# 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

### Problem Definition

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

اعتبر المدير العام → اِعْتَبَرَ الْمُدِيرُ الْعَامُّ

AEtbr Almdyr AlEAm      AiEotabara Almudiyru AlEAm~

### Ambiguity

- Arabic words are highly ambiguous without diacritics:

| Word | Gloss |
|---|---|
| *Ealima* | he knew |
| *Eulima* | it was known |
| *Eal~ama* | he taught |
| *Eilomu* | knowledge (def.nom) |
| ... | ... |
| *EalamK* | flag (indef.gen) |

Possible diacritized forms for علم *Elm*.

### Arabic Diacritics

| Diacritic | Transliteration | Transcription |
|---|---|---|
| دَ | *a* | /a/ |
| دُ | *u* | /u/ |
| دِ | *i* | /i/ |
| دً | *F* | /an/ |
| دٌ | *N* | /un |
| دٍ | *K* | /in/ |
| دّ | ~ | Gemination |
| دْ | *o* | No vowel |

## 3. Approach

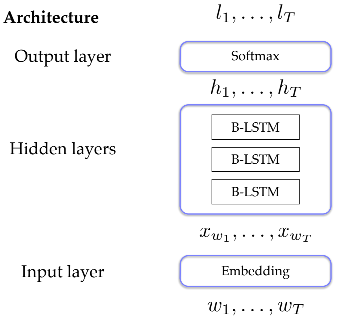
### Diacritization as sequence classification

- 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



$l_1, \ldots, l_T$

Output layer: Softmax

$h_1, \ldots, h_T$

Hidden layers: B-LSTM, B-LSTM, B-LSTM

$x_{w_1}, \ldots, x_{w_T}$

Input layer: Embedding

$w_1, \ldots, w_T$

## 4. Experiments

### Data

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

| | Train | Dev | Test |
|---|---|---|---|
| Words | 470K | 81K | 80K |
| Letters | 2.6M | 438K | 434K |

Arabic diacritization corpus statistics.

### 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)
- LSTM beats competitor lexical MaxEnt with access to same information
- LSTM rivals MaxEnt with access to a segmenter and part-of-speech tagger

| | DER | | |
|---|---|---|---|
| Model | All | End | # params |
| Feed-forward | 11.76 | 22.90 | 63K |
| Feed-forward (large) | 11.55 | 23.40 | 908K |
| 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 |

Diacrtic error rates (DERs) on the Dev set, over all diacritics and only at word ending.

| | |
|---|---|
| MaxEnt (only lexical) | 8.1 |
| MaxEnt (full) | 5.1 |
| 3-layer B-LSTM | 4.85 |

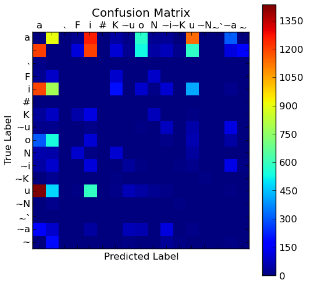
Results (DER) on the Test set. MaxEnt results from (Zitouni and Sarikaya, 2009).

### Error Analysis

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

| Model | Diacritization |
|---|---|
| Gold | AiEotabara Almudiyru AlEAm~u l " Aln~ahAri " juborAn tuwayoniy~ Aan Alt~a$okiylAti AlqaDA}iy~apa jA'at litamoyiyEi milaf~i [...] |
| Feed-forward | AiEotabara Almudiyru AlEAm~u l " Aln~ahAr_ " jaborAn tuwayoniy_ Aan Alt~a$okiylAti AlqaDA}iy~api jA'at litamayiyEi malaf~i [...] |
| B-LSTM | AiEotabara Almudiyru AlEAm~u l " Aln~ahAri " juborAn t_wayoniy_ Aan Alt~a$okiylAti AlqaDA}iy~apa jA'at litamoyiyEi milaf~i maHaT~api [...] |

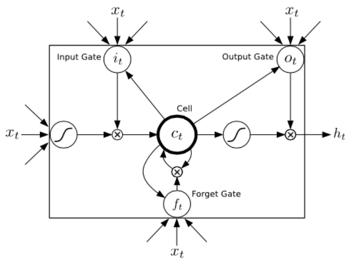
Errors by two diacritization models. Wrong diacritics underlined in red. Translation: "The editor [. . . ] thought that the judicial formations came to dilute the issue of [. . . ]".



A confusion matrix of errors made by our system.

## 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 Currennt (Weninger et al. 2015)
- **LSTM hidden layers**: memory cells reuse long term dependencies over the sequence (Graves et al. 2013)



## 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

- Graves 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*.

**Acknowledgments**: This work was supported by the Qatar Computing Research Institute (QCRI). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.