POSTER Improving Neural Morphological Segmentation for Minimal-Resource Languages

Katharina Kann, Manuel Mager, Ivan Meza, and Hinrich Schütze

1 CIS, LMU Munich, Germany
2 Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México

Morphological Segmentation

The segmentation task aims to split a word into the surface forms of its smallest meaning-bearing units, its morphemes, i.e.: nepitikuye kai (wixarika), (English translation).

Research Questions

1. How can we successfully segment words in polysynthetic languages?
2. Which supervised methods are applicable in minimal-resource settings and how can they be improved?

Polysynthetic Languages

Polysynthetic languages are languages which are highly synthetic, i.e., single words can be composed of many individual morphemes. We experiment on four languages of the Yuto-Aztecan:

- Mexicanero
- Nahuatl
- Wixarika
- Yorem Nokki

Architecture

Attention-based encoder-decoder gated recurrent neural network (Bahdanau et al., 2015).

Hyperparameters

- 100-dimensional hidden layers in encoder and decoder
- 300-dimensional embeddings
- Training: stochastic gradient descent, Adadelta
- Mini-batch size 20

Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Mexicanero</th>
<th>Nahuatl</th>
<th>Wixarika</th>
<th>Yorem Nokki</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2S</td>
<td>0.8051</td>
<td>0.6004</td>
<td>0.5895</td>
<td>0.6856</td>
</tr>
<tr>
<td>CRFS</td>
<td>0.7837</td>
<td>0.6444</td>
<td>0.5866</td>
<td>0.6596</td>
</tr>
<tr>
<td>MTT-U</td>
<td>0.7611</td>
<td>0.5541</td>
<td>0.5425</td>
<td>0.6212</td>
</tr>
<tr>
<td>MTT-R</td>
<td>0.7504</td>
<td>0.5585</td>
<td>0.5754</td>
<td>0.6569</td>
</tr>
</tbody>
</table>

Amount of Additional Data

We treat the amount of additional and artificial data as an hyperparameter. Values we experiment with are

\[ m \times \text{amount of instances in the original training set}, \]

with \( m \in \{1, 2, 4, 8\} \).

Conclusions

- We investigated the applicability of neural encoder-decoder models for surface segmentation.
- We proposed 2 novel multi-task approaches and 2 novel data augmentation methods for surface segmentation.
- We investigated the applicability of neural encoder-decoder models.

Acknowledgements

Support of Gerardo Sierra, 2016-01-2225. We also thank the To CONACyT (Program No. FC-0699).

Contact

Katharina Kann: kann@cis.lmu.de
Manuel Mager: mmager@turing.iimas.unam.mx

Improving Neural Segmentation

We extend the available training data with new examples from unlabeled data set (DA-U) and random strings (DA-R), such that

\[ \mathcal{L}(w, c) = \sum_{e(w)} \log p(c | e(w)) + \sum_{a} \log p(a | e(a)) \]

Using Random Strings (MTT-R) and unlabeled words (DA-R), our model performs better than the original input string.