Learning Meters of Arabic Poems with Recurrent Neural Networks

A step forward for language understanding and synthesis

Mostafa A. Mahmoud

A Thesis Presented to the Faculty
of Information Technology and Computer Science
Nile University
In Partial Fulfillment
of the Requirements for the Degree
of Master of Science

Supervised By,
Prof. Samhaa El-Beltagy
Assoc. Prof. Waleed A. Yousef

Table of Contents I

Introduction

But ... What is poetry?

Definition

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

Example

وُلِدَ الهُدى فَالكائِناتُ ضِياءُ وَفَمُ الزَمانِ تَبَسُّمٌ وَثَناءُ

Arabic Prosody العَرُوض

Al-Farahidi (718 – 786 CE) analyzed the Arabic poetry, then he discovered the **Patterns** which is the succession of consonants and vowels.

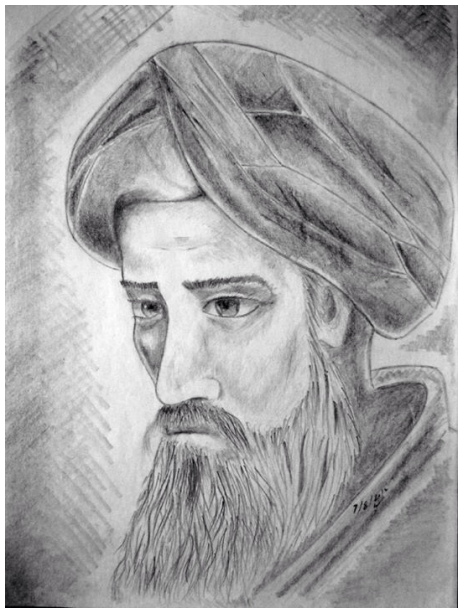


Figure: *Al-Farahidi*

figure taken from `https://goo.gl/ZJySa8`.

Arabic Prosody العَرُوض

- A **poem** is a set of verses having the same meter and rhyme.

Arabic Prosody العَرُوض

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ◌ِ ◌ُ ◌َ which represent the following short vowels /a/, /u/, /i/.

Arabic Prosody العَرُوض

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ◌َ ◌ُ ◌ِ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ◌ْ.

Arabic Prosody العَرُوض

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ◌ِ ◌ُ ◌َ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ◌ْ.
- **Shadaa** indicates the letter is doubled ◌ّ.

Arabic Prosody العَرُوض

- A **poem** is a set of verses having the same meter and rhyme.
- **Vowels** carry one of ◌ِ ◌ُ ◌َ which represent the following short vowels /a/, /u/, /i/.
- **Consonants** carry ◌ْ.
- **Shadaa** indicates the letter is doubled ◌ّ.
- **Tanween** *harakah* and *Noon* letter with consonant to the end of the word. It sounds /n/.

Arabic Prosody العَرُوض

وَمَنْ دَعَا النَّاسَ إِلَى ذَمِّهِ			ذَمُّوهُ بِالْحَقِّ وبِالْبَاطِلِ		
وَمَنْ دَعَ نَـ	نَاسَ إِلَى	ذَمْمِهِي	ذَمُّوهُ بِلْـ	حَقْقِ وبِلْـ	بَاطِلِي
0//0//	0////0/	0//0/	0//0/0/	0////0/	0//0/
مُتَفْعِلُنْ	مُسْتَعِلُنْ	مَفْعُلا	مُسْتَفْعِلُنْ	مُسْتَعِلُنْ	مَفْعُلا

- A **foot**(*tafa'ilah* التفعيلة) : is an **ordered** sequence of vowels and consonants.
- **Meter** البحر: is an **ordered** sequence of feet.

Meter Name	Meter *feet combination*
al-Wafeer	مُفَاعَلَتُنْ مُفَاعَلَتُنْ فَعُولُنْ
al-Taweel	فَعُولُنْ مَفَاعِيلُنْ فَعُولُنْ مَفَاعِلُنْ
⋮	⋮
al-Moktadib	مَفْعُولَاتُ مُسْتَفْعِلُنْ مُسْتَفْعِلُنْ
al-Modar'e	مَفَاعِيلُنْ فَاعِلَاتُنْ مَفَاعِيلُنْ

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///

Thesis Working Steps.

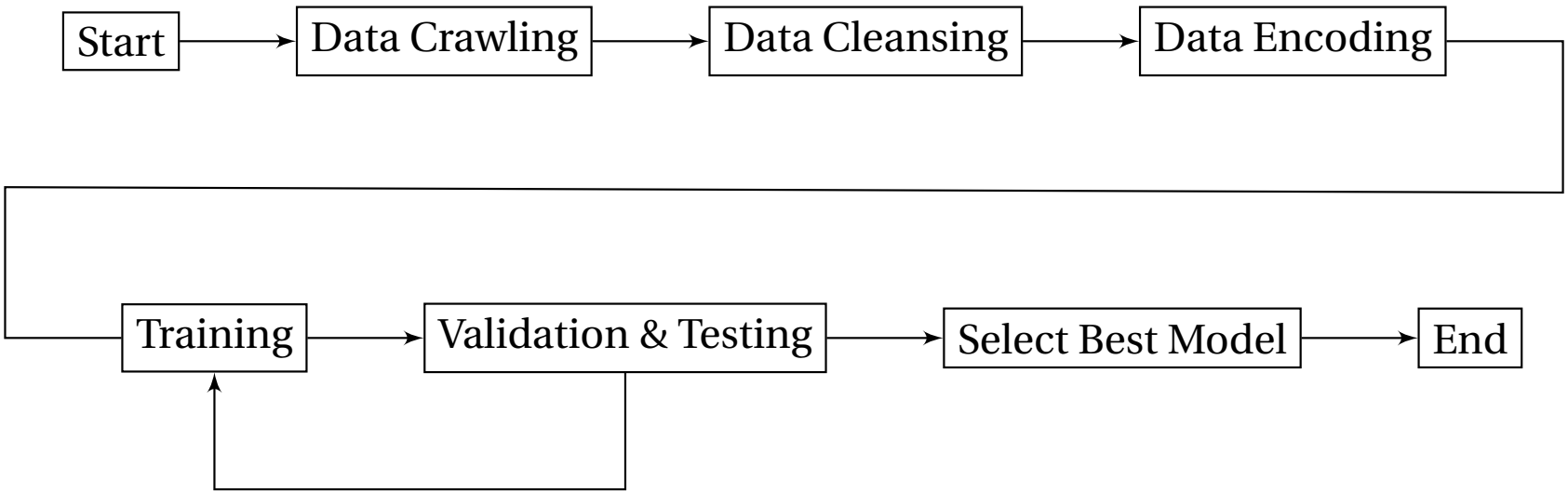


Figure: Thesis Working Steps.

Literature Review

Detecting Arabic poems meters

Deterministic Approach

There is some literature on recognizing the meters of written Arabic poem using rule-based deterministic algorithms

Machine Learning Approach

Learning and classifying poems to the right meter has not been addressed before!

Deterministic Approach

- **Abuata and Al-Omari [1]:**

Deterministic Approach

- **Abuata and Al-Omari [1]:**
  - Deterministic 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.

Deterministic Approach

- **Abuata and Al-Omari [1]:**
  - Deterministic 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.

Deterministic Approach

- **Abuata and Al-Omari [1]:**
  - Deterministic 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 [2]**, similar approach; Context-Free Grammar; 75% correctly classed from 128.

Our point of departure

Issues;

- Dataset size.
- Accuracies: (75%, 82%) tested on (128, 417) verses respectively.
- Diacritics are a must.
- Converting verses into Al-Arud writing style is probabilistic.
- For detecting meters, all models are so **naive and primitive**. They do no have any clue about the real pattern.

Our Approach

- We present:
  1. Machine Learning approach.

Our Approach

- We present:
  1. Machine Learning approach.
  2. Big Dataset.

Our Approach

- We present:
  1. Machine Learning approach.
  2. Big Dataset.
  3. New encoding technique.

Our Approach

- We present:
  1. Machine Learning approach.
  2. Big Dataset.
  3. New encoding technique.
  4. Results.

Datasets Design

Dataset

Our dataset consists of **1,722,321** labeled data points.

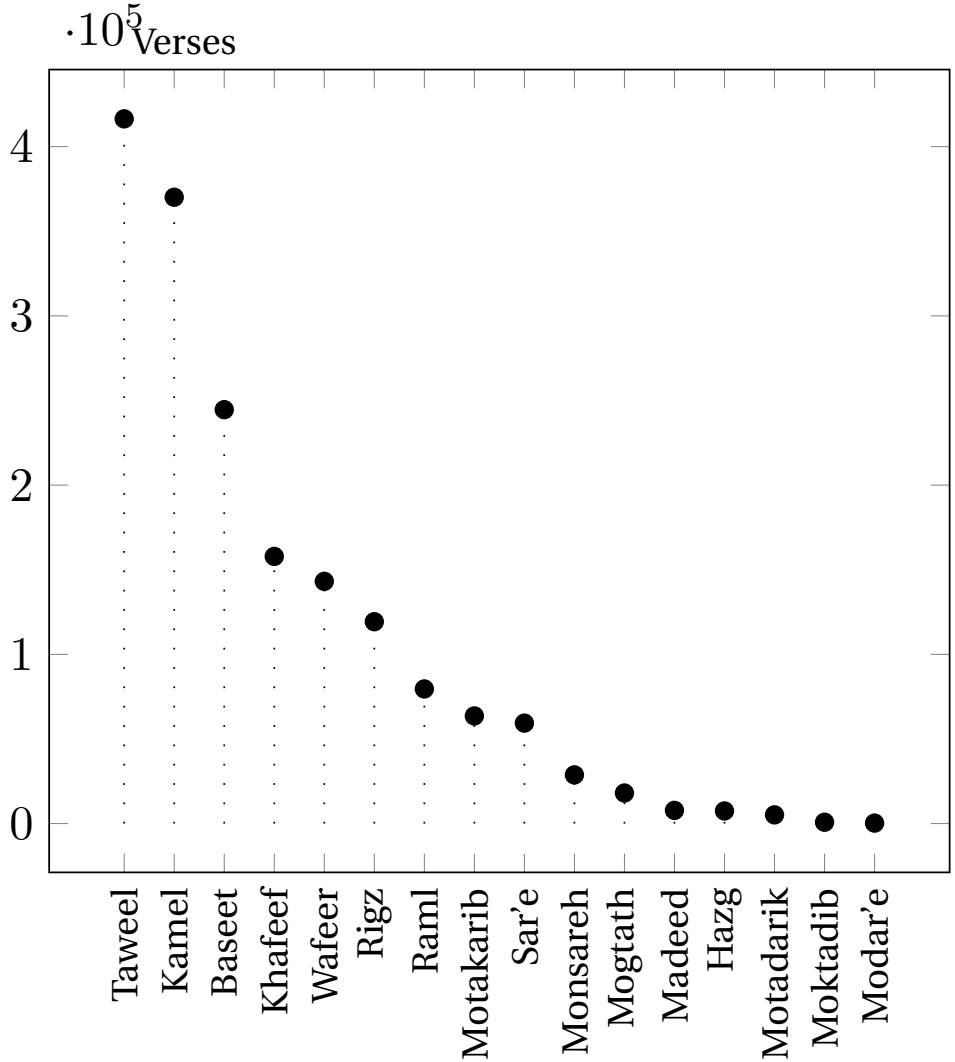


Figure: Number of verses per *meter* ordered descendingly on y-axis vs. meter name on x-axis.

Dataset

Basic cleansing rules:

- Filtering the 16 classic meters.
- Removing many unnecessary white spaces.
- Removing non-Arabic characters.
- Factoring Shadaa and Tanween.

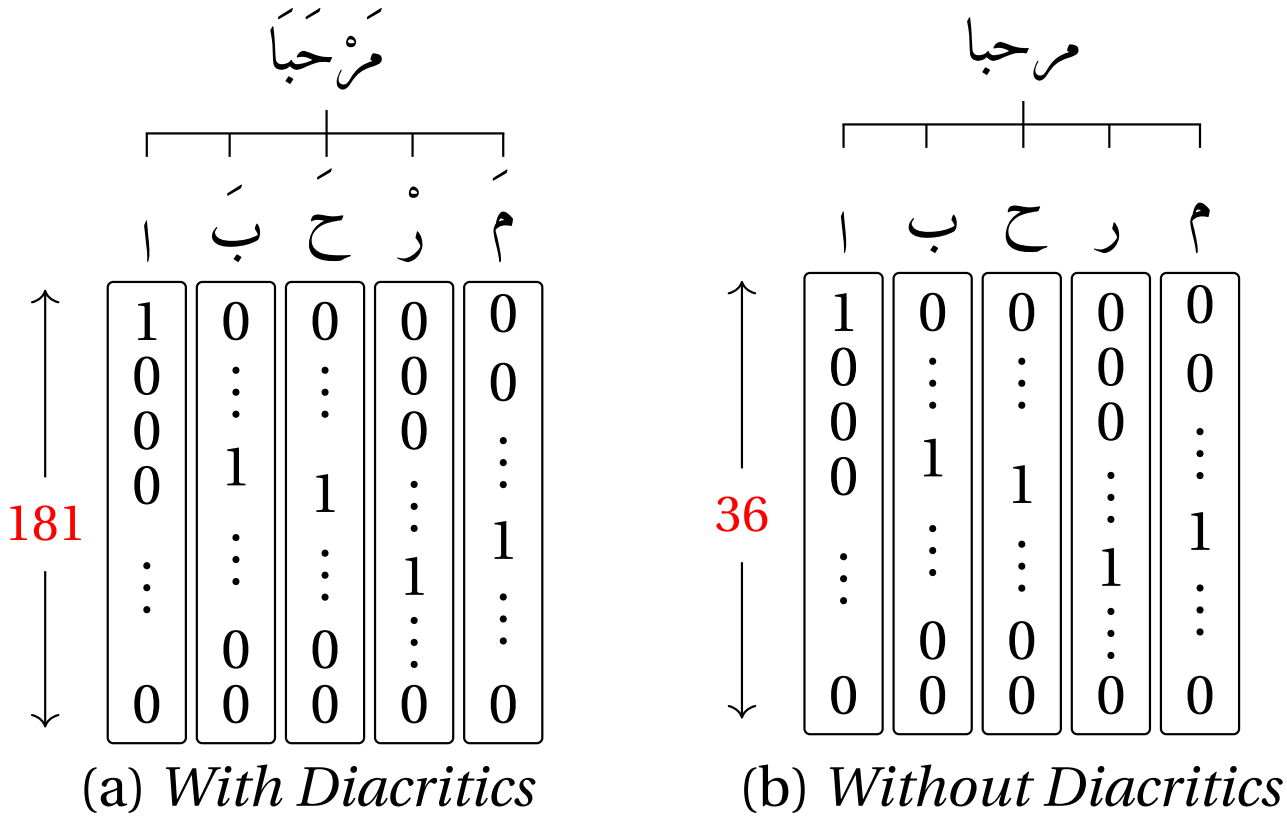
Diacritics	*With Shadda*	*Without shadaa*	*With tanween*	*Without tanween*
Shape	دَّ	دْدَ	دً	دَ+نْ

Table: Diacritics on the letter د

Data Representation

An Issue:

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



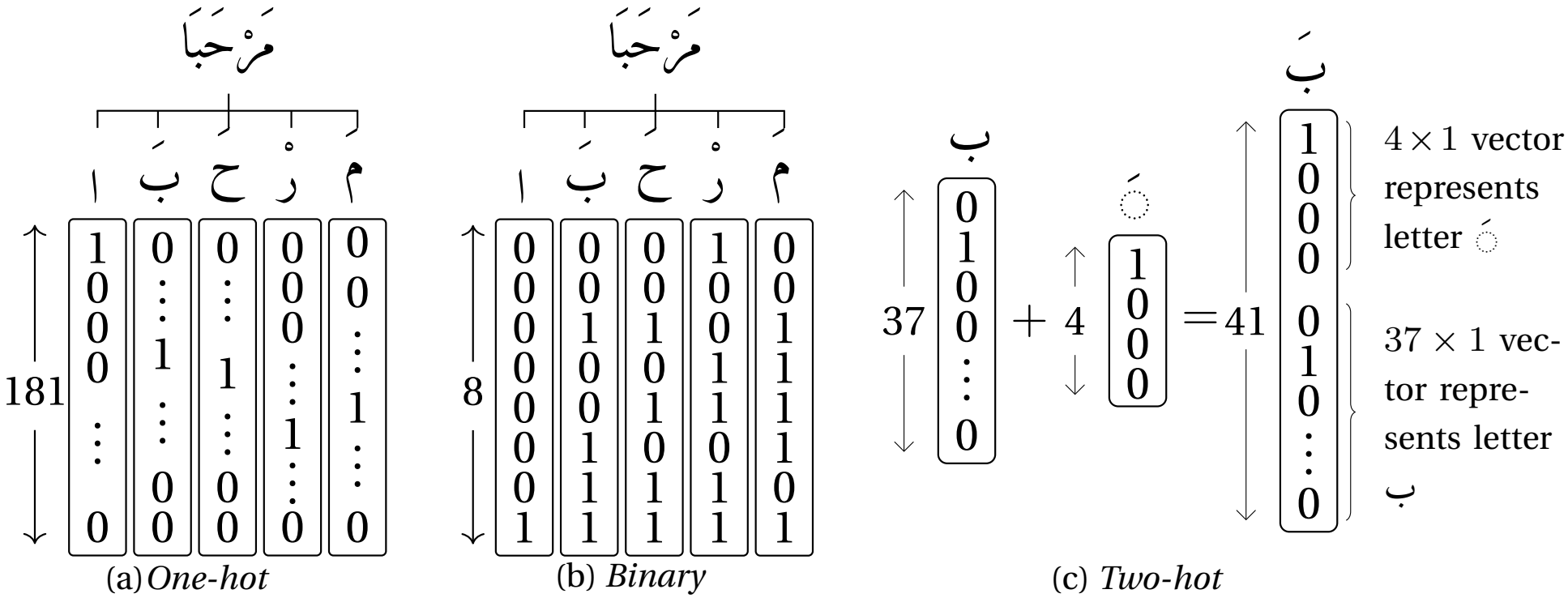
(a) *With Diacritics* (b) *Without Diacritics*

Figure: Example of *One-hot* encoding for Arabic word (مرحبا).

Encoding Techniques

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

Encoding Techniques



(a) *One-hot* (b) *Binary* (c) *Two-hot*

Figure: Different encoding mechanisms

Training and Architecture

Which Network!

- **Pattern**: is a sequence of characters.
- Unlike feedforward neural networks, RNNs can use their internal state (memory) to process sequences of inputs.
- In theory, RNNs are capable of handling long-term dependencies. However, in practice they do not, due to the exploding gradient problem
- LSTMs was designed to solve the long-term dependency problem using internal memory gates.

RNN, Architectures

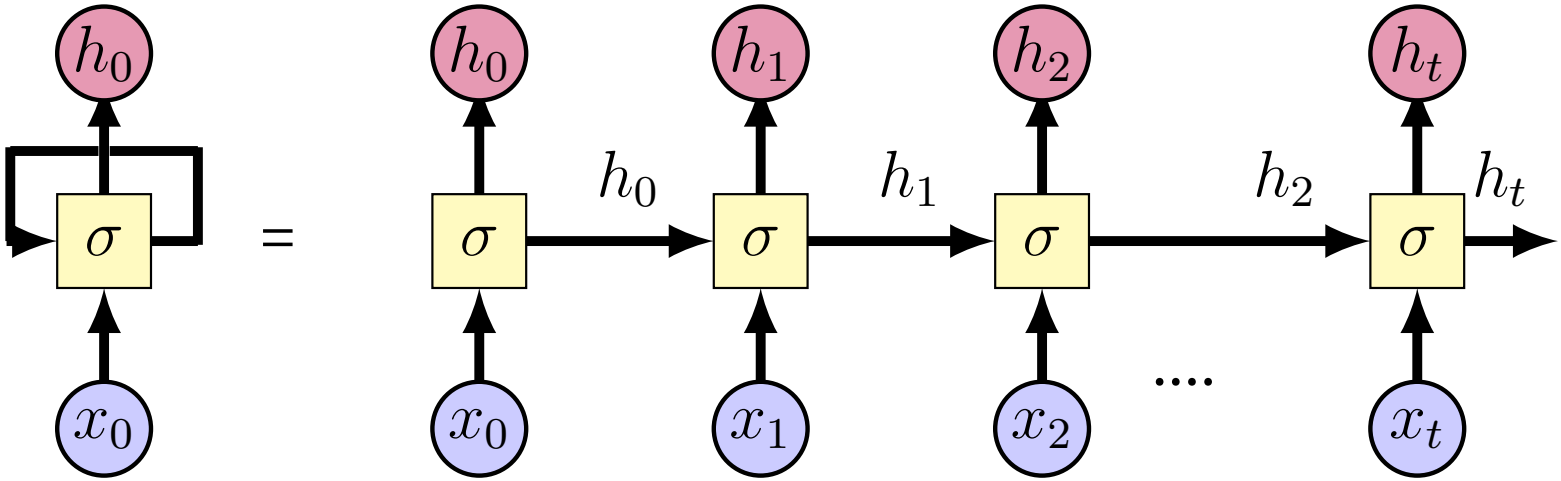


Figure: Recurrent Neural Networks Loops adapted from [3]

RNN, Architectures

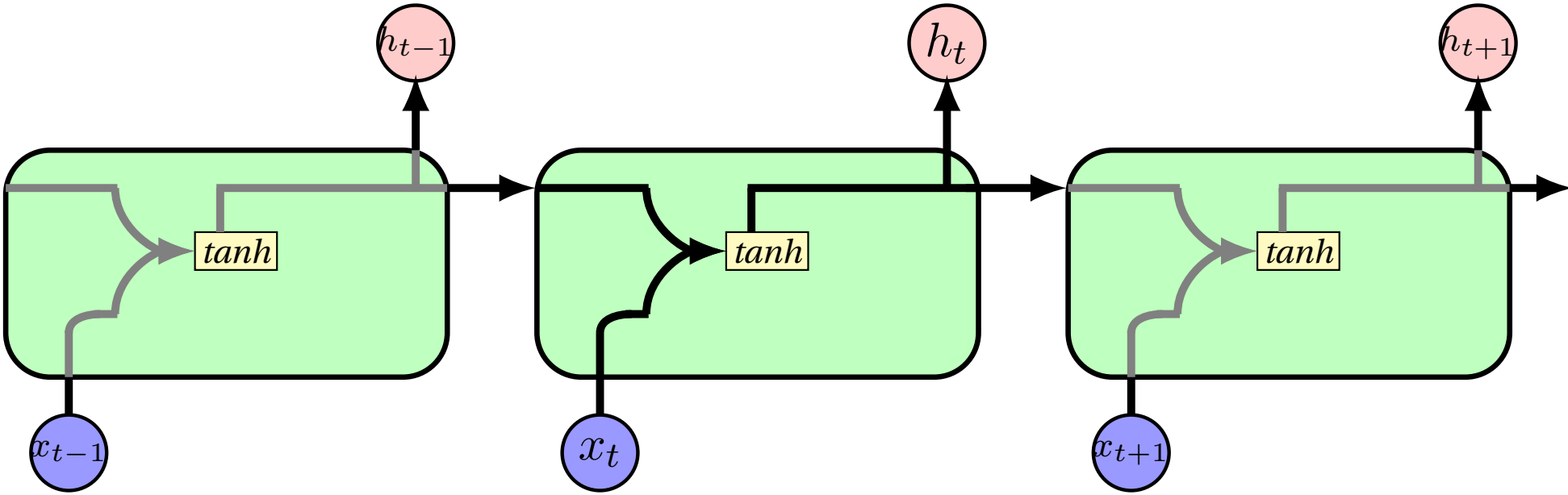


Figure: A single recurrent layer adapted from [3]

LSTM Architectures

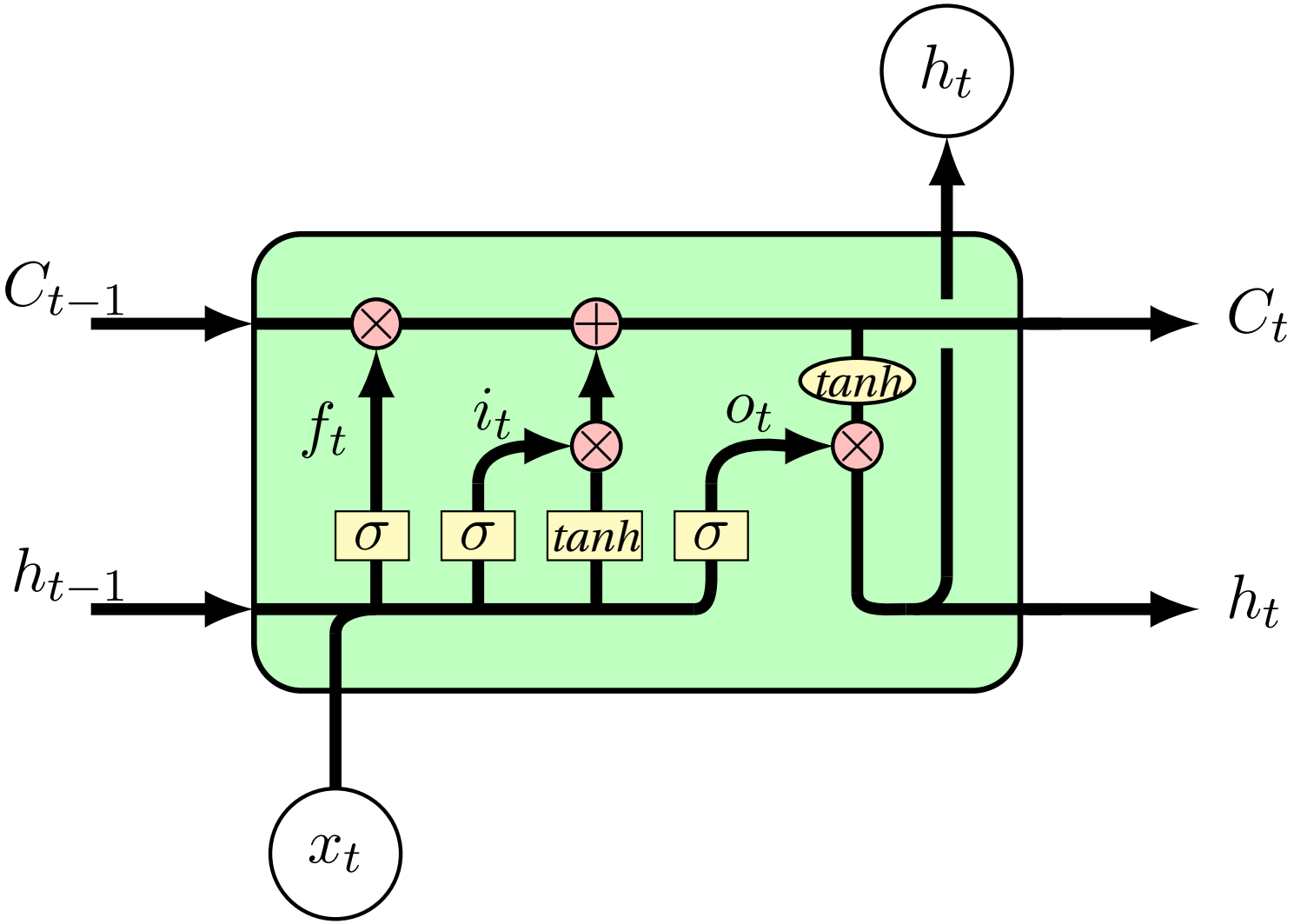


Figure: LSTM internal cell adapted from [3]

LSTM Architectures

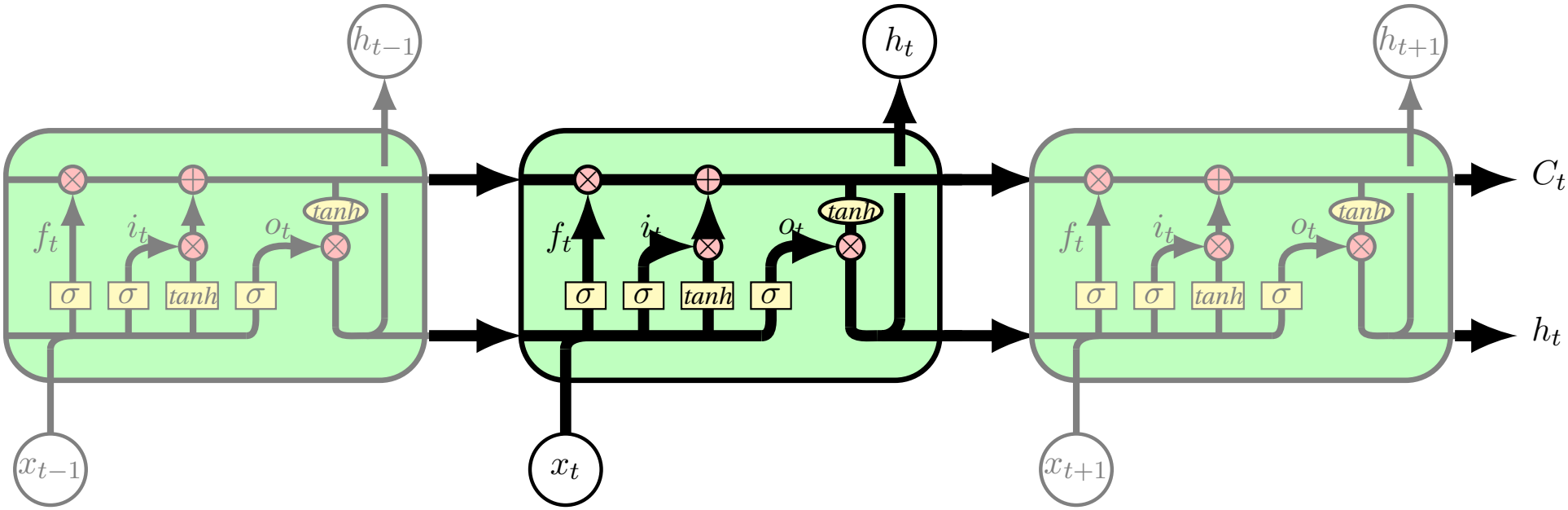


Figure: Unfold LSTM adapted from [3]

LSTM Architectures

Bi-LSTM Motivation

- ***Harry*** is the king, and he will travel next week.
- The new book which makes the big sale is named ***Harry*** Potter.

- Bi-LSTM models always outperform LSTM models.
- It means that models can't learn the pattern from one direction, it should be two directions together.

LSTM Architectures

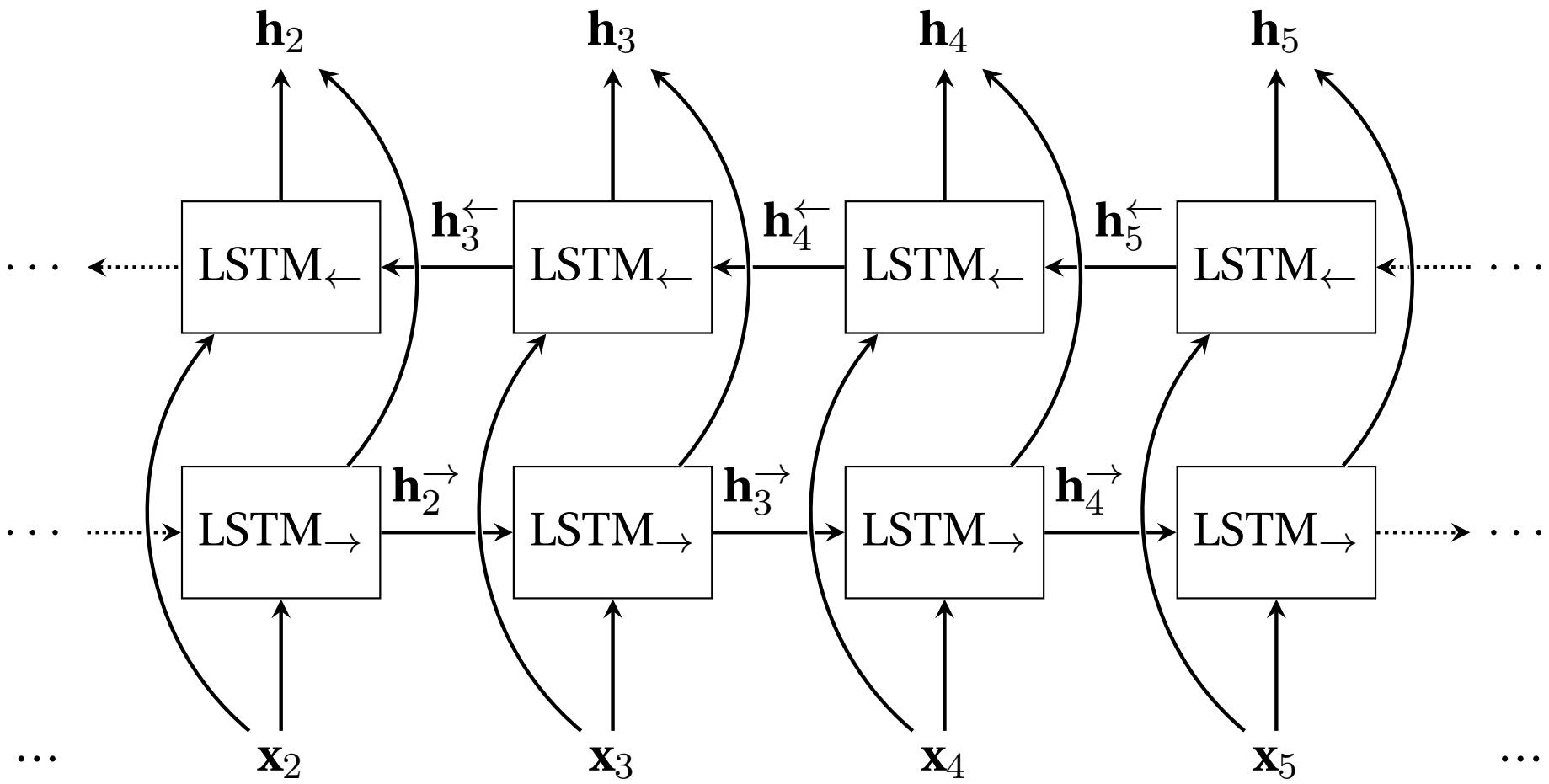


Figure: bidirectional long short-term memory [4]

Experiments and Results

Experiments Parameters

- **Dataset Configurations**:
  - Encoding technique: BinE, OneE, TwoE.
  - Diacritics: 0D, 1D.
  - Trimming: 0T, 1T.
- **Network Configurations**:
  - Loss functions: *Weighted* or *Non-Weighted* **(1, 0)** respectively.
  - The number of layers: nL.
  - The number of cell units: nU.
  - Cell type: LSTM, Bi-LSTM.

Overall Accuracy!

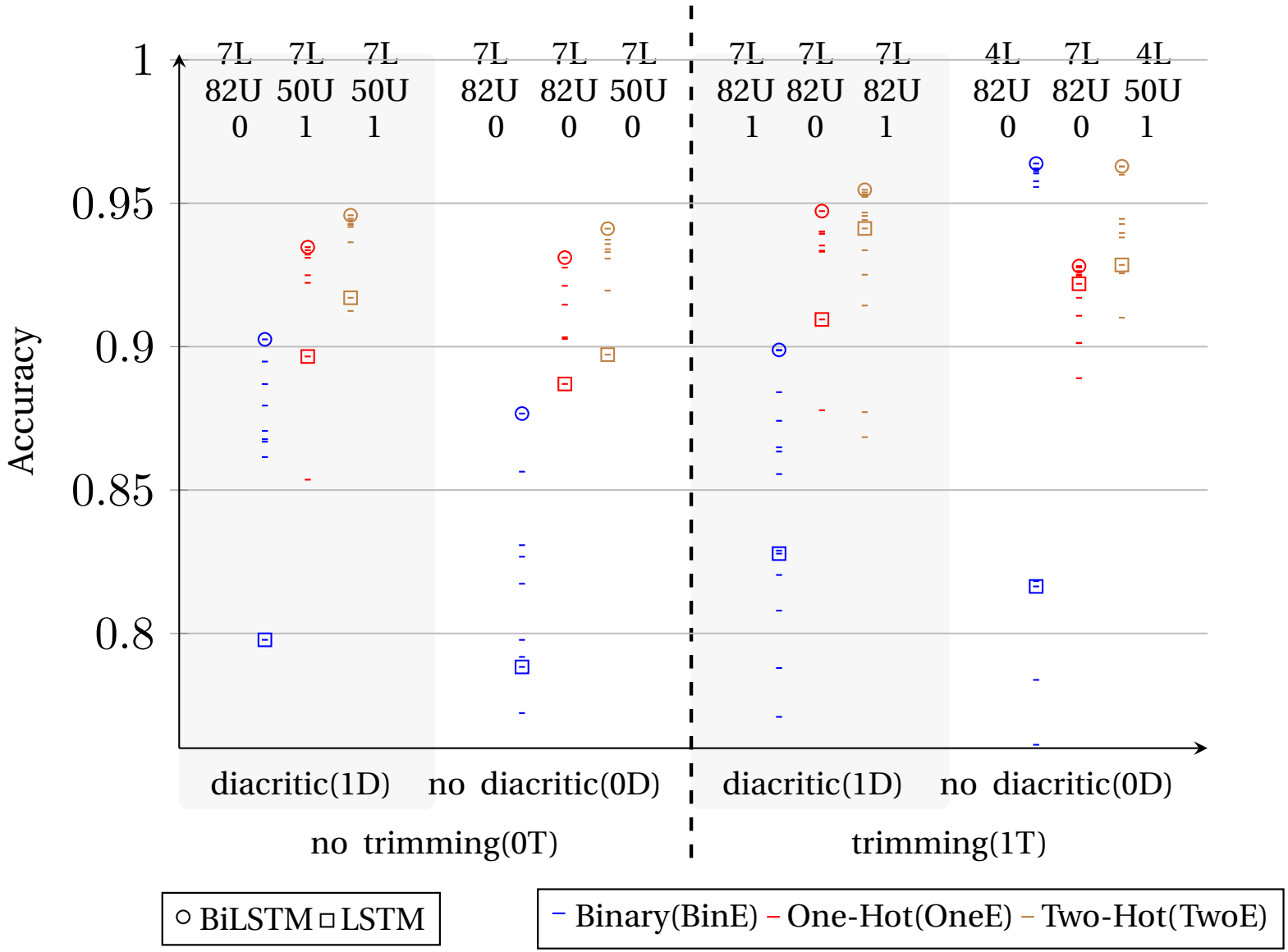


Figure: Overall accuracy of the 192 experiments

Comparison with related works

Ref.	Accuracy	Test Size
[2]	75%	128
[1]	82.2%	417
This article	96.38%	150,000

Table: Overall accuracy of this article compared to literature.

Per-class Accuracy!

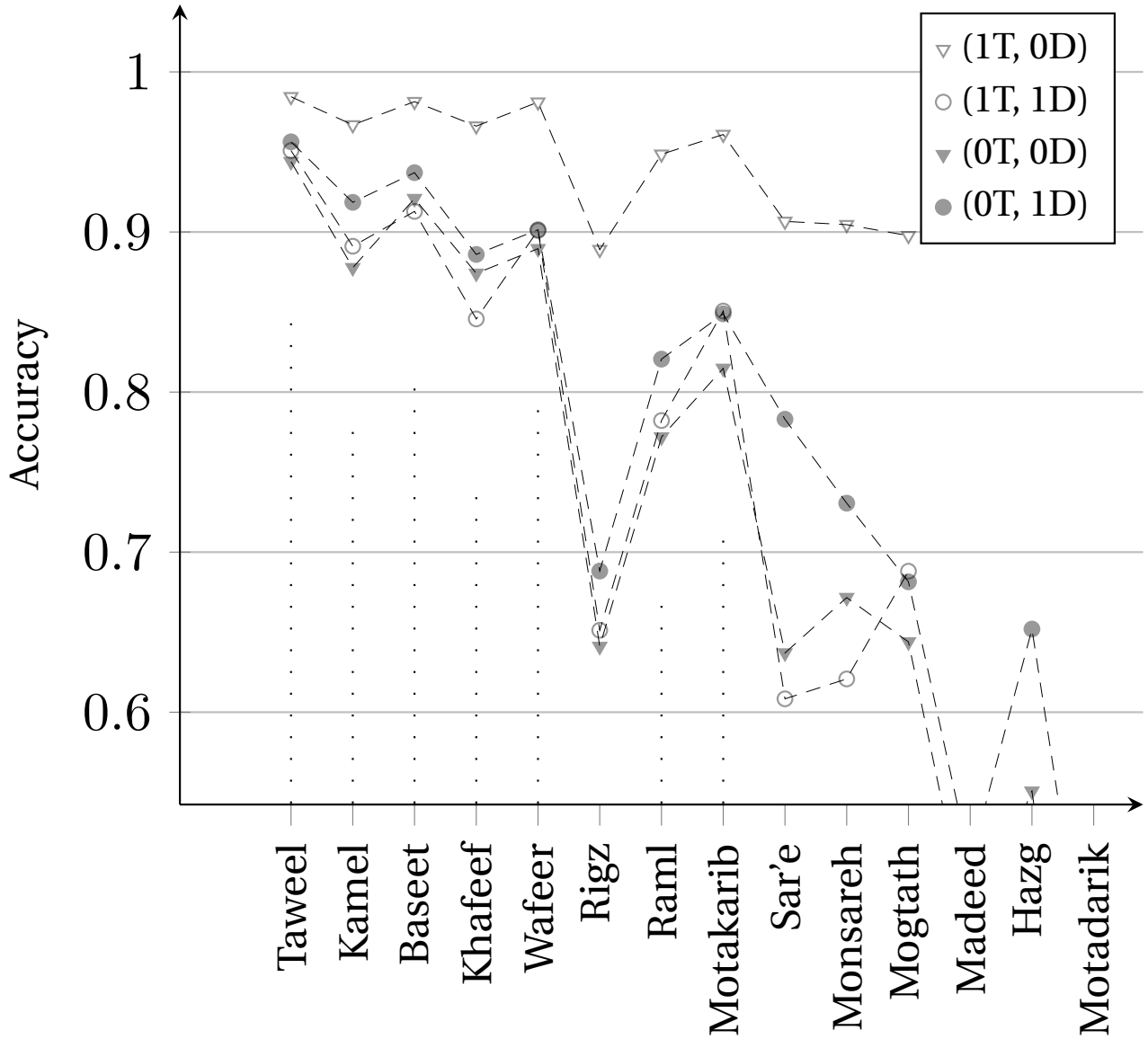


Figure: The per-class accuracy score of the best four models.

Discussions

Encoding effect

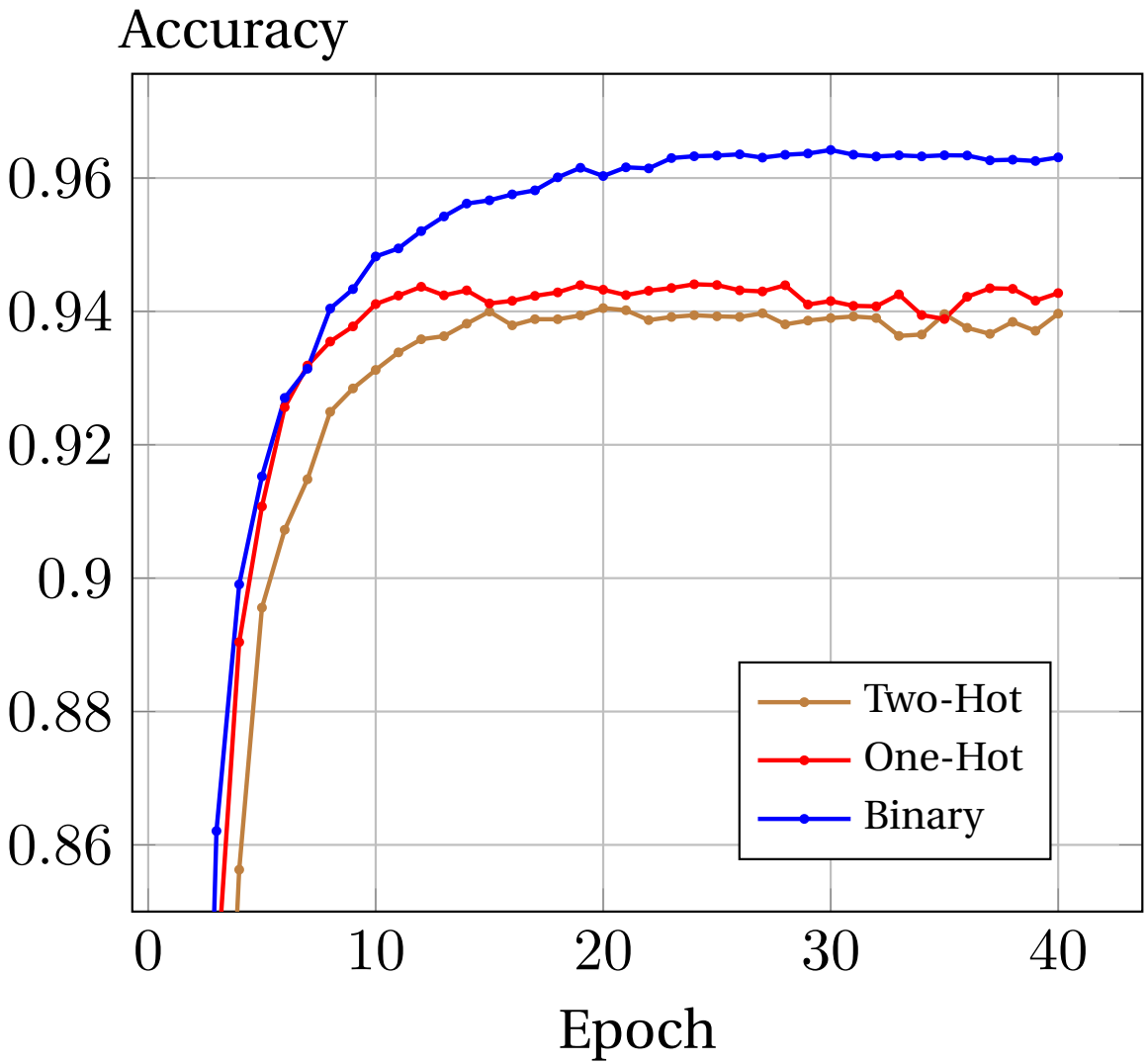


Figure: Encoding effect on Learning rate with the best model (1T, 0D, 4L, 82U, 0W, BinE) and when using the two other encodings instead of BinE.

Encoding effect

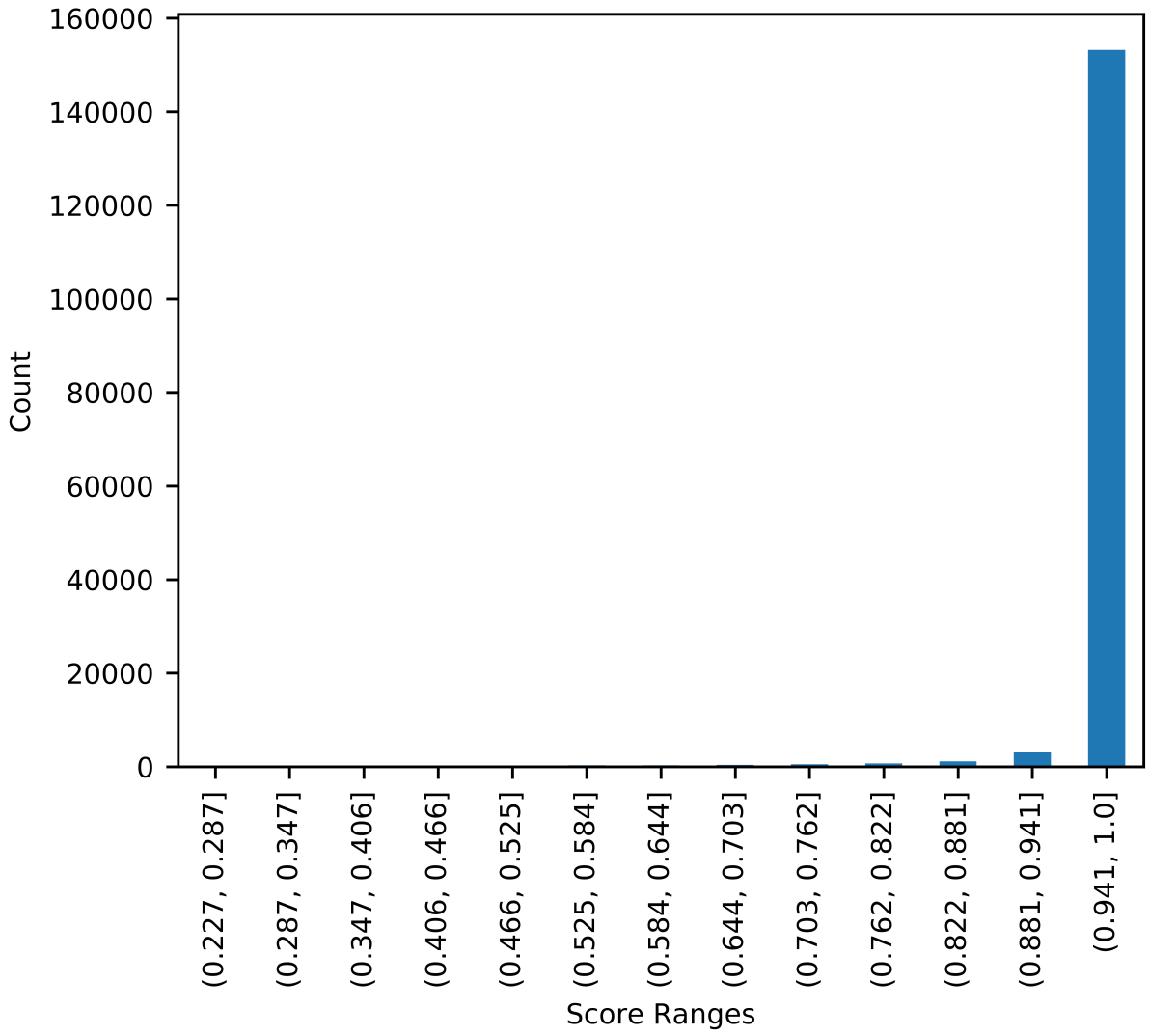
Encoding

- The encoding method is a transformer function $\mathcal{T}$ which transform a discrete input values X.
- If the network η_1 is the most accurate network which can "decode" $\mathcal{T}(X)$.
- If we have another encoding function $\mathcal{T}_2$ and we tried to use the same network η_1 for the $\mathcal{T}_2$ as $\eta_1\left(\mathcal{T}_1(X)\right) = \left(\eta_1 \cdot \mathcal{T}_1 \cdot \mathcal{T}_2^{-1}\right)\left(\mathcal{T}_2(X)\right)$. This network may be of complicated architecture to be able to "decode" a terse or complex pattern $\mathcal{T}_2(X)$.

Classifying Arabic Non-Poem Text

خلال	المباراة	التي	جمعتهما	مساء السبت بالجولة	الـ26 من المسابقة
خلالَلْ	مباراتل	لتيج	معتهما	مساءَسْ سبتبلجو	لتلمنل مسابقه
0/0//	0/0/0//	/0//	0//0//	0/0/// 0/0/0//	0//0// 0//0//
0/0//	0/0/0//	/0//	0/0///	0/0/// 0/0/0//	0/0// 0//0//
فَعُولُنْ	مَفَاعِيلُنْ	فَعُولُ	مَفَاعِلُنْ	فَعُولُنْ مَفَاعِيلُنْ	فَعُولُنْ مَفَاعِيلُنْ

Classifying Arabic Non-Poem Text



;

Figure: Testing data score ranges distribution.


Abuata, Belal and Al-Omari, Asma
A Rule-Based Algorithm for the Detection of Arud Meter in Classical Arabic Poetry
International Arab Journal of Information Technology. (2017), 15.


Alnagdawi, Mohammad and Rashaideh, Hasan and Aburumman, Ala
Finding Arabic Poem Meter Using Context Free Grammar
J. of Commun. & Comput. Eng. (2013), 3, 52-59.


Colah
Understanding Lstm Networks
http://colah.github.io/posts/2015-08-Understanding-LSTMs/ , 2015.


Petar Veličković
Collection of Latex Tikz figures
https://github.com/PetarV-/TikZ.

Questions!

Questions.