LREC-COLING 2024

aixplain

Arabic Diacritization Using Morphologically Informed Character-Level Model

Muhammad Elmallah, Mahmoud Reda, Kareem Darwish, Abdelrahman El-Sheikh, Ashraf Elneima, Murtadha Aljubran, Nouf Alsaeed, Reem Mohammed, Mohamed Al-Badrashiny

Joint work with SDAIA الهيئة السعودية للبيانات والذكاء الاصطناعي

Motivation

aixplain

- Arabic words are composed of letters and short vowels called diacritics.
- Diacritics:
  - Typically attached to consonants
  - Generally omitted in writing
  - Speakers need to reintroduce them to properly pronounce words
  - Have two types:
    - Core-word diacritics: disambiguate words in context
    - Case endings: appear at the end of words and often disambiguate syntactic role
  - Example: ذهب الولد ⇒
    - ذَهَبُ الوَلَدِ جَمِيلٌ – The boy's gold is beautiful
    - ذَهَبَ الوَلَدُ جَمِيلٌ – The boy, Jameel, went
- Automatic diacritic recovery is critical for tasks such as Text-To-Speech (TTS) and language learning.

Prior Approaches

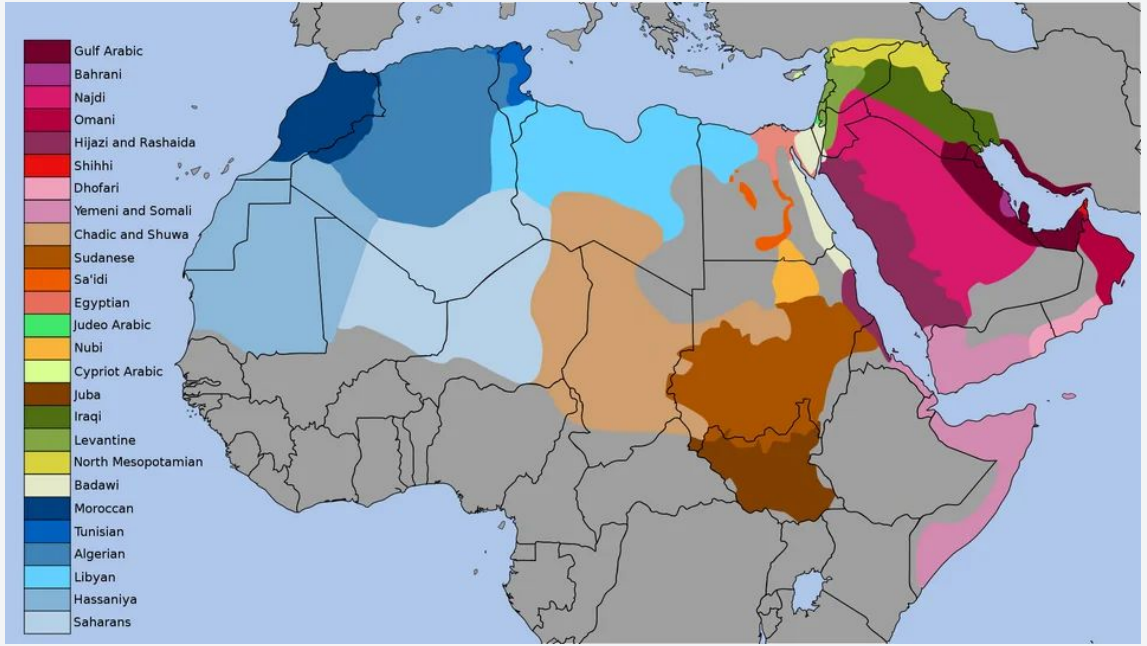
aiXplain

- Classical approaches (before 2002): rule based
- Classical ML approaches:
  - Finite-state automata/hidden Markov model (@ word-level) for core-word diacritics
  - Sequence labeling (ex. SVM) for case ending
- Deep learning approaches:
  - Decoupling of core-words and case endings: ex. HMM for core words + RNN for case ending
  - Core-words + case endings:
    - Character level RNN
    - Character level seq-to-seq translation from letters to diacritized letters

Varieties of Arabic

aixplain

- Classical Arabic:
  - Classical and religious texts
- Modern Standard Arabic:
  - Formal language in most of the Arab World
- Dialectal Arabic:
  - Daily language that differs from region to the next



Available Data

aiXplain

- Modern Standard Arabic:
  - Arabic Penn Treebank – roughly 400k words
  - Proprietary datasets – > 7 million words
- Classical Arabic:
  - Tashkeela – 2.3 million words
  - Shamela – partially diacritized – > 100 million words (partially diacritized)
- Dialectal Arabic:
  - Kurras (Palestinian)
  - Bible (Moroccan & Tunisian)
  - All a few thousand words each

Our Methodology

aiXplain

- Approach: Morphologically informed character-level RNN model
- Since, Arabic words are composed of concatenated clitics; and (in limited cases for MSA/common for dialects) concatenated words:
  - Alwld (الولد) ⇒ Al+wld (the+boy)
  - wktAbhm (وكتابهم) ⇒ w+ktAb+hm (and+book+their – and their book)
  - qAlly (قاللي) ⇒ qAl+ly (he said+to me) – dialectal
- Most important departure from prior methods:
  - **Inserted clitic delimiters inside the words – to inform morphology.**
  - **Developed a robust word tokenization model.**

Our Segmentation Model

aiXplain

- Training data:
  - For MSA & Classical Arabic, we used 1.5 million unique words from Arabic Wikipedia segmented using Farasa
  - For dialects, we used a dataset composed of 27,366 words (Gulf, Egyptian, Levantine, and Maghrebi).
- Given a sequence of input letters, the output for each letter would be:
  - B – beginning of clitic
  - I – intermediate in clitic
  - E – end of clitic
  - S – single character clitic
  - Example: wAlwld (w+Al+wld) ⇒ SBEBIE

Our Segmentation Model

aixplain

- Model:
  - Input: characters (dim: 15)
  - Character Embeddings Layer (dim: 50)
  - 2 biLSTM layers (dim: 100)
  - 4 biLSTM layers (dim: 400)
  - CRF layer
- Results:
  - For MSA & Classical Arabic: 98.4% (Farasa is 99%)
  - For dialects: 93.1%

Our Diacritization Model

aiXplain

- Given a sequence of input letters, the output for each letter would be:
  - Diacritic
  - Null – for segment separator or letters without diacritics
  - Example: w+Al+wld ⇒ {a, null, null, o, null, a, a, u} (wa+Alo+waladu)
- Training data:

Split	Classical	MSA	Moroccan	Tunisian
Train	2,458,113	4,157,656	116,400	114,037
Test	125,098	18,300	29,130	28,501
Dev	119,958	18,017	12,933	12,671

Our Diacritization Model

aiXplain

- Model:
  - Input: characters (dim: 200)
  - Character Embeddings Layer (dim: 50)
  - 2 biLSTM layers (dim: 512 – 25% drop out)
  - Dense layer (dim: 512)
  - Dense layer (dim: 256)
  - Softmax
- Results:

	w/o Segmentation				w/ Segmentation			
	WER		CER		WER		CER	
	w/o CE	w/ CE	w/o CE	w/ CE	w/o CE	w/ CE	w/o CE	w/ CE
MSA	3.4	5.7	1.1	1.4	1.9	3.4	1.3	1.4
Classical	4.5	7.3	2.0	2.4	2.7	5.4	2.5	2.8
Moroccan	4.1	-	1.5	-	2.0	-	1.6	-
Tunisian	9.6	-	4.0	-	3.1	-	2.6	-

Our Diacritization Model

aixplain

- Comparison to SOTA results:

MSA System	WER
(Belinkov and Glass, 2015)	30.5
(Pasha et al., 2014)	19.0
(Obeid et al., 2020)	15.6
(Rashwan et al., 2015)	16.0
(Darwish et al., 2017)	12.8
(Mubarak et al., 2019b)	4.5
Ours	**3.4**

Classical Arabic	WER
(Fadel et al., 2019)	11.2
(AlKhamissi et al., 2020) (d2/d3)	5.5/**5.3**
Ours	5.4
Moroccan	**WER**
(Darwish et al., 2018a)	2.9
(Mubarak et al., 2019b)	**1.4**
Ours	2.0
Tunisian	**WER**
(Darwish et al., 2018a)	3.8
(Mubarak et al., 2019b)	**2.5**
Ours	3.1

Contributions

aixplain

- Built novel morphological segmentation model for Arabic (MSA/Classical/Dialectal)
- Trained a morphologically informed Arabic diacritization model that works well across different varieties of Arabic
- Achieved results that beat or match SOTA results

Thanks

kareem.darwish@aixplain.com