data

Classifier	Training Time	Error Rate
Primal MLR	1 hour 9 mins	17.81%
Dual MLR ϵ:0.1	18 minutes	17.95%
Dual MLR ϵ:1.0	13 minutes	21.13%
Dual MLR ϵ:0.01	22 minutes	17.89%

BLEU scores based on NIST sets

Ar-En Translation System	MT06	MT08
Baseline + Lexical. model	30.86	34.22
Baseline + Primal MLR	31.37	34.85
Baseline + Dual MLR ϵ:0.1	31.36	34.87

Conclusion

- MLR reordering model is better than the lexicalized one. However, MLR is computationally expensive due to iterative learning.
- Dual MLR with shrinking method is almost four times faster than the primal MLR and much more memory-efficient.
- Dual MLR is very beneficial when data cannot fit in memory since primal MLR will take long time due to severe disk-swapping.
- The method is applicable for many classification problems in NLP.

Main References

M. Collins, A. Globerson, T. Koo, X. Carreras, and P. L. Bartlett. 2008. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. JMLR

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning, 85(1-2)

[a] King Abdulaziz City for Science and Technology (KACST), Saudi Arabia, asrajeh@kacst.edu.sa

[b] University of Southampton, United Kingdom, {asar1a10, mn}@ecs.soton.ac.uk