Learning Meters of Arabic and English poems

With Recurrent Neural Networks

Prof. Waleed A. YOUSEF
The Team

June 19, 2018

Computer Science department
Faulty of Computers and Information, Helwan University

Table of contents

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

ودعْ عنك آراءَ الرجالِ وقولَهم   فقولُ رسولِ الله أزكى وأشرحُ

- A **poem** is a collection of verses.
- **Vowels** carry one of ◌ِ ◌ُ ◌َ.
- **Consonants** carry ◌ْ.
- A **foot** التفعيلة: is an **ordered** sequence of vowels and consonants.

Feet	**Scansion**
فَعُولُنْ	0/0//
فَاعِلُنْ	0//0/
مُسْتَفْعِلُنْ	0//0/0/
مَفاعِيلُنْ	0/0/0//
مَفْعُولاَت	0//0///
فَاعِلاَتُنْ	0/0//0/
مُفَاعَلَتُنْ	0///0//
مُتَفَاعِلُنْ	0//0///

- **Meter البحر**: is an **ordered** sequence of feet.

ويسْأل فىْ الْحوادْث ذوْ صوابٍ

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

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

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:

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

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!

- Probabilistic operation

ويسْأل فىْ الْحواْدث ذوْ صواْبٍ

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

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. & Tizhoosh and Dara & Almuhareb et al.

Metric or **Free-Verse**:

- Verses are represented as vectors of statistical features.
  - Average number of feet per line.
  - Longest run of a single foot.
  - Percentage of foot changes.
- He has used an stress-annotated dataset, which means that the pattern is already detected!
- A couples of tries for detecting poem inside documents depending on visual features, for both Arabic and English.
  - Averge of line legth.
  - Averge number of block.
  - etc ...

Our point of departure

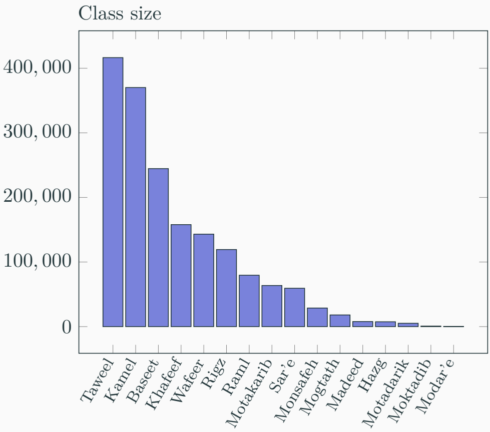
- Detecting poems inside documents by its visual feature represent the majority of the research related to poetry.
- For detecting meters, all models are so naive and primitive. They do no have any clue about the real pattern.

Datasets

Datasets

Arabic Dataset:

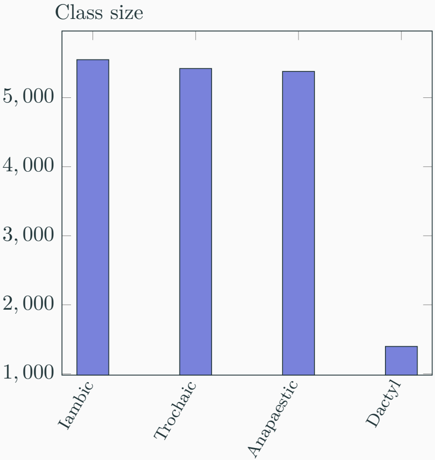
- Our dataset consists of 1,722,321 data points labeled by the 16 meters.
- Verses are for 3701 poets from 11 ages.



Datasets

English Dataset:

- Our dataset consists of 17,744 verses only, labeled by 4 meters.
- This is what we have found after an extensive search.



Motivation

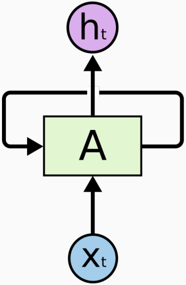
- The effect of the diacritics over the learning curve.
- The effect of the dataset size over the learning curve.
- Best technique to represent character-level text.

Methodology

Which Network!

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

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

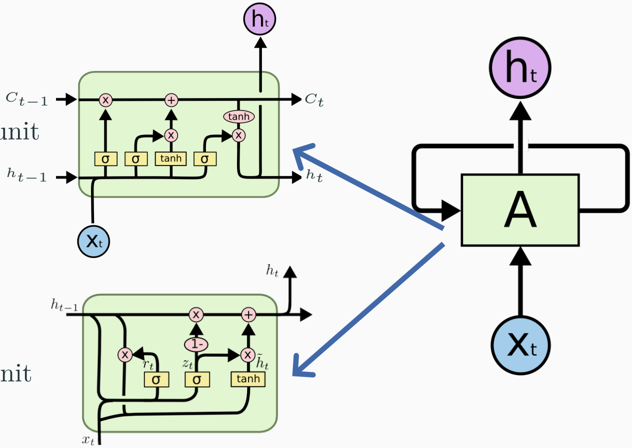


Rolled Rnn unit

RNN, Architectures

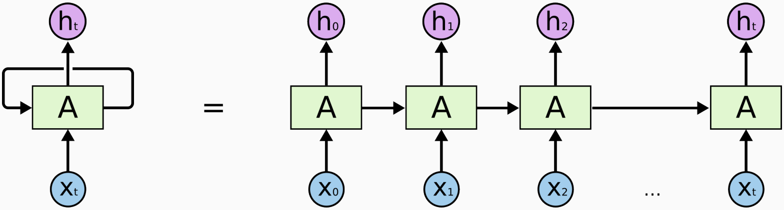
LSTM unit

GRU unit

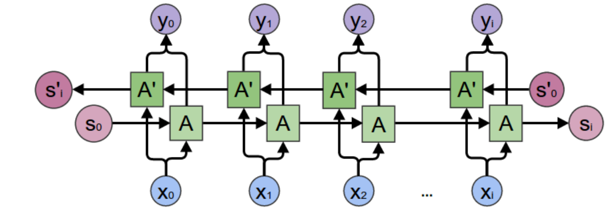


- Two variants of *recurrent units.*
- LSTM and GRU are designed for the same reason, to avoid vanishing problem and long term dependency problem .

RNN, Architectures



Unidirectional RNN



Bidirectional RNN

Data Representation

An Issue:

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

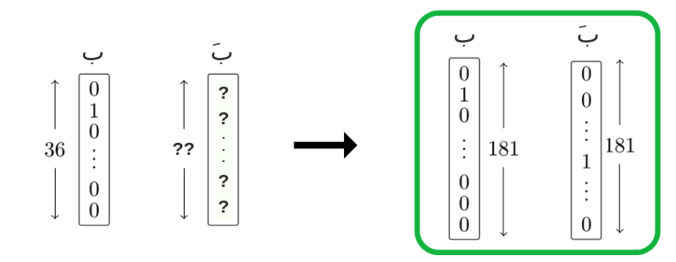
Benefits:

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

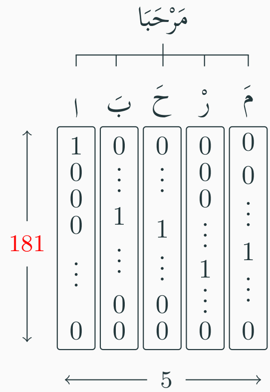


(a) not consistent representation with characters that have diacritic.

(b) consistent representation for diacritics and non-diacritics.

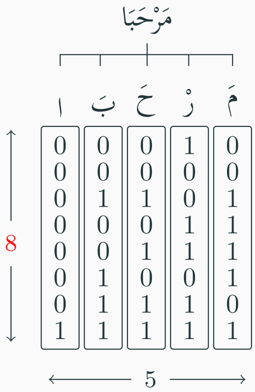
181 is the number of all combination between letters and diacritics.

$181 = 36 + 36 \times 4 + 1$



For English, vector length is 28.

Let n be the vector length.
$n = \lceil \log_2 l \rceil$ $l \in \{181, 28\}$ for Arabic and English alphabet, respectively.

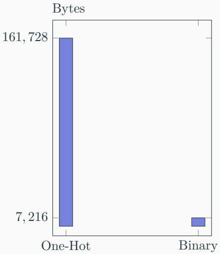


Pros and Cons

One-Hot:

- 181×1 too long vector.
- Wast of memory, compared towith Binary.

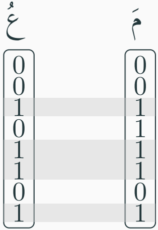
But it's very informative and easy to learn.



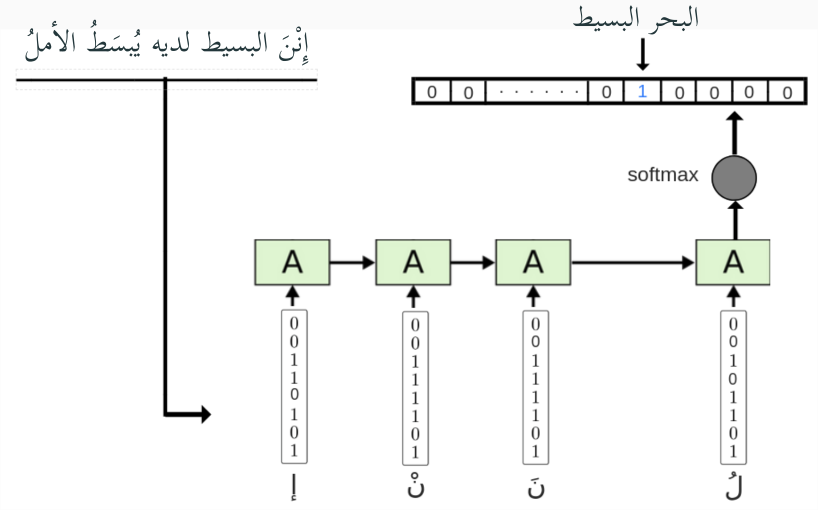
Binary:

- The problem of a common feature, the optimizer updates weights of this two common features to reduce the error of one character of them it will affect the rest characters that have the same feature.

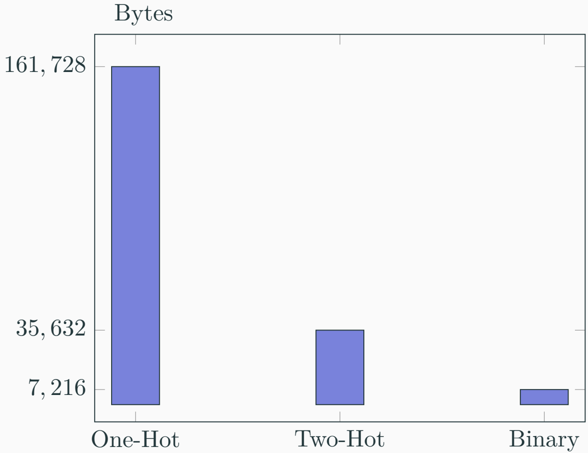
But, memory is well consumed.



Feeding verse to model

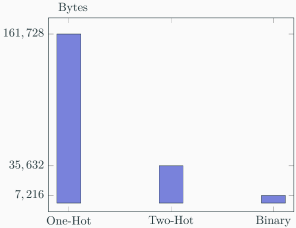


Space Comparison



Two-Hot Encoding

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



- Memory is well consumed.
- Have only one common feature.
- For Arabic only.

Results

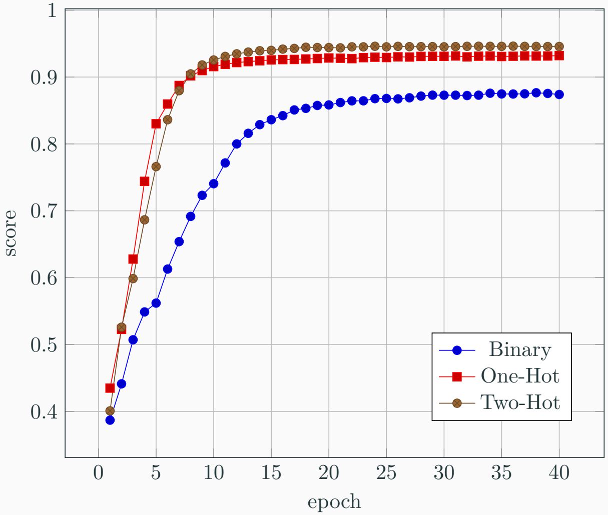
Arabi Results

#	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?