Standard Machine Learning Techniques for investigating relationships: Oracle Bone Scripts and Modern Chinese Characters

Tyler Sledge

Dr. Martin Zhao

Mercer University

CSC 485

07/29/2021

Table of Contents

Abstract

Learning foreign language is very important but hard undertaking. One difficult aspect for English-speaking students to master the Chinese language (which is rendered in a logographic writing system) is to recognize and memorize thousands of Chinese characters. Abstract forms (radicals) with straight strokes are widely used in modern Chinese characters to replace the pictorial features in ancient script forms, such as Oracle Bone Scripts, the earliest known form of Chinese writing dated back to the late 2nd millennium BC. This project has focused on making the connections between modern and ancient forms to help ease the learning curve. We have used standard Machine Learning and newer data preprocessing techniques in order to investigate the relationships between the Oracle Bone Scripts and the modern Chinese characters and improve upon an existing software architecture. Significant improvements were made in developing a non-relational database to facilitate working with large amounts of data and data sets. Initial results demonstrated that the use of recurrent neural networks on Modern Chinese Language forms have shown significant character stroke analysis. Our current goal is to recognize meaningful blocks (radicals) in the Oracle Bone Scripts. The results can help generate visual aids for learning through web based user interfaces and hopefully interpret those forms that are currently still not recognizable.

1 Introduction

The investigation of the Oracle Bone Scripts provides a lot of useful information needed for understanding how modern Chinese characters were created. The original Oracle Bone Scripts found carved into the bones of animals provides us with one of the earliest forms of written language that is still relevant to this day. Having this record of the ancient, researchers can trace the evolution from ancient to modern scripts over time. It is still important to make note of the ancient scripts as each individual script and character has a meaning behind it and reason for its existence. Some of the meaning can be lost and confused as the language evolves, like with any language, however modern Chinese is unique. Being a non-phonetic, logographic language, every single stroke, every direction of a stroke, and the order of each stroke differs vastly from the letters and characters of phonetic languages such as English. For language learners, the culture of the language provides a guide on how to use the language. With modern techniques, such as Natural Language Processing, we can attempt to analyze the uncovered Oracle Bone Scripts and find the similarities between the ancient and modern scripts.

2 Analysis of Existing Work

During the Fall 2020 Semester, a group of six students began a Software Engineering project to develop a web application for a English speakers wanting to learn the Chinese language. This project was developed using the React.js framework for the front-end, the Flask microframework for the backend and server side development, and a small builtin Python module called SQLite3 for the database. It was important for the team to use a Python-based framework on the backend so that it would be easier to integrate any Python related libraries for data preprocessing. The front end was made to include a small video game developed in

Javascript to help the user understand how different Chinese characters of complex strokes can be split into their respective simple strokes. The model-template-view architecture was used when developing and designing the REST APIs on the backend in Flask.

The team also used a small, relational database management system, SQLite3, when creating the database for the web application. The schema for which was relatively simple as it only included a single table with the script ID and an image of the script. This database only consisted of modern Chinese characters as we had a separate dataset made with some Python data processing tools.

Due to time constraints, the team was unable to complete any character recognition methods into the functionality application. The image processing development was done on a separate branch in the GitHub repository not used in the main working branch for the application's deployment. The developer's code does successfully generate processed, cropped images of scripts and other linguistics from Chen Nianfu's book of Oracle Bone Scripts, but there was not enough time allocated for use as a working unit within the system.

3 Software Integration Recommendations

From the research methods above, software engineers should try to use a native Python framework such as the Flask micro web development framework or the Django framework, which offers much more features. For larger projects, use Django, for smaller projects or the development of quicker prototypes use the Flask framework with whichever front end framework is the most familiar with the team.

Machine learning integration in production can be difficult to accomplish, often companies have to allocate more funds in order to support the storage of large datasets on company databases. MongoDB is a NoSQL database that stores objects in JavaScript Object Notation (JSON) documents. With this database tool, a team can store all of their datasets as tabular, non relationship databases. Training and testing datasets have the potential to start out as large datasets and continue to grow. This makes building a relational database for Oracle Bone Scripts very difficult. If a relational database is used, the team can quickly run into errors trying to query large amounts of data and face errors from having different data models. NoSQL databases offer flexible data models, which may be needed for different object types in our software. The utilization of the Flask framework for the back-end and MongoDB for the database offers a technical stack with a desirable amount of flexibility in design and development. With over 8000 simplified Chinese characters and over 30,000 distinct Oracle Bone Scripts, creating a flexible data pipeline with a horizontally scaled database management system is recommended.

4 Improved Development

Improvements were made using a slightly newer and upgraded technical stack with a focus being on facilitating the database development to be able to use efficiently by other tasks on the application. This provides a solid foundation to conduct future data processing and analytics work. The Django framework was selected for use as it follows the model-template-views architectural pattern and has a larger amount of features available from

the start of development. Being a Python based framework, it allows the use of existing Python-based data analysis tools as well as other popular machine learning libraries.

4.1 Database Development

The first team utilized a Python based module called SQLite3 that is automatically installed when using the Flask microframework. However, with the expected magnitude of data that would be expected to be entered into our system will be very large, with over 30,000 Oracle Bone Scripts. This does not include any Simplified Chinese (Hanzi Scripts) or other generations of Chinese Scripts. For analysis purposes, a team may want to analyze the differences in scripts over each generation to examine how they have changed. Designing the database easily can easily turn into a big-data problem.

The newer database was developed using a non-relational, NoSQL database called MongoDB. In this database, a large number of tables were created to hold values for each radical. Unlike the English, the radicals act as a way of indexing each class of scripts. Additional databases were created to hold any training and testing data sets either from other sources or ones that are created locally[1].

5 Conclusion and Future Works

One of the goals with this research project was to develop a database system that could be used for a language learning application. The database system was able to be improved from the

[1] https://doi.org/10.18653/v1/2020.emnlp-demos.29

original design, using a small relational database, to a non-relational, document based database. The flexibility of a non-relational database allows for developers to easily add large amounts of data, whether it be datasets for training and testing large datasets or maintaining a large dictionary of characters.

With Django as a backend, a software development team might also connect a mobile user interface and front end to it through the addition of more REST APIs. The touch screen of mobile phone or tablet would enable users to physically practice writing Oracle Bone Scripts and Modern Chinese characters in the correct stroke order.

For future development of this application, it would be useful to implement more machine learning capabilities into the language learning features. Several sources tested pattern and character recognition based neural networks for Oracle Bone Scripts. A source from the Shanghai Maritime University[2] confirmed the effectiveness of the development of a recurrent neural network model for modern Chinese characters. This neural network architecture type is popular for Natural Language Processing and translating between languages. This approach can be paired with an image based convolutional neural network to analyze handwriting of different strokes to provide a way to get feedback in real time without the need of a teacher. These neural networks types should also be able to be configured with datasets of Oracle Bone Scripts, allowing a better comparison of handwriting Oracle Bone Scripts and Modern Chinese characters.

[2] https://doi.org/10.1007/978-981-10-7605-3_146

References

1. Han, X., Bai, Y., Qiu, K., Liu, Z., & Sun, M. (2020). Isobs: An information system for

   oracle bone script. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. https://doi.org/10.18653/v1/2020.emnlp-demos.29

2. KOSTADINOV, S. I. M. E. O. N. (2018). Chapter 1: Introducing Recurrent Neural

   Networks; Chapter 2: Building Your First RNN with TensorFlow. In *Recurrent neural networks with python quick start guide: Sequential learning and language ... modeling with tensorflow* (pp. 1–33). essay, PACKT Publishing Limited.

3. Li, B., Dai, Q., Gao, F., Zhu, W., Li, Q., & Liu, Y. (2020). HWOBC-A handwriting oracle

   Bone character recognition database. *Journal of Physics: Conference Series*, *1651*, 012050. https://doi.org/10.1088/1742-6596/1651/1/012050

4. Lin, L., Liu, J., Gu, Z., Zhang, Z., & Ren, H. (2017). Build Chinese language model with

   recurrent neural network. *Advances in Computer Science and Ubiquitous Computing*, 920–925. https://doi.org/10.1007/978-981-10-7605-3_146

5. Liu, D., Yang, K., Qu, Q., & Lv, J. (2020). Ancient–modern chinese translation with a

   new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *19*(1), 1–13. https://doi.org/10.1145/3325887

6. Tao, J., Fang Zheng, Li, A., & Ya Li. (2009). Advances in Chinese natural language

   processing and language resources. *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*. https://doi.org/10.1109/icsda.2009.5278384

7. Xing, J., & Xiong, J. (2020). Spatial pyramid block for Oracle Bone Inscription

   Detection. *Proceedings of the 2020 9th International Conference on Software and Computer Applications*. https://doi.org/10.1145/3384544.3384561