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)

2. Diacritization

Problem Definition
- Given a training text with diacritics, learn a model that will predict diacritics in a test text without diacritics.

Ambiguity
- Arabic words are highly ambiguous without diacritics:

| Word | Class |
|------|-------|
| Eita | unknown |
| Eita | he taught |
| Eita | knowledge (def noun) |
| EitaK | flag (inf def gen) |

Arabic Diacritics

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

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

Results
- LSTM outperforms simple feed-forward networks
- Bidirectional LSTM is better than unidirectional
- Deeper models are better than shallow ones
- LSTM better at case endings (long dependencies)

| Model | All Test | End Test | # params |
|-------|----------|----------|----------|
| Feed-forward | 11.76 | 22.90 | 6K |
| Feed-forward (large) | 11.55 | 23.40 | 90K |
| LSTM | 6.98 | 10.36 | 838K |
| B-LSTM | 6.16 | 9.85 | 518K |
| 2-layer B-LSTM | 5.77 | 9.18 | 916K |
| 3-layer B-LSTM | 5.08 | 8.14 | 1,498K |

Diacritic error rates (DERs) on the Test set, all diacritics and only at word ending.

Error Analysis
- Most errors from confusing short vowels
- Qualitative analysis shows how LSTM captures long-distance dependencies like case endings

| Model | Diacritization |
|-------|----------------|
| LSTM | 9.18 |
| Feed-forward | 11.55 |
| B-LSTM | 6.16 |

5. Implementation Details

- Stack previous and future letter vectors in a context window
- Linear projection after input layer learns new representation
- Cross-entropy objective, optimized with SGD
- Hyper-parameters tuned on the development set
- Implemented with Current (Weninger et al. 2015)

6. Future Work

- Experiment with other languages, genres, and dialects
- Incorporate diacritizer in a speech recognizer
- Replace external tools like MADA (Al Hanai and Glass 2014)

References

- Belinkov et al. 2013. Speech recognition with deep recurrent neural networks. ICASSP.
- Weninger et al. 2015. Introducing CURRENNT. JMLR.
- Al Hanai and Glass. 2014. Lexical Modeling for Arabic ASR. INTERSPEECH.