Low-Resource Neural Machine Translation

Alexandra Birch, Kenneth Heafield, Proyag Pal

University of Edinburgh

4/12/2017
goal

subword segmentation that:

- uses a closed vocabulary of subword units
- can represent open vocabulary (including unknown words)
- minimizes the sequence length (given the vocabulary size)

solution

- greedy compression algorithm: byte pair encoding (BPE) [Gage, 1994]
- we adapt BPE to word segmentation
- hyperparameter: vocabulary size

<table>
<thead>
<tr>
<th>vocabulary size</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>t+ h+ e i+ n+ d+ o+ o+ r t+ e+ m+ p+ e+ r+ a+ t+ u+ r+ e i+ s v+ e+ r+ y p+ l+ e+ a+ s+ a+ n+ t</td>
</tr>
<tr>
<td>1300</td>
<td>the in+ do+ or t+ em+ per+ at+ ure is very p+ le+ as+ ant</td>
</tr>
<tr>
<td>10300</td>
<td>the in+ door temper+ ature is very pleasant</td>
</tr>
<tr>
<td>50300</td>
<td>the indoor temperature is very pleasant</td>
</tr>
</tbody>
</table>
Deeper models

Figure: Alternating stacked encoder [Zhou et al., 2016].
Figure: Illustration of BiDeep RNN architecture: stack of layers of recurrent cells; each cell is composed of multiple GRU transitions.
why?

monolingual data

- is much less sparse than parallel data
- may be used for domain adaptation

why is this hard?

- standard in SMT: monolingual LM as feature in linear model
- linear combination of NMT and LM barely effective [Gülçehre et al., 2015]

our solution

end-to-end training of NMT model with parallel and monolingual data
NMT is a conditional language model

\[ p(u_i) = f(z_i, u_{i-1}, c_i) \]

**Problem**

for monolingual training instances, source context \( c_i \) is missing
Monolingual Training Instances

solutions: missing data imputation for $c_i$

- missing data indicator: $\overrightarrow{0}$
  → works, but danger of catastrophic forgetting
- impute $c_i$ with translation model
  → we do this indirectly by back-translating the target sentence
- make $c_i$ a copy of the target
  → works especially for names and copied terms
### Transfer learning

Harness resources from other language pairs

- Transfer learning takes knowledge gained while solving one task and applies it to a different but related task.
- Learning from parallel data in other languages: apply domain adaptation techniques to fine-tune multilingual MT models

### Examples

- Zoph et al. (2016) train a high-resource language pair, then fine-tune on a low-resource language pair
  → French-English parent helps Hausa Turkish and Uzbeck to English child languages

- Nguyen and Chiang (2017) initially train on low-resource but related language pair
  → Uzbek-English parent helps and Turkish and Uyghur child languages
