Arabic Diacritization with Recurrent Neural Networks

Yonatan Belinkov, James Glass
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
{belinkov,glass}@mit.edu

1. Overview

- Arabic, Hebrew, and similar languages are typically written without diacritics.
- Diacritization is important for core tasks like speech recognition and morphological analysis.
- Previous work relied on external resources (e.g., morphological analyzers).
- We develop a recurrent neural network (RNN) for diacritization, with long short-term memory (LSTM), trained solely from diacritized texts.
- We achieve state-of-the-art results without relying on external resources.

2. Diacritization

- Arabic words are highly ambiguous without diacritics.
- Previous work relied on external resources (e.g., morphological analyzers).
- We develop a recurrent neural network (RNN) for diacritization, with long short-term memory (LSTM).

3. Approach

- Diacritization as sequence classification:
  - Map character sequence to label sequence
    \( (w_1, \ldots, w_T) \rightarrow (l_1, \ldots, l_T) \)
  - A label can be 0, 1, or more diacritics.

- RNN Architecture
  - Output layer
  - Hidden layers
  - Input layer

4. Experiments

- Diacritic texts extracted from the Arabic Treebank.
- Diacritic combinations treated as separate label.

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>End</th>
<th># params</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feed-forward</td>
<td>11.76</td>
<td>22.90</td>
<td>6.6K</td>
</tr>
<tr>
<td>LSTM</td>
<td>11.55</td>
<td>23.40</td>
<td>90.8K</td>
</tr>
<tr>
<td>2-layer B-LSTM</td>
<td>6.16</td>
<td>9.85</td>
<td>518K</td>
</tr>
<tr>
<td>3-layer B-LSTM</td>
<td>5.77</td>
<td>9.18</td>
<td>916K</td>
</tr>
</tbody>
</table>

5. Implementation Details

- LSTM hidden layers: memory cells reuse long term dependencies over the sequence (Graves et al. 2013)
- Experiment with other languages, genres, and dialects.
- Incorporate diacritizer in a speech recognizer.

References

- Weninger et al. 2015. Introducing CURRENT. JMLR.
- Al Hanai and Glass. 2014. Lexical Modeling for Arabic ASR. INTERSPEECH.