Machine Translation
07: Open-vocabulary Translation

Rico Sennrich
University of Edinburgh

Text Representation

how do we represent text in NMT?

- 1-hot encoding
  - lookup of word embedding for input
  - probability distribution over vocabulary for output
- large vocabularies
  - increase network size
  - decrease training and decoding speed
- typical network vocabulary size: 10,000–100,000 symbols

Problem

translation is open-vocabulary problem

- many training corpora contain millions of word types
- productive word formation processes (compounding; derivation) allow formation and understanding of unseen words
- names, numbers are morphologically simple, but open word classes

<table>
<thead>
<tr>
<th>vocabulary</th>
<th>representation of &quot;cat&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>the 0</td>
</tr>
<tr>
<td>1</td>
<td>cat 1</td>
</tr>
<tr>
<td>2</td>
<td>is 0</td>
</tr>
<tr>
<td>1024</td>
<td>mat 0</td>
</tr>
</tbody>
</table>
Non-Solution: Ignore Rare Words

- replace out-of-vocabulary words with UNK
- a vocabulary of 50,000 words covers 95% of text

this gets you 95% of the way...
... if you only care about automatic metrics

why 95% is not enough
rare outcomes have high self-information

Solution 1: Approximative Softmax

approximative softmax [Jean et al., 2015]
compute softmax over "active" subset of vocabulary
→ smaller weight matrix, faster softmax

- at training time: vocabulary based on words occurring in training set partition
- at test time: determine likely target words based on source text (using cheap method like translation dictionary)

limitations
- allows larger vocabulary, but still not open
- network may not learn good representation of rare words

Solution 2: Back-off Models

back-off models [Jean et al., 2015, Luong et al., 2015]
- replace rare words with UNK at training time
- when system produces UNK, align UNK to source word, and translate this with back-off method

limitations
- compounds: hard to model 1-to-many relationships
- morphology: hard to predict inflection with back-off dictionary
- names: if alphabets differ, we need transliteration
- alignment: attention model unreliable
Refresher

Open-vocabulary models

- Non-Solution: Ignore Rare Words
- Solution 1: Approximative Softmax
- Solution 2: Back-off Models
- Solution 3: Subword NMT
- Solution 4: Character-level NMT

Subwords for NMT: Motivation

MT is an open-vocabulary problem

- compounding and other productive morphological processes
  - they charge a carry-on bag fee.
  - sie erheben eine Hand|gepäck|gebühr.
- names
  - Obama (English; German)
  - Обама (Russian)
  - オバマ (o-ba-ma) (Japanese)
- technical terms, numbers, etc.

Subword units

segmentation algorithms: wishlist

- **open-vocabulary NMT**: encode all words through small vocabulary
- encoding generalizes to unseen words
- small text size
- good translation quality

our experiments [Sennrich et al., 2016]

- after preliminary experiments, we propose:
  - character n-grams (with shortlist of unsegmented words)
  - segmentation via byte pair encoding (BPE)

Byte pair encoding for word segmentation

bottom-up character merging

- starting point: character-level representation
  - computationally expensive
- compress representation based on information theory
  - byte pair encoding [Gage, 1994]
- repeatedly replace most frequent symbol pair ('A', 'B') with 'AB'
- hyperparameter: when to stop
  - controls vocabulary size

<table>
<thead>
<tr>
<th>word</th>
<th>freq</th>
<th>vocabulary:</th>
</tr>
</thead>
<tbody>
<tr>
<td>'l o w&lt;w&lt;/w&gt;'</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>'l o w e r&lt;w&lt;/w&gt;'</td>
<td>2</td>
<td>l o w&lt;w&lt;/w&gt; w e r&lt;w&lt;/w&gt; n s t&lt;w&lt;/w&gt; i d</td>
</tr>
<tr>
<td>'n e w e s t&lt;w&lt;/w&gt;'</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>'w i d e s t&lt;w&lt;/w&gt;'</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
Byte pair encoding for word segmentation

bottom-up character merging

- starting point: character-level representation
  → computationally expensive
- compress representation based on information theory
  → byte pair encoding [Gage, 1994]
- repeatedly replace most frequent symbol pair ('A', 'B') with 'AB'
- hyperparameter: when to stop
  → controls vocabulary size

<table>
<thead>
<tr>
<th>word</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>'low'</td>
<td>5</td>
</tr>
<tr>
<td>'lower'</td>
<td>2</td>
</tr>
<tr>
<td>'newest'</td>
<td>6</td>
</tr>
<tr>
<td>'widest'</td>
<td>3</td>
</tr>
</tbody>
</table>

vocabulary:

vw

e

why BPE?

- open-vocabulary:
  operations learned on training set can be applied to unknown words
- compression of frequent character sequences improves efficiency
  → trade-off between text length and vocabulary size

<table>
<thead>
<tr>
<th>word</th>
<th>freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>'low'</td>
<td>5</td>
</tr>
<tr>
<td>'lowest'</td>
<td>2</td>
</tr>
<tr>
<td>'newest'</td>
<td>6</td>
</tr>
<tr>
<td>'widest'</td>
<td>3</td>
</tr>
</tbody>
</table>

es → es
esest → est
lo → lo
**Byte pair encoding for word segmentation**

**why BPE?**
- open-vocabulary: operations learned on training set can be applied to unknown words
- compression of frequent character sequences improves efficiency → trade-off between text length and vocabulary size

'es' → 'es'
es t</w> → est</w>'
'l o' → 'lo'

**Evaluation: data and methods**

**data**
- WMT 15 English → German and English → Russian

**model**
- attentional encoder–decoder neural network
- parameters and settings as in [Bahdanau et al, 2014]
Subword NMT: Translation Quality

BLEU

EN-DE EN-RU

0.0 10.0 20.0 22.0

19.1 20.4

| word-level NMT (with back-off) | [Jean et al., 2015] |
| subword-level NMT: BPE |

Subword NMT: Translation Quality

NMT Results EN-RU

NMT Results EN-RU

BLEU

1.0

0.8

0.6

0.4

0.2

0.0

50 000 500 000

training set frequency rank

1.0

0.8

0.6

0.4

0.2

0.0

50 000 500 000

training set frequency rank

1.0

0.8

0.6

0.4

0.2

0.0

50 000 500 000

training set frequency rank

Examples

system sentence
source health research institutes
reference Gesundheitsforschungsinstitute
word-level (with back-off) Forschungsinstitute
character bigrams Forschungs|inst|it|ut|io|ne|n
BPE Gesundheits|forsch|ungsin|stitute

source rakfisk
reference пакфиска (rakfiska)
word-level (with back-off) rakfisk → UNK → rakfisk
character bigrams rakf|l|s|k → пакфиска (rakfiska)
BPE rak|f|isk → пакфиска (rakfiska)

MT – 2018 – 07

Open-vocabulary models

1. Refresher

2. Open-vocabulary models

- Non-Solution: Ignore Rare Words
- Solution 1: Approximative Softmax
- Solution 2: Back-off Models
- Solution 3: Subword NMT
- Solution 4: Character-level NMT
Character-level Models

- advantages:
  - (mostly) open-vocabulary
  - no heuristic or language-specific segmentation
  - neural network can conceivably learn from raw character sequences
- drawbacks:
  - increasing sequence length slows training/decoding
    (reported x2→x4 increase in training time)
  - naive char-level encoder-decoders are currently resource-limited
    [Luong and Manning, 2016]
- open questions
  - on which level should we represent meaning?
  - on which level should attention operate?

Character-level Models

classification output [Chung et al., 2016]

- no word segmentation on target side
- encoder is BPE-level
- good results for EN→{DE,CS,RU,FI}
- long training time (≈ x2 compared to BPE-level model)

Character-level Models

- hierarchical model: back-off revisited [Luong and Manning, 2016]
  - word-level model produces UNKS
  - for each UNK, character-level model predicts word based on word hidden state
  - pros:
    - prediction is more flexible than dictionary look-up
    - more efficient than pure character-level translation
  - cons:
    - independence assumptions between main model and backoff model

Character-level Models

character-level input [Ling et al., 2015]

hierarchical representation: RNN states represent words, but their representation is computed from character-level LSTM

The figure shows a diagram of a character-level model with LSTM layers and character vectors.
Fully Character-level NMT [Lee et al., 2016]

- goal: get rid of word boundaries
- character-level RNN on target side
- source side: convolution and max-pooling layers

Bibliography I


Bibliography II


Conclusion

- BPE-level subword segmentation is currently the most widely used technique for open-vocabulary NMT
- character-level models are theoretically attractive, but currently require specialized architectures and more computational resources
- the presented methods allow open vocabulary; how well we generalize is other question
- next lecture: morphology