Data-Driven Supervised Morpheme Extraction

Yuntae Kim, Ramya Narayanaswamy, Dariusz Kuc
Mentor : Ahmed El-kishky
Text Segmentation for Data Mining

- 2.5 Quintillion bytes a day
  - Unstructured data, predominantly text

- Text Segmentation
  - Input text à characters, subwords, words, phrases and sentences
  - Segmented text à features for downstream data mining applications

- Example
  - Features(characters, words) à entity recognition as classification
What are Morphemes

- Morpheme
  - Subword, smallest meaningful unit

- Morpheme Extraction
  - Segmentation task

- aerophobia -> aero + phobia
- acrophobia -> acro + phobia
Motivation for Morpheme-Based Segmentation

- Natural Language Processing
  - Better word embedding

- Machine Translation
  - Alignment in morphologically rich languages
  - Rare and out-of-vocabulary words

- Information Retrieval
  - Medical documents
Notable Existing Morpheme Extraction Methods

- Byte-Pair Encoding (BPE)
- Unigram Language Model (ULM)
- Local minima of substring transition probability
- Viterbi for MDL and ML cost optimization

All are unsupervised methods
Supervised Morpheme Extraction

- Label for every split point of a word
  - “aerophobia” : [0 0 0 1 0 0 0 0 0]

- Features for every split point of a word
  - Statistical Features
    - Identify change in values, potentially signaling a split boundary
  - Character N-gram

- Probabilistic random forest classifier
  - Threshold for label decision
  - “aerophobia” : RF invoked 10 times for every index
Morpheme Extraction Framework

Feature Generation
For each word/reverse word

For every index
Entropy, subword, char n-gram features

Classifier
For each word/reverse word

For every index
Probabilistic Random Forest Classifier, Thresholding, Decision for every index

Extracted Morphemes for every word

Text Corpus

Vocabulary List

Labels from knowledge bases
Classification Features

• Substring transition entropy (pre, current, post)

\[ H(X) = \sum_{i} \frac{f(Y_i)}{f(X)} \log \frac{f(Y_i)}{f(X)} \]

• Substring count (pre, current, post)
• Character n-grams (uni-, bi-, tri- to left and right)
• Word features à Prefix identification
• Reverse word features à suffix identification
## Illustration of Split-Points and Features

<table>
<thead>
<tr>
<th>Substring entropy</th>
<th>Substring count</th>
<th>Character N-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ae_roplane, enalpor_ea</strong></td>
<td><strong>ae_roplane, enalpor_ea</strong></td>
<td><strong>ae_roplane, enalpor_ea</strong></td>
</tr>
<tr>
<td><strong>Word:</strong></td>
<td><strong>Word:</strong></td>
<td><strong>Word:</strong></td>
</tr>
<tr>
<td>$H(a^*/a): Pre$</td>
<td>$Count(a): Pre$</td>
<td>$(ae, ae, e, r, ro, rop)$</td>
</tr>
<tr>
<td>$H(ae^*/ae): Current$</td>
<td>$Count(ae): Current$</td>
<td></td>
</tr>
<tr>
<td>$H(aer^*/aer): Post$</td>
<td>$Count(aer): Post$</td>
<td></td>
</tr>
<tr>
<td><strong>Reverse word:</strong></td>
<td><strong>Reverse word:</strong></td>
<td><strong>Reverse word:</strong></td>
</tr>
<tr>
<td>$H(enalpo^*/enalpo): Pre$</td>
<td>$Count(enalpo): Pre$</td>
<td>$(por, po, r, e, ea, ea)$</td>
</tr>
<tr>
<td>$H(enalpor^*/enalpor): Current$</td>
<td>$Count(enalpor): Current$</td>
<td></td>
</tr>
<tr>
<td>$H(enalpore^*/enalpore): Post$</td>
<td>$Count(enalpore): Post$</td>
<td></td>
</tr>
</tbody>
</table>
Results

Accuracy

BPE  Unigram  N-grams  Subword count  Entropy  Subword + N-grams  Entropy + Subword  Entropy + N-grams  Entropy + Subword + N-grams
### Effects of threshold

<table>
<thead>
<tr>
<th>Test Set Parameters (%)</th>
<th>Threshold @ 0.5</th>
<th>Threshold @ 0.4</th>
<th>Threshold @ 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Accuracy</td>
<td>49.5</td>
<td>59.24</td>
<td>65.43</td>
</tr>
<tr>
<td>No-split</td>
<td>28.8</td>
<td>15.66</td>
<td>5.75</td>
</tr>
<tr>
<td>Wrong-split</td>
<td>21.7</td>
<td>25.1</td>
<td>28.82</td>
</tr>
</tbody>
</table>

#### Accuracy

![Accuracy Graph](image)
## Sample words and roots

<table>
<thead>
<tr>
<th>Word</th>
<th>Root</th>
<th>BPE</th>
<th>ULM</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td>inwards</td>
<td>in</td>
<td>in-wards</td>
<td>in-wards</td>
<td>in-wards</td>
</tr>
<tr>
<td>xylophone</td>
<td>xyl</td>
<td>x-y-lo-phone</td>
<td>x-y-lo-phone</td>
<td>xylo-phone</td>
</tr>
<tr>
<td>heterozygous</td>
<td>zyg</td>
<td>heter-o-zy-g-ous</td>
<td>hetero-zy-g-ous</td>
<td>hetero-zyg-ous</td>
</tr>
<tr>
<td>preponderance</td>
<td>ponder</td>
<td>pre-p-on-der-ance</td>
<td>pre-pon-der-ance</td>
<td>preponder-ance</td>
</tr>
<tr>
<td>carpology</td>
<td>carp</td>
<td>car-p-o-logy</td>
<td>car-p-o-logy</td>
<td>carp-o-logy</td>
</tr>
<tr>
<td>commemorate</td>
<td>com</td>
<td>comm-em-or-ate</td>
<td>co-mme-mo-rate</td>
<td>com-memorate</td>
</tr>
<tr>
<td>osteoporosis</td>
<td>osteo</td>
<td>O-st-eo-por-osis</td>
<td>O-ste-op-or-osis</td>
<td>Osteo-porosis</td>
</tr>
</tbody>
</table>
Future Work and Conclusion

- Iterative classification
  - Like SegPhrase+

- To prevent under-splitting
  - Binary recursive classifier

- Supervised min accuracy (~36%) > Unsupervised max accuracy (~20%)
- Supervised best accuracy (~65%)
  - 225% increase
Thank you!