Translating Old Persian cuneiform by artificial intelligence (AI)

Author: Shaghayegh Rahmani (Melanee)
Email: melaneepython@gmail.com
GitHub: https://github.com/Electronic-Old-Persian-Library
Iran, Tehran
24 July 2024

Keywords
Old Persian, translating, cuneiform, artificial intelligence, OCR, NLP

Abstract
In this project, I developed an optical character recognition (OCR) model to decipher Old Persian language to modern languages. The process of converting an image text to machine-readable text is called OCR. To the best of my knowledge, I am the first one who is developing this language in this scale in the world. The primary raw data is collected from all over the world. For instance, the British museum collection, cuneiform digital library Initiative (CDLI), Livius and my personal photography from the national museum of Iran and Takht-e-Jamshid (Persepolis). I have trained my data with some OCR models, the best one is based on the EasyOCR model.

Introduction

Old Persian cuneiform is a semi-alphabetic cuneiform script that was the primary script for the Old Persian language. Texts written in this cuneiform were found in Persepolis, Susa, Hamadan, Armenia, and along the Suez Canal (Kent 1950).They were mostly inscriptions from the time period of Darius the Great and his son Xerxes the Great kings of Achaemenid Empire (Kent 1950).

The Achaemenid Persian empire was the largest that the ancient world had seen, extending from Anatolia and Egypt across western Asia to northern India and Central Asia. Its formation began in 550 B.C (The Achaemenid Persian Empire 2000).

Being able to read and decipher ancient scripts has always been of much interest and importance for human-being. Many valuable historical secrets were revealed by

archaeologists through the time, which made us aware of the culture and civilization of our ancestors.
Saving the history of my country, Iran, is a very important and valuable issue to me. Sometimes I think if there is no one in this world who can translate the Old Persian language, what will happen to this language, how will the next generation decipher new found inscriptions or tablets, therefore, this new brainstorm crosses my mind to develop a new AI model to achieve this goal and keep this ancient language alive forever.

Previous research

In June 2024, I joined Professor Enrique Jiménez team. He and his team at Ludwig Maximilian University (LMU) was utilising advanced artificial intelligence (AI) and computer vision techniques to analyse and decipher Babylonian and Akkadian inscriptions, including the Epic of Gilgamesh. They have developed several tools and algorithms to facilitate this process, which are available in their GitHub repositories and electronic babylonian library (eBL). My project is inspired by the eBL project. Professor Enrique Jiménezhas has developed models for Babylonian cuneiform but I am going to develop my models for Old Persian cuneiform.

Developing models

I have developed three OCR models in this project:
- yolo_cnn_old_persian
- tessearct_old_persian
- easyocr_old_persian

I have developed an “easyocr_old_persian” model from scratch and trained it with 42 images of the last 12 lines of the great Darius's inscription in Persepolis, DPd inscription as primary data. This model deciphers Old Persian cuneiform to English transcription and is published on the Old-Persian-Cuneiform-OCR repository on Github. This repository is a part of Electronic Old Persian Library (EOPL) organisation. Since I was looking for an OCR model for the Old Persian language, I have not implemented image pre-processing for my models yet and they work on just black and white images. Additionally, my models are still under developing. In the future, I will train the “easyocr_old_persian” model with more real data to get better results.

The “tessearct_old_persian” model is a pre-trained model developed by S. Muhammad Hossein Mousavi. I just developed a code for evaluating this pre-trained model in Python programming language.

.

The “yolo_cnn_old_persian” model is not completed yet.

Results and discussion

Evaluating “tessearct_old_persian” model code, I did input image 1 to this model:

```
!pip install pytesseract
!apt-get install tesseract-ocr
!pip install pillow
%cd /usr/share/tesseract-ocr/4.00/tessdata/
!mv/content/drive/MyDrive/Persiancuneiform1/tesseract/myLang.traineddata
/usr/share/tesseract-ocr/4.00/tessdata

from PIL import Image
import pytesseract

pytesseract.pytesseract.tesseract_cmd = r'/usr/bin/tesseract'

img_path ='/content/drive/MyDrive/Persiancuneiform1/DPd.png'
img = Image.open(img_path)
text = pytesseract.image_to_string(img, lang='myLang')
print(text)
```

Input:

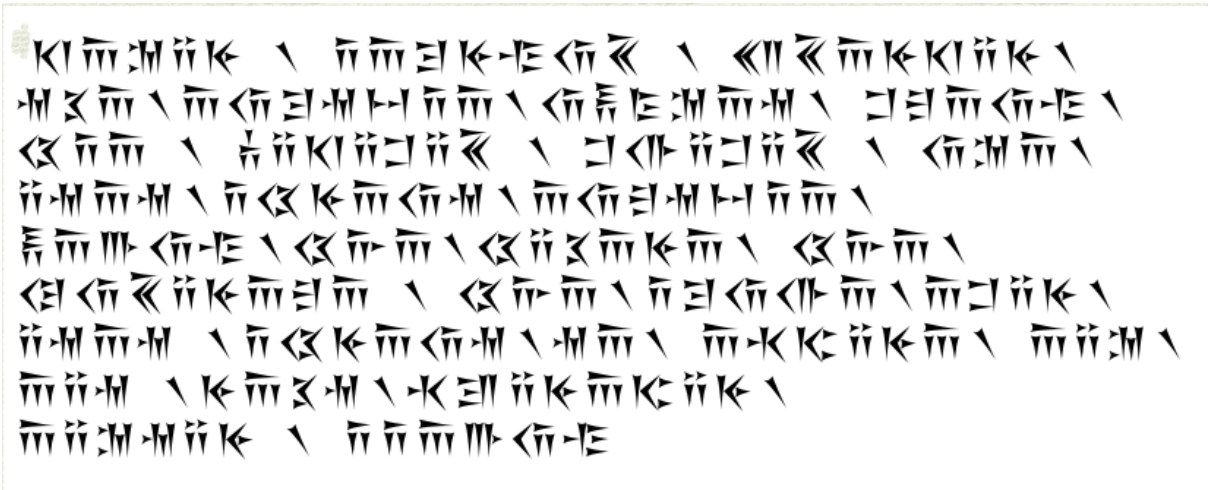

Image 1. Input cuneiform to “tessearct_old_persian” model

Output:

Zittiy ; iaryvuS ; xrSayZiy;

mnc;aurmzia;upstam; rlauv;

hia ; ViZiriS ; rgiriS ; uta;

im am ; i h yaum ; au lm z i a ;

pitTucs;hca;hinaya; hca;

QuSiyala ; hca;iruga;ariy;

imam ;ihyaum;ma; ajMiya; ait;

aim ;yanm;jDiyaMiy;

aitmiy ; iiaTuv

In the next step. I translate that Old Persian transcription to modern languages by Chat-GPT (GPT-4 version):

Translate to Modern Persian:

"این منم داریوش شاهنشاه؛ به لطف اهورامزدا، من این را بنا کردم؛ من این امپراتوری را بنیان نهادم و آن را نیرومند ساختم. باشد که اهورامزدا من و پادشاهی مرا محافظت کند؛ باشد که برای همیشه پایدار بماند؛ و باشد که از دروغ در امان باشد؛ این است آنچه من انجام دادم؛
این است آنچه من می‌گویم."

Translate to Modern English:

"This is me, Dariush king; By the grace of Ahura Mazda, I have built this; I founded this empire and made it strong. May Ahuramazda protect me and my kingdom; may it last forever; and it would be safe from lies; that is what I did;
That is what I am saying."

But in the future, I will develop my own natural language processing (NLP) model to translate the output instead of using ChatGPT.

Results for “easyocr_old_persian” model are show in table 1 and image 2:

Step	num_iter	valInterval	Confidence score	Execution time(min)
1	300	200	0.0032	5
2	3000	2000	0.5	15
3	10000	50000	0.45	102

Table 1. Training results for “easyocr_old_persian” model

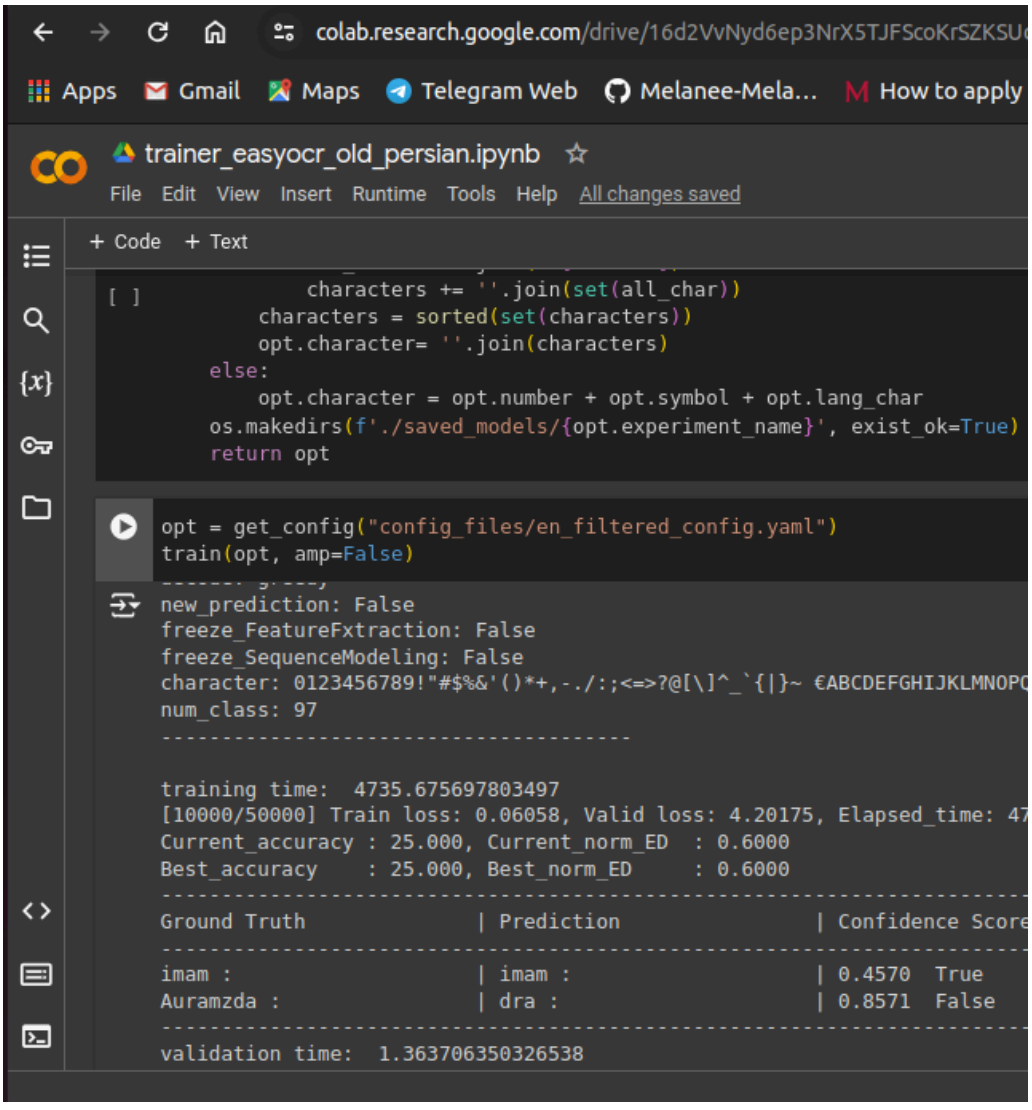

Image2. Training results for “easyocr_old_persian” model in my notebook

All my models have been developed on Google Colab notebook infrastructure (table 2):

Operating system	version	GPU	CPU	RAM
Linux, Ubuntu	22.04 LTS	T4GPU	-	16G

Table 2. Infrastructure of my Google Colab notebook

Future Research

- This “easyocr_old_persian” model can be implemented for other languages, for example, Akkadian, Pahlavi, Avestan, Sogdian and so on.

- By deciphering Old Persian cuneiform in trilingual inscriptions (like image 3) or tablets, we can decipher Babylonian and Elamite cuneiform because they have same meaning.

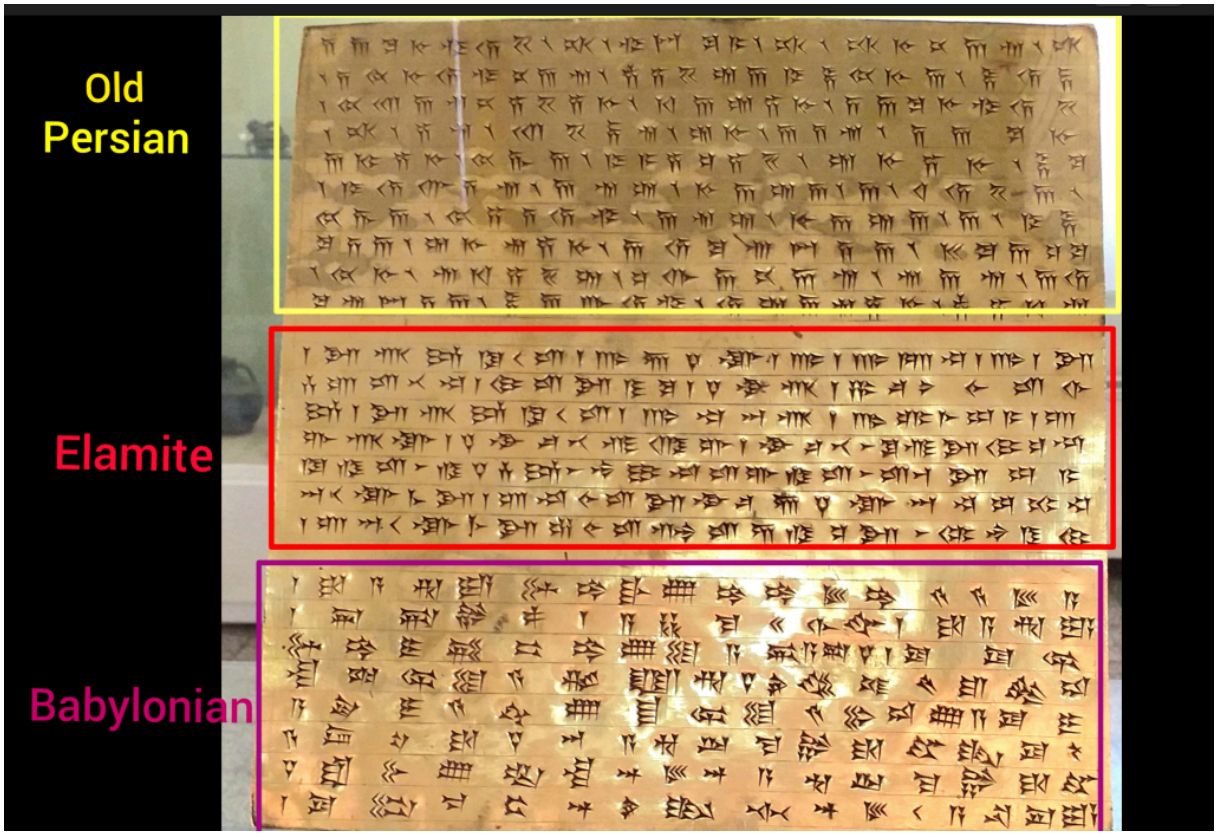

Image 3. Trilingual gold plate of king Darius, photo is taken from national museum of Iran

- Matching broken tablets or inscriptions, we can use OCR to convert the cuneiform signs from images into machine-readable text. Then we will use Prof Enrique Jiménez’s NLP project to apply algorithms to detect and match segments of different tablets or inscriptions, aiding in the reconstruction of fragmented texts (image 4).

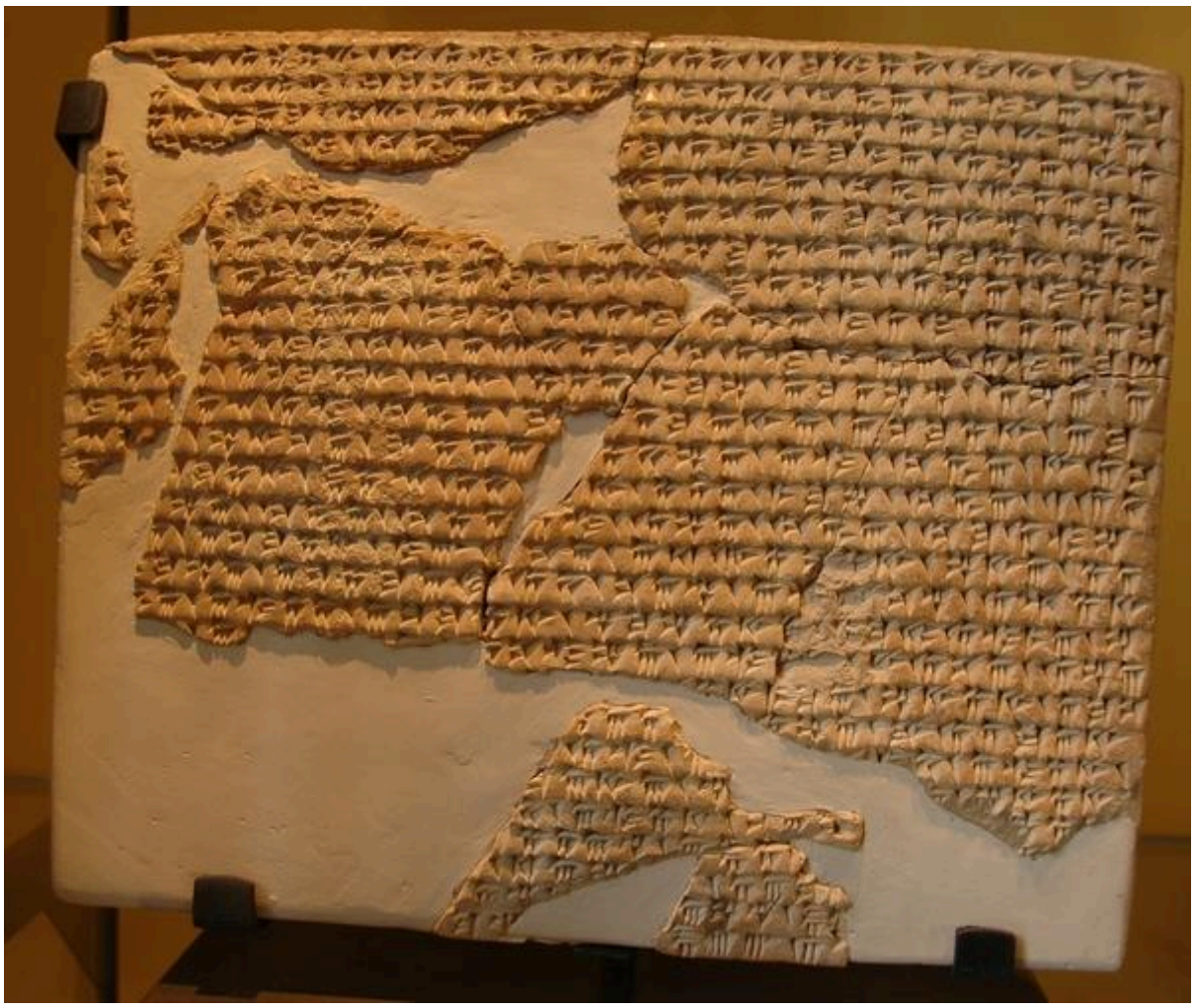

Image 4. Photo is from Apadana Castle Shush, DSf inscription

Future work

- Developing image pre-processing code for the raw image dataset for "yolo_cnn_old_persian" model to convert real images to black and white images and omit noises from background (image 5)

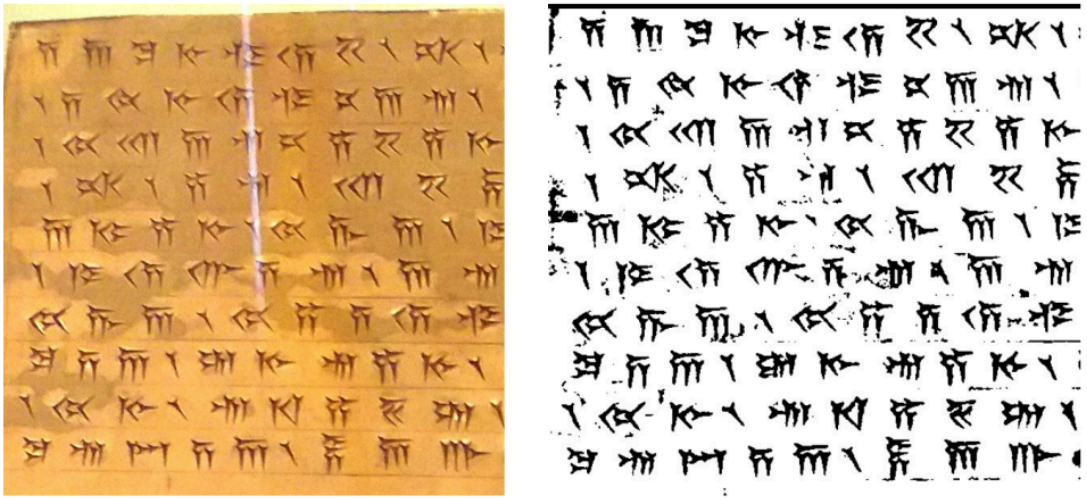

Image 5. Image pre-processing

- Training “easyocr_old_persian” model with huge images
- Developing NLP models to translate English transcriptions to modern Persian

Conclusion

Acquired results of the evaluation indicate that my models will be able to properly translate Old Persian cuneiform. The acquired results are promising that they are able to make and improve NLP in this area.

References

Gutherz, G., Gordin, S., Sáenz, L., Levy, O., and Berant, J. (2023). Translating Akkadian to English With neural machine translation. PNAS Nexus 2. https://doi.org/10.1093/PNASNEXUS/PGAD096

Kent, R., 1950. Old Persian Grammar Texts Lexicon, American Oriental Society, New Haven.

Mostofi, Fahimeh, and Adnan Khashman (2015). Intelligent Recognition of Ancient Persian Cuneiform Characters. https://doi.org/10.5220/0005035401190123

Mousavi, Seyed Muhammad Hossein and Vyacheslav Lyashenko (2017). "Extracting old persian cuneiform font out of noisy images (handwritten or inscription)". In: 2017 10th Iranian Conference On Machine Vision and Image Processing (MVIP). IEEE, 241–246 https://doi.org/10.1109/IranianMVIP.2017.8342358