Learning Meters of Arabic and English poems

With Recurrent Neural Networks

Prof. Waleed A. YOUSEF

The Team

June 18, 2018

Computer Science department
Faulty of Computers and Information, Helwan University

Table of contents

1. Introduction

2. Literature Review

3. Datasets

4. Methodology

5. Results

Introduction

Hello, Arabic

ودعْ عنك آراءَ الرجالِ وقولَهم   فقولُ رسولِ الله أزكى وأشرحُ

But ... What is poetry?

General Definition:

- **Poetry** is a piece of writing or speaking, which **MUST** follow specific **Patterns**.

Example, *English verse*:

That **time** of **year** thou **mayst** in **me** be**hold**

To detect poems' meters, we need to learn those **Patterns**.

Arabic Prosody العَرُوض

- **Foot التفعيلة**: is a sequence of vowels and consonants.

Feet	Scansion
فَعُولُنْ	0/0//
فَاعِلُنْ	0//0/
مُسْتَفْعِلُنْ	0//0/0/
مَفَاعِيلُنْ	0/0/0//
مَفْعُولَاتُ	0//0///
فَاعِلَاتُنْ	0/0//0/
مُفَاعَلَتُنْ	0///0//
مُتَفَاعِلُنْ	0//0///

Arabic Prosody العَرُوض

Arabic Patterns/Meters بحور الشعر:

- **Meter البحر**: is a sequence of feet.

Meter Name	**Meter** *feet combination*
al-Wafeer	مُفَاعَلَتُن مُفَاعَلَتُن فَعُولُن
al-Taweel	فَعُوْلُنْ مَفَاْعِيْلُنْ فَعُوْلُنْ مَفَاْعِلُنْ
⋮	⋮
al-Moktadib	مَفْعُوْلاتُ مُسْتَفْعِلُنْ مُسْتَفْعِلُنْ
al-Modar'e	مَفَاْعِيْلُنْ فَاْعِلاتُنْ مَفَاْعِيْلُنْ

Arabic Prosody, example!

From بحر الوافر:

ويسْأل فىْ الْحواْدث ذوْ صواْبٍ

صوابن	حوادث ذو	ويسأل فل
0/0//	0///0//	0///0//
فعوْلنْ	مفاْعلتنْ	مفاْعلتنْ

English Prosody

English Meters Building Blocks:

- Syllables: /ˈwɔːtə/ = /ˈwɔː/ + /tə(r)/.
    - **stressed** + unstressed.
- Foot: is a combination of stressed and unstressed syllables.

Feet	Stresses Combination
Iamb	$\times/$
Trochee	$/\times$
Dactyl	$/\times\times$
Anapest	$\times\times/$
Pyrrhic	$\times\times$
Amphibrach	$\times/\times$
Spondee	$//$

Meter: is repeating a foot n times; where $n \in [1, 8]$.

English Patterns

Iambic pentameter verse:

$$\underbrace{\text{That }\textbf{time}}_{\text{Iambic Foot}}\quad \overbrace{\text{of }\textbf{year}}^{\text{2nd}}\quad \overbrace{\text{thou }\textbf{mayst}}^{\text{3rd}}\quad \overbrace{\text{in }\textbf{me}}^{\text{4th}}\quad \overbrace{\text{be}\textbf{hold}\text{.}}^{\text{5th}}$$

Literature Review

Detecting Arabic poems' Meters

Abuata and Al-Omari:

- Five-step Algorithm
  1. Getting the input, carrying full diacritics.
  2. Metrical scansion rules are applied to the Arud writing. 0/0/..
  3. Grouping zero and ones to feet تفعيلات.
  4. A class is assigned to the input.
- **Results**: 82.2% of 417 verses.

Alnagdawi et al, similar approach; Context-Free Grammar; 75% correctly classed from 128.

example!

ويسْأل فىْ الْحواْدث ذوْ صواْبٍ

ويسأل فل	حوادث ذو	صوابن
0///0//	0///0//	0/0//
مفاْعلتنْ	مفاْعلتنْ	فعوْلنْ

Abuata and Al-Omari && Alnagdawi et al; Problems

Issues;

- A huge constrain. Diacritics are a must.
- Converting the text into pronounced text is probabilistic.
  - اثبات الحروف المحذوفة خطاً
  - التصرف فى التقاء الساكنين

Tanasescu et al.

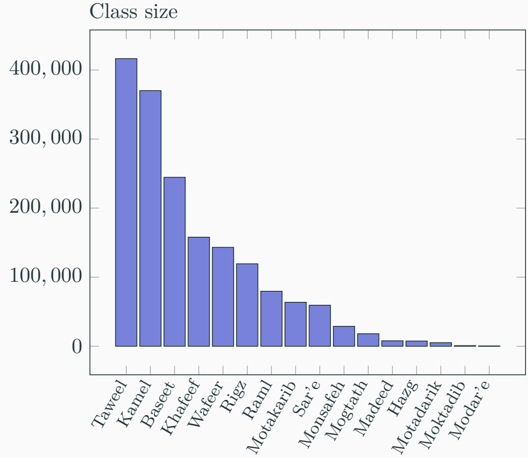
Binary Classification; Metric or Free-Verse:

- verses are represented as vectors of statistical features.

Datasets

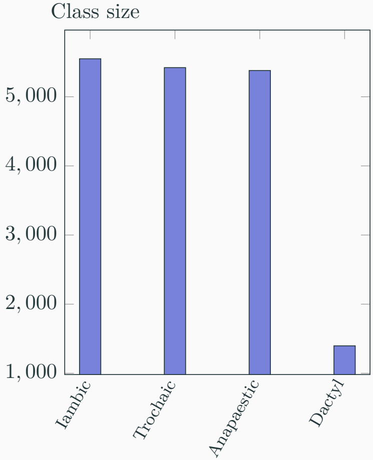
Datasets

Arabic Dataset:



Datasets

English Dataset:

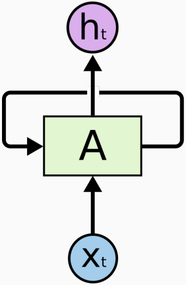


Methodology

Which Network!

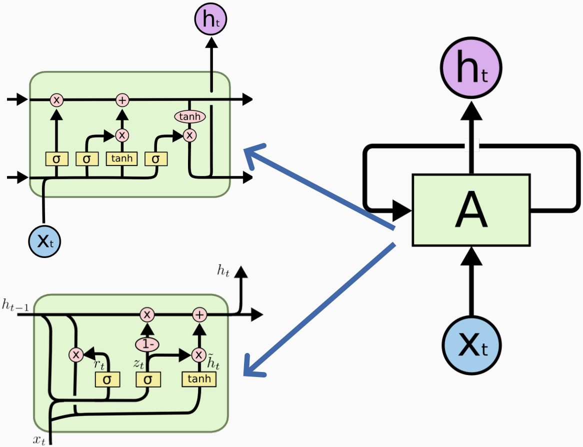
– **Pattern**: is a sequance of characters.

- Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.



Rolled Rnn unit

RNN, Architectures



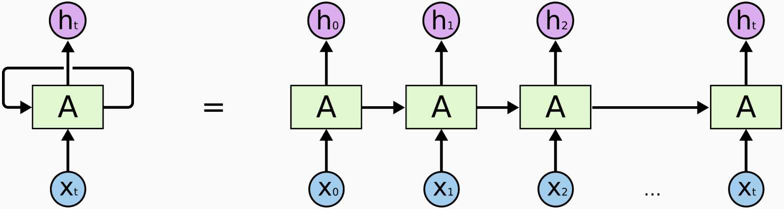
LSTM unit

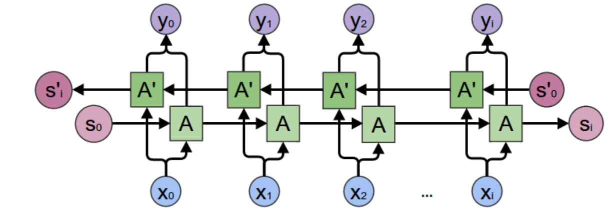
GRU unit

- Two variants of unidirectional *recurrent units*.

RNN, Architectures

Unidirectional & Bidirectional RNN:





Data Representation

An Issue:

- Diacritics are standalone characters!
  - `len` مرحبا ≠ `len` مَرْحَبَاً
  - We have represented the letter and its diacritic as a one character.

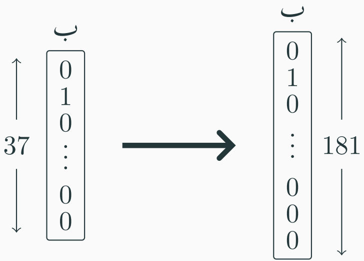
Benefits:

1. Verse's length is fixed, regardless the diacritic states.
2. Saving more space, by shorten the length of full diacritic verses.
3. Models can be tested on both diacritic or non-diacritic data.

Encoding Techniques

1. One-Hot
2. Binary
3. Two-Hot (new technique)

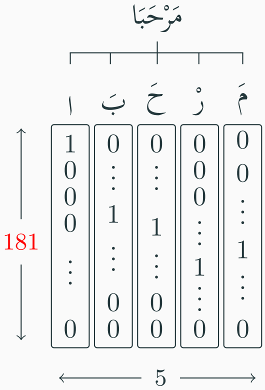
One-Hot



One-Hot Vector: from 37×1 to 181×1

181 is the number of all combination between letters and diacritics.
$181 = 36 + 36 \times 4 + 1$

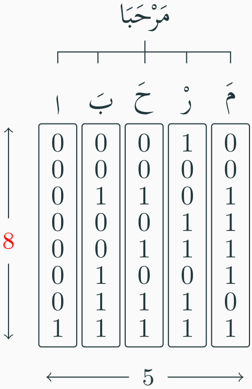
One-Hot, example



Binary

Let n be the vector length.

$n = \lceil \log_2 l \rceil \; l \in \{181, 28\}$

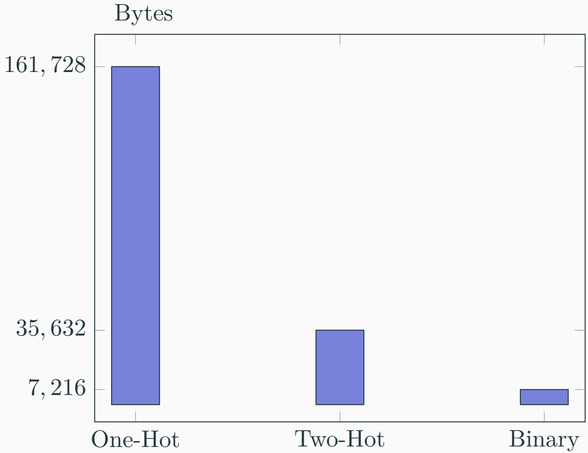


ا	بَ	حَ	رْ	مَ
0	0	0	1	0
0	0	0	0	0
0	1	1	0	1
0	0	0	1	1
0	0	1	1	1
0	1	0	0	1
0	1	1	1	0
1	1	1	1	1

Two-Hot

$$37\updownarrow\underset{m}{\overset{\text{ب}}{\boxed{\begin{matrix}0\\1\\0\\0\\\vdots\\0\end{matrix}}}} + \underset{k}{\overset{\text{◌َ}}{\boxed{\begin{matrix}1\\0\\0\\0\end{matrix}}}} = 41\updownarrow\overset{\text{بَ}}{\boxed{\begin{matrix}1\\0\\0\\0\\0\\1\\0\\\vdots\\0\end{matrix}}}\begin{matrix}\left.\vphantom{\begin{matrix}1\\0\\0\\0\end{matrix}}\right\}\begin{matrix}4\times1\text{ diacritic vector}\\\text{which represents ◌َ}\end{matrix}\\\left.\vphantom{\begin{matrix}0\\1\\0\\\vdots\\0\end{matrix}}\right\}\begin{matrix}37\times1\text{ letter vector}\\\text{which represents بَ}\end{matrix}\end{matrix}$$

Space Comparison



Results

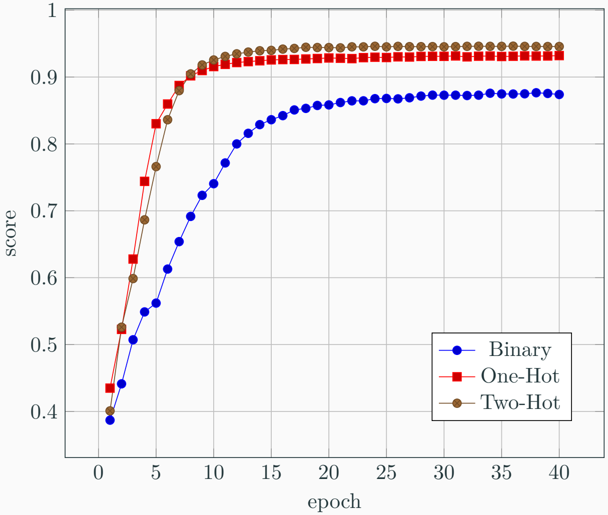
Arabi Results

#	data size	encoding	diacritic	archit.	f1
1	full data	two-hot	Yes	7L, 50U, 0	95.79%
2	full data	two-hot	No	7L, 50U, 0	95.43%
3	full data	binary	Yes	7L, 81U, 0	95.51%
4	full data	binary	No	10L, 30U, 0	93.2%
5	full data	one-hot	Yes	7L, 50U, 1	95.32%
6	full data	one-hot	No	7L, 82U, 0	93.94%
7	eliminated	two-hot	Yes	7L, 81U, 1	95.88%
8	eliminated	two-hot	No	4L, 50U, 1	96.29%
9	eliminated	binary	Yes	7L, 81U, 1	94.87%
10	eliminated	binary	No	4L, 82U, 0	96.38%
11	eliminated	one-hot	Yes	7L, 75U, 0	95.65%
12	eliminated	one-hot	No	7L, 50U, 0	95.04%

English Results

id	encoding	cell type	f1 test
1	one-hot	GRU	81.35%
2	one-hot	LSTM	80.34%
3	binary	LSTM	75.43%
4	binary	GRU	75.04%

Encoding Effect



Binary Encoding Problem

عُ	مَ
0	0
0	0
1	1
0	1
1	1
1	1
0	0
1	1

Questions?