Optical character recognition for ancient non-alphabetic scripts

openaccessgovernment.org

30 September 2022

Shai Gordin, Senior Lecturer at Digital Pasts Lab in Ariel University, looks at the deciphering of ancient non-alphabetic scripts, and the technology we use to understand it

How do we get from an ancient cuneiform tablet with non-alphabetic scripts, written thousands of years ago, to a digital representation of its text? Thus, making it available for further computational analysis like quantitative methods or natural language processing (NLP) models.

Cuneiform is one of the earliest writing systems in the world, invented at the end of the fourth millennium BCE. It is usually written by pressing a stylus on moist clay tablets, creating a three-dimensional script. The script is logo-syllabic, like the Chinese or Japanese writing systems, meaning the same sign can be read logographically, as a word, as syllables, or as determinatives (ie semantic classifiers). The correct reading depends on the context. There are close to a thousand cuneiform signs, not all of which were used simultaneously; usually about 200-300 signs were used at once.

Optical character recognition (OCR)

Optical character recognition (OCR) and NLP methods for such non-alphabetic scripts and writing systems are still complex and error-prone. Practically, any attempt at computational analysis that is not for the English language, employs a script other than the Latin script or requires a writing direction that is not left-to-right will not be able to use most out-of-the- box implementations and tools.

Non-alphabetic scripts present additional difficulties for digitisation and analysis, particularly those whose origins can be traced back hundreds, if not thousands, of years. Texts can be written on different types of materials, such as stone, clay or metal. Historical documents can be eroded, partially broken or even completely erased.

How do experts read a cuneiform tablet? First, they need access to the original tablet or a good-quality image on which the text is inscribed. Then, the expert deciphers the signs. This requires knowledge of the cuneiform writing system, its various signs for the period at hand and the geographical area from which the tablet comes, as well as the different readings of the signs.

The expert also needs good knowledge of the language in which the texts are written, like Akkadian or Sumerian. Furthermore, knowledge of similar or equivalent texts, and an understanding of the culture from which the text arose, is often needed to validate the

best decipherment of the text at hand.

AI and machine learning models

For the OCR to be helpful for experts and laypeople alike, all these tasks need to be tackled computationally. However, the process of training an AI model cannot fully imitate the scholarly process of reading a text. On the other hand, machine learning models can often detect useful patterns overlooked by humans.

In recent years, difficult OCR tasks, like nearly all computer vision problems, have been solved with deep learning models. These models considerably improved performance in such a way that OCR for common languages such as English can be reasonably considered solved.

OCR for cuneiform tablets is more challenging for two main reasons: First, cuneiform writing and all the languages using it are low-resource languages, meaning we have a very small amount of labelled data for these languages. Second, cuneiform signs overlap to a much greater degree than standard print text.

3D models of cuneiform tablets

Attempts to perform OCR on cuneiform texts have tackled these complex issues from various directions. Several research groups in Germany have developed programs for manipulating 3D models of cuneiform tablets, including stroke extraction, and joining broken tablet fragments.

Unfortunately, such 3D scans are not numerous and 2D images of cuneiform tablets are far more ubiquitous and are available in higher quality. Therefore, these scholars have also turned to machine learning models to do OCR from 2D tablet images. The best results were achieved in sign detection on Neo-Assyrian tablet images.

My research team working in the Babylonian Engine project created a set of CuRe (Cuneiform Recognition) tools, an online interactive platform for scholars. The machine learning models are envisioned as “co-workers”, which provide likely suggestions to the user, aiding the process of cuneiform scholarly edition publication and improving as the user corrects them. This way, it is not only the machine learning models that benefit from the corrections and tagged data created by experts but also the experts can enjoy a designated work environment for cuneiform studies and download the results of their work, which is already advancing cuneiform scholarship.

The Babylonian Engine project has been building towards an OCR pipeline from 2D images of cuneiform signs to transliteration and translation. Our first attempt focused on stroke identification and vectorisation from 2D images of cuneiform signs. Identifying the strokes, the constituent parts which make up a cuneiform sign, is drastically simpler than identifying the whole. While the signs changed drastically in the 3,000 years in which cuneiform was in use, the strokes and writing technique remained similar.

Is it easier to train a machine learning model?

Furthermore, since there are only three main stroke types (horizontal, vertical and oblique), it is much quicker to obtain a large corpus of labelled examples of each, as compared to collecting enough samples for every sign and its variant forms.

As a parallel step, we also began work with hand-copies of cuneiform tablets. Hand-copies exist in large numbers and the cuneiform glyphs are much more distinct and easier to classify compared to real images of tablets. Therefore, it will be easier to successfully train a machine learning model on this task.

While developing the CuRe OCR, we identified the need for another type of OCR model, one that can enlarge the available number of digital transliterations. This is aimed primarily to help scholars and interested persons with knowledge in Akkadian to create their own digital dataset for computational research.

Results as highly successful as 89%

We trained a custom deep learning- based OCR model on Latin transliterations of Akkadian and achieved results as high as 89%. CuReD also provides a platform for correcting the model's initial results and fine-tuning the model on new, unfamiliar types of transliterations. A minimal amount of 10 texts can suffice to significantly improve the model's initial results.

To conclude, designing machine learning models as part of a human- in-the-loop pipeline application for OCR of ancient non-alphabetic scripts has the following benefits: (a) people are incentivised to create the data needed for training; (b) breaking down the "reading" process of the OCR, allows for less mistakes in the final digitised text; (c) the machine learning models are used from their inception in a real-world scenario, providing real- world value to the research community.

Please Note: This is a Commercial Profile