# Named-Entity Recognition of Chinese Genealogical Records

Authors: Rex Gao and Sam Meng

## Abstract

*With no formal research nor datasets found on NER labeling of Chinese genealogical records, we used the BERT-Ancient-Chinese model to create the framework for our custom Chinese genealogy dataset, manually edit the labels, then use it to fine-tune the same model. We have found that fine-tuning on our genealogy dataset has resulted in ~10% increase of accuracy from a generic Chinese NER dataset and an increase of 0.0895 of the F1 score, with the total accuracy being 91% (from a baseline of 53%). This promising result shows that including more properly labeled data can make the BERT-Ancient-Chinese model a viable option as the NER in a data pipeline that takes raw text from genealogical records and extract useful information for building a family tree.*

## Introduction

With mental health issues on the rise in the world today, researchers have been looking for ways to improve mental health and resilience in young adults. One such method is the discovery of one's roots. As summarized by Driessnack (2017), intergenerational narratives create an expanded sense of self "referred to as our intergenerational self. Developing an intergenerational self not only grounds an individual but also provides a larger context for understanding and dealing with life's experience(s) and challenges".

Traditional genealogical records known as Jiapu are an important part of Chinese culture that helps people of Chinese descent to understand their roots. These records include stories of ancestors and a lineage to them like a family tree. Websites such as FamilySearch.org are digitizing these genealogical records, but do not have a way to store them as indexable family trees without manual effort. With many of these records written in ancient Chinese, we need a machine learning model that can parse out important information from these records to easily make them searchable and available to all.

Although to fully solve the problem of indexing scanned Chinese genealogical records requires accuracy of a data pipeline in OCR (to parse scanned text), NER (to recognize entities), and relations tagging (to recognize relationships between entities), the scope of this project is on the NER automation of this pipeline. Our goal is to research whether fine-tuning existing language models trained on classical Chinese literature can help an existing Chinese named-entity recognition model to be able to extract names, places, and dates from Chinese Jiapu books.

# Background

We were not able to find an existing Chinese Jiapu dataset to work with, nor any prior research done on Chinese genealogical records. However, we have found research on training large language models to understand classical Chinese.

We have first found a reference dataset we used is from Xu et al. (2017). They published a Classical Chinese dataset that followed the IOB format in labeling entities. We have found that about 70% of the text is not labeled, making it as the baseline of models trained on that dataset. At that time, the best model for completing the NER task was the Conditional Random Field (CRF) model, with an F1 score of 71.33.

We found multiple attempts to improve on the score after the release of BERT, including the Siku-BERT model (Wang et al., 2021). Most attempts increase the vocabulary size of BERT by expanding the dataset due to the higher variety of characters seen in ancient Chinese literature. The model we use for this project is the one proposed in Wang & Ren (2022), known as BERT-Ancient-Chinese: a two-stage BERT model that takes the preliminary predictions and uncertain components from the first stage as input to the second stage, called a “Knowledge Fusion Model”. The Knowledge Fusion BERT model then outputs the final prediction results. Appending the uncertain components determined from Monte Carlo dropout methodology to the preliminary outputs increases the overall accuracy of the model.

# Methods

To achieve our objective on fine-tuning a model to recognize names, locations, and dates in a Chinese Jiapu record, we created a dataset using the IOB2 format in labeling the entities in text:

- PERSON: A proper noun describing a person
- LOC: A proper noun describing a location
- TIME: Text describing an absolute date/period in history.
- O: No label

Each labeled entity (PERSON, LOC, TIME) is marked with a prefix B- or I- representing the first token and all subsequent tokens of each entity respectively. This helps differentiate between adjacent chunks of tokens, which is useful because Classical Chinese is often unpunctuated so using the IOB2 format can help us identify individual entities in, for instance, lists of people.

Our baseline we will use is the prediction of the most common classification in the dataset, which in our case is “O”. This is a common baseline for NER problems and can help us know whether our model will bring any value to NER on Chinese Jiapu data.

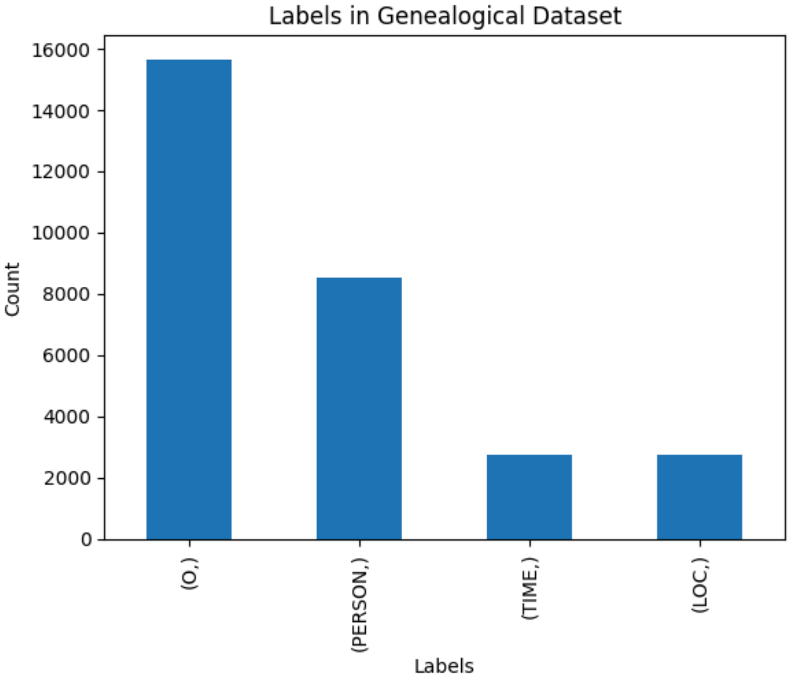
## Fine tuned BERT-Ancient-Chinese Model

For our main model, we will be fine tuning BERT-Ancient-Chinese, a BERT model which is pre-trained on classical Chinese. Out of all the available pre-trained models for classical

Chinese, BERT-Ancient-Chinese had the most potential, scoring the highest on precision, recall, and F1 score (Feng et al.). We will also compare the performance of the BERT-Ancient-Chinese base model prior to fine tuning, to serve as a control.

For both the control and fine tuned model, we will first train the model on an existing Chinese literature corpus (Chinese-Literature-NER-RE-Dataset), containing just over 1 million tokens of training data, for 3 epochs. For the fine tuned model, we will then fine tune on a curated dataset of Jiapu genealogies. Because of the limitations and constraints on size of this dataset, we will run 5 rounds of cross validation for the fine tuning, at 5 epochs each, which we found to be a good sweet spot for our validation loss and accuracy. For both models, we use a max length of 200 tokens as our window, a batch size of 50 tokens, a learning rate of 0.0001, and a dropout rate of 0.3.

## Genealogical Dataset

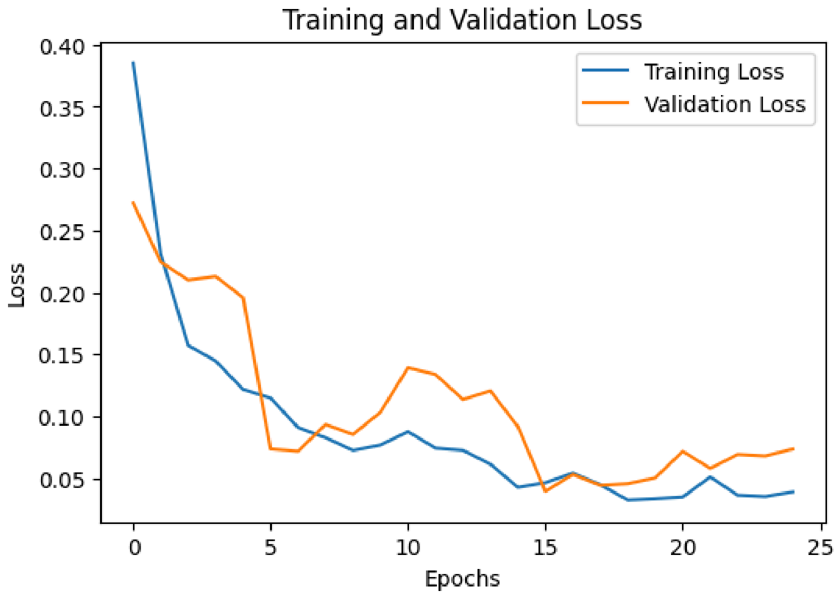


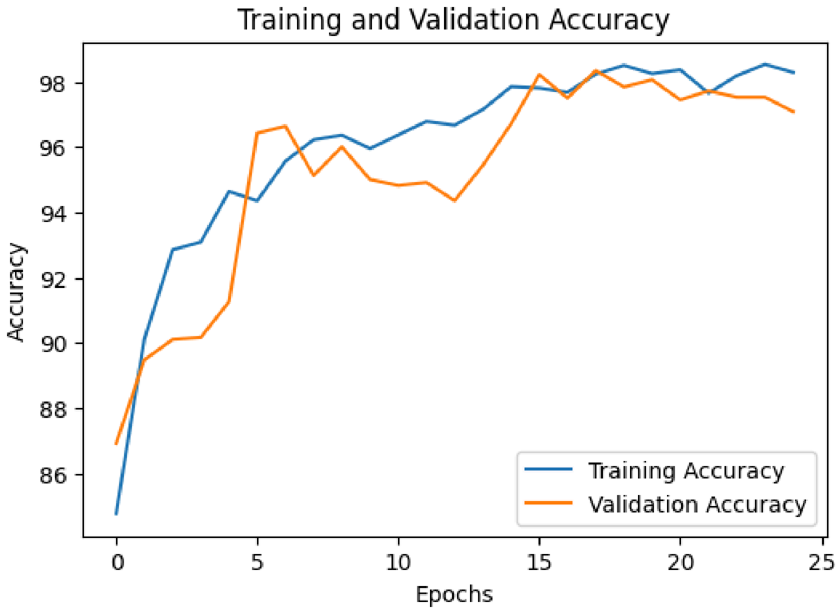
To create our own genealogical dataset to finetune the BERT-Ancient-Chinese model, we first took images from FamilySearch.org and Internet Archive and passed the images through Google Cloud's Vision API for OCR. We have found that the OCR API yields the results above 95% with minimal effort from us.

We then passed the raw text into the BERT-Ancient-Chinese model to retrieve the entity labels, then store it using a similar format as done by Xu et al. (2017). We were able to process 29,608 Chinese characters, with 15,648 characters without a label, or 53% of the dataset. Our baseline is therefore 53%. Our dataset has considerably more labels of persons, which is expected of genealogical records. However, this skew of the data will affect the model's performance of predicting a PERSON label versus a LOC label.

Our hypothesis is that by fine tuning the BERT-Ancient-Chinese model to our correct labels, it should be able to recognize the three entities with similar performance to the classical Chinese dataset. Our focus will be on the F1 score due to the skewness of our dataset. The F1 score will measure how well our model is performing in respect to precision and recall.

# Results and Discussion





Looking at the graphs for both loss and accuracy during our fine-tuning, we can see that while loss and accuracy for validation was somewhat turbulent, we were able to stop right as or right before they reached their cusp. We also see that the loss and accuracy curves matched quite well, with training and validation both remaining quite close to one another throughout the fine-tuning. This indicates a fairly good training, without any overfitting or underfitting.

As expected, the pre-trained BERT model performed much better than our baseline, but our fine tuned model performed even better than the pre-trained model, with an increase in accuracy of about 11%. The increase in F1 score shows the overall improvement in our fine-tuned model, with a moderate improvement in precision, but a significant improvement in recall, indicating our model is better at identifying the correct entities. Considering the marked improvement even with our fairly limited dataset, we believe that this is a worthwhile approach, and with an even larger dataset, we are optimistic that results can improve further with more research.

Looking more closely at the individual labels, we see somewhat poorer results for location entities. One possible explanation for this is that many locations are relatively small and/or obscure, and also have frequent name changes over time, possibly tens over the course of history. In comparison, people tend to stick with one or two names throughout their lives, at maximum only a few. We also see poorer precision for time entities, meaning that we see more false positives. Some of these may be due to confusion from era names, which generally correspond to each Chinese emperor's reign and in some cases used to refer to them by extension.

### Results Tables

Overall entity recognition metrics:

| | Baseline | BERT-Ancient-Chinese | Fine-tuned BERT-Ancient-Chinese |
|---|---|---|---|
| Accuracy | 0.5285 | 0.8197 | 0.9104 |
| F1 | - | 0.7630 | 0.8525 |
| Precision | - | 0.7987 | 0.8314 |
| Recall | - | 0.7304 | 0.8748 |

Fine-tuned model metrics per label:

| | No label ('O') | Person | Location | Time |
|---|---|---|---|---|
| F1 | 0.9290 | 0.8877 | 0.7941 | 0.8246 |
| Precision | 0.9480 | 0.9095 | 0.7876 | 0.7877 |
| Recall | 0.9107 | 0.8668 | 0.8331 | 0.8806 |

# Conclusion

We have found the BERT-Ancient-Chinese model to be as robust as described in the paper by Wang, P., & Ren, Z. (2022). We were able to fine-tune it with good performance on our dataset, but we did see that labels with data sparsity like location and time did not do well. Therefore, more research and data labeling is needed to see if it is because of a lack of data or the complexity of that labeling. Classical Chinese is also an information dense but highly contextual and ambiguous language, making NLP tasks in general more difficult than most. Nevertheless, we remain optimistic and believe that our results show the potential of the BERT-Ancient-Chinese model as a base for a more refined NER model that can extract names, locations, and dates to help make ancestor stories and lineage quickly and easily accessible to all.

# References

Deng, Y., Deng, Y., & Deng, J. (Eds.). (2002). 安仁里鄧氏族譜 - 村版本 (S. Deng, Y. Deng, Y. Deng, J. Deng, J. Deng, B. Deng, Z. Deng, & F. Deng, Compilers). 廣東省非營利性出版. Retrieved from Internet Archive website: https://archive.org/details/genealogydeng1

Driessnack, M. (2017). "Who Are You From?": The Importance of Family Stories. *Journal of family nursing*, *23*(4), 434-449. https://doi.org/10.1177/1074840717735510

FamilySearch.org. (n.d.). *Guangdong, China records*. FamilySearch. Retrieved July 25, 2024, from https://www.familysearch.org/ark:/61903/3:1:3Q9M-CSMN-V344-Y?view=explore

Feng, P., Kang, J., Huang, B., Wang, S., & Li, J. (2022). "A method of Named Entity Recognition in Classical Chinese based on Bert-Ancient-Chinese," *2022 4th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 137-141, https://doi.org/10.1109/MLBDBI58171.2022.00033.

Wang, D., Liu, C., Zhu, Z., Feng, J., Hu, H., Shen, S., & Bin. (2021, 06 04). *Construction and Application of Pre-training Model of "Siku Quanshu" Oriented to Digital Humanities*. Retrieved August 4, 2024, from http://www.kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/2021-06-04/Wang2021-05.pdf

Wang, P., & Ren, Z. (2022). The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 164-168. https://aclanthology.org/2022.lt4hala-1.25.pdf

Xu, J., Wen, J., Sun, X., & Su, Q. (2017). A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text. *CoRR*, *abs/1711.07010*. http://arxiv.org/abs/1711.07010