**Highlights**
- Up to 4% improvement over a strong character-level sequence-to-sequence (cED) baseline for three languages
- Improvement over the previous state-of-the-art for two languages, while eliminating the need for external resources such as large dictionaries.
- Including corpus counts is beneficial to both encoder-decoder and classical statistical machine translation systems

**Approach**

Task: züricher → zürich

1. Combine cED with Language Model over morphemes (LM):
   
   **Assume corpus contains:**
   
   - zuōrīn → zuōrī[ə̯n]
   - cED: zuōrī[ə̯n]
   - cED+LM: zuōrī

2. Add Length Control (LC) as a difference in length between the input and its prediction:

   **Assume corpus contains:**
   
   - zuōrīn → zuōrī[ə̯n]
   - zuōrī → zuōrī
   - cED+LM: zuōrī
   - cED+LM+LC: zuōrī

**Morphological Segmentation Problem**

**Example 1 (Chintang)**

- cuwa thaptic
- cuwa thapt -a -khag -a
- water move -IMP -see -IMP
- across
- ‘Bring some water over here!’

   **Canonical segmentation:**

   thaptic → thap t -a | -khag | -a

**Decoding Algorithm**

**Beam Search:** expansion by morphemes with combined score CS (weighted cED,LM,LC).

**Input:** Input word \( x \), beam size \( n \)

**Output:** Predicted segmentation \( h \)

1. Initialize Hypotheses \( = ["<s>" ] \)
2. while not 1-best\(_{cED}(Hypotheses)\) is closed with '<s>':
   
   3. \( \text{New Hypotheses} = [] \)
   4. foreach \( h \in \text{Hypotheses} \)
   5. \( \text{New Hypotheses}.append(h_{i1}, \ldots, h_{in}, \text{from SyncBeam}(h_{i})) \)
   6. end
   
   7. Hypotheses = \( n \)-best\(_{cED}(\text{New Hypotheses})\)
   8. end
   
   9. return 1-best\(_{cED}(\text{Hypotheses})\)

**Scores Synchronization**

**SynCB:ED** expansion by characters with cED score

**Input:** Partial hypothesis \( h \) for an input word \( x \), beam size \( n \)

**Output:** \( n \)-best expansions of \( h \), closed with '<t> or '<s>'.

1. Initialize Hypotheses\(_{[t]}\)
2. while not all \( h \in \text{Hypotheses} \) are closed with '<t>' or '<s>':
   
   3. \( \text{New Hypotheses} = [] \)
   4. foreach \( h \in \text{Hypotheses} \)
   5. \( \text{New Hypotheses}.append(h_{i1}, \ldots, h_{in}, \text{from SyncBeam}(h_{i})) \)
   6. end
   
   7. Hypotheses = \( n \)-best\(_{cED}(\text{New Hypotheses})\)
   8. end
   
   9. return Hypotheses

**Table 2:** Performance on the task of canonical segmentation for Chintang. Comparative setting for cED: training in types regime for the same number of iterations as in the individual setting of token regime.

<table>
<thead>
<tr>
<th>No. of</th>
<th>Error Rate (%)</th>
<th>Tokens Regime</th>
<th>Error Rate (%)</th>
<th>Types Regime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>cED</td>
<td>LM Baseline</td>
<td>cED</td>
</tr>
<tr>
<td>Total</td>
<td>24,606</td>
<td>0.19</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>Seen</td>
<td>19,920</td>
<td>0.13</td>
<td>0.12</td>
<td>0.18</td>
</tr>
<tr>
<td>New</td>
<td>3,959</td>
<td>0.44</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>New morph</td>
<td>727</td>
<td>0.53</td>
<td>0.57</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Table 1:** Performance on the task of canonical segmentation for English, German and Indonesian. cED+LM - character based encoder-decoder model fused with morpheme based language model. Baseline models: cED - character based encoder-decoder model, cSMT - character based statistical machine translation model. For reference only: Joint* - model of [1], cED+RR* - model of [2], not directly comparable since using external dictionary information.

**References**
