Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation

Rui Wang†,*, Hai Zhao†, Bao-Liang Lu†, Masao Utiyama‡, Eiichiro Sumita‡
† Shanghai Jiao Tong University
‡ National Institute of Information and Communications Technology
* Part of this work is finished when Rui Wang visited NICT

Introduction

Background: How to construct efficient large LM is an important topic in SMT. Most of the existing LM growing methods need an extra monolingual corpus, where additional LM adaption technology is necessary.

Main Contributions: We propose a novel neural network based bilingual LM growing method, only using the bilingual parallel corpus in SMT. The results show that our method can improve both the perplexity score for LM evaluation and BLEU score for SMT, and significantly outperforms the existing LM growing methods without extra corpus.

Bilingual LM Growing

Connecting Phrases $w^k_1$ if:

1. $w^k_1$ is the right (rear) part of one phrase $\beta w^k_1$ in the phrase table, or
2. $w^{k+1}_1$ is the left (front) part of one phrase $w^{k+1}_1 \gamma$ in the phrase table.

**Ranking the Connecting Phrases (Grown N-grams):**

$$P_{\text{target}}(e) = \sum_{f} P_{\text{source}}(f) \times P(e|f)$$

$$P_{\text{connect}}(w^k_1 w^{k+1}_1) = \sum_{f=1}^{k+1} \sum_{\beta} P_{\text{target}}(\beta w^k_1) \times \sum_{\gamma} P_{\text{target}}(w^{k+1}_1 \gamma)$$

Calculating Probabilities of Grown N-grams Using CSLM:

Corpus

Phrase Table

Monolingual

CSLM

Input

CSLM

Output

Interpolate

Grown n-grams

with Probabilities

BNLM

Text Data

Converting

Entropy Pruning

2-gram CONV

Renormalized back-off weights

3-gram BNLM

3-gram CONV

Renormalized back-off weights

4-gram BNLM

4-gram CSLM

Append

As BLM

2-gram BNLM

3-gram BNLM

4-gram BNLM

5-gram BNLM

2-gram CSLM

3-gram CSLM

4-gram CSLM

5-gram CSLM

CONV

BNLM

Text Data

Split

Calculate

Replace

Normalize

$P_{\text{shortlist}}(w^k_1) = \{ P_c(w^k_1 | h^k) \times (1 - P_c(o|h^k)) P_o(h^k), \text{otherwise}\}$

Experiments and Results

Corpus:

1. NTCIR-9: 1 million sentences from Chinese to English
2. TED: 186K sentences from Chinese to English (additional monolingual corpus is hard to obtain)

SMT Results:

<table>
<thead>
<tr>
<th>LMs</th>
<th>N-grams</th>
<th>PPL</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNLM</td>
<td>73.9M</td>
<td>108.8</td>
<td>32.19</td>
</tr>
<tr>
<td>CSLM-RE</td>
<td>N/A</td>
<td>97.5</td>
<td>32.42</td>
</tr>
<tr>
<td>Wang2013</td>
<td>73.9M</td>
<td>104.4</td>
<td>32.60</td>
</tr>
<tr>
<td>Arsoy-1</td>
<td>217.6M</td>
<td>103.3</td>
<td>32.55</td>
</tr>
<tr>
<td>Arsoy-2</td>
<td>458.5M</td>
<td>103.0</td>
<td>32.39</td>
</tr>
<tr>
<td>Arsoy-3</td>
<td>712.2M</td>
<td>102.5</td>
<td>32.49</td>
</tr>
<tr>
<td>BI-1</td>
<td>223.5M</td>
<td>101.3</td>
<td>32.02</td>
</tr>
<tr>
<td>BI-2</td>
<td>464.5M</td>
<td>100.6</td>
<td>33.25++</td>
</tr>
<tr>
<td>BI-3</td>
<td>705.5M</td>
<td>100.1</td>
<td>33.24++</td>
</tr>
</tbody>
</table>

NTCIR-9

TED

Decoding Time on Test Data:

<table>
<thead>
<tr>
<th>LMs</th>
<th>Decoding Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNLM</td>
<td>15.3</td>
</tr>
<tr>
<td>CSLM</td>
<td>186.5</td>
</tr>
<tr>
<td>Bilingual Grown LM (BI-2)</td>
<td>16.5</td>
</tr>
</tbody>
</table>