When Less Is More
Logits-Constrained Framework with RoBERTa for Ancient Chinese NER

Wenjie Hua
School of Chinese Language and Literature
Wuhan University, China
huawenjie@whu.edu.cn

Shenghan Xu
Yuanpei College
Peking University, China
xsh2022@stu.pku.edu.cn

Abstract

This report presents our team's work on ancient Chinese Named Entity Recognition (NER) for EvaHan 2025[1]. We propose a two-stage framework combining GujiRoBERTa with a Logits-Constrained (LC) mechanism. The first stage generates contextual embeddings using GujiRoBERTa, followed by dynamically masked decoding to enforce valid BMES transitions. Experiments on EvaHan 2025 datasets demonstrate the framework's effectiveness. Key findings include the LC framework's superiority over CRFs in high-label scenarios and the detrimental effect of BiLSTM modules. We also establish empirical model selection guidelines based on label complexity and dataset size.

1 Introduction

Named Entity Recognition (NER) is basically a task to identify and classify named entities in texts, such as person name, geographical location, and time expression. It is a crucial research topic in NLP. NER in Ancient Chinese is particularly challenging due to the complex semantic properties of words, which can lead to errors in label sequence predictions. To address this, our model integrates the Logits-Constrained Framework with GujiRoBERTa[2], effectively reducing such errors.

2 Related Work

2.1 RoBERTa

Large-scale pre-trained language models (PLMs) based on Transformer architectures (Vaswani et al., 2023) have revolutionized sequence labeling tasks. RoBERTa (Liu et al., 2019), an optimized variant of BERT (Devlin et al., 2019), steadily improved Ancient Chinese NER accuracy. GujiRoBERTa, pre-trained on a large corpus of traditional Chinese texts, serves as the backbone model in our EvaHan 2025 close-modality setting and is a fine-tuned version of SikuRoBERTa.

2.2 Transition Constraints in Sequence Labeling

Sequence labeling tasks require strict adherence to structural constraints defined by tagging schemes. For instance, under the BMES scheme where valid label sequences must conform to $S_3 = \mathrm{Perm}(\{B, M, E\})$, the transition (B, M, E) is the only valid transition in S_3. Traditional approaches employ Conditional Random Fields (CRFs) (Lafferty et al., 2001) with bidirectional LSTMs (BiLSTMs)(Huang et al., 2015) to globally normalize label transition probabilities during inference. However, these methods depend on manually designed transition matrices and often produce illegal paths when decoding under low-resource or label-sparse senarios.

Recent work explores alternative constraint mechanisms. For example, Jiang et al. (2021) proposes a constrained transition framework that dynamically masks invalid transitions during training and inference. Similarly, Wei et al. (2021) develops a masked transition learning approach that implicitly encodes tagging scheme rules through auxiliary language modeling objectives. Our work extends these paradigms by directly incorporating transition constraints into the model's parameterized decision boundary, which eliminates heuristic post-processing while maintaining theoretical guarantees of valid output structures.

3 Method

3.1 Pre-processing

Punctuation marks provide potential entity boundary information, and preserving and correctly segmenting them can enhance NER performance (Ge, 2022) . Considering the characteristics of punc-

[1] https://github.com/GoThereGit/EvaHan
[2] https://huggingface.co/hsc748NLP/GujiRoBERTa_jian_fan

tuation in the EvaHan 2025 training sets, we adopt different sentence segmentation strategies. Specifically, `trainset_c` only considers primary sentence-ending punctuation: "。", "！", and "？". In contrast, `trainset_a` and `trainset_b` additionally account for "」" and "』", as well as """ and "" as special sentence-final markers.

3.2 Framework

Motivated by the Occam's razor principle – that simpler hypotheses consistent with observations are preferable (MacKay, 2003) – we propose a minimally invasive two-stage architecture that maintains model simplicity while enforcing structural constraints. Our design philosophy consciously avoids stacking complex components like CRFs or BiLSTMs, which may introduce interference patterns during learning.Just as illustrated in Figure 1, the framework operates through.

3.2.1 Stage 1: Contextual Encoding with GujiRoBERTa

The pre-trained GujiRoBERTa model generates contextualized embeddings $\mathbf{h}_i \in \mathbb{R}^d$ for each token x_i, capturing ancient linguistic patterns through its 12-layer transformer architecture. A linear projection layer then computes initial label logits:

$$\mathbf{l}_i = \mathbf{W}\mathbf{h}_i + \mathbf{b} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{k \times d}$ maps to k possible labels. Training uses standard cross-entropy loss without explicit transition modeling.

3.2.2 Stage 2: Logits-Constrained Decoding

We introduce a constraint matrix $\mathbf{M} \in \{0, 1\}^{k \times k}$ encoding valid BMES transitions (e.g., B-PER can only transition to M-PER or E-PER). During inference, we modulate the logits sequence $\{\mathbf{l}_1, ..., \mathbf{l}_n\}$ through masked autoregressive refinement:

$$\mathbf{l}'_t = \mathbf{M}[y_{t-1}] \odot \mathbf{l}_t + (1 - \mathbf{M}[y_{t-1}]) \cdot (-\infty) \tag{2}$$

where y_{t-1} denotes the previous token's predicted label. This differentiable masking ensures structurally valid outputs without additional trainable parameters.

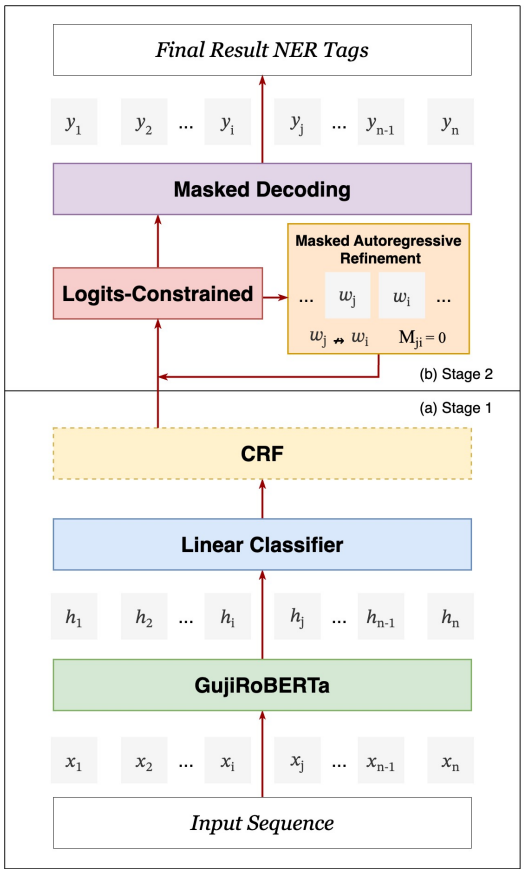

Figure 1: Framework Overview

4 Experiments

Following EvaHan 2025 guidelines, we use three training sets—trainset_a, trainset_b, and trainset_c—annotated with 6, 3, and 6 NER categories, respectively, plus a non-NER label "O." The {B, M, E, S} scheme marks entity positions as Begin, Middle, End, or Single. Since trainset_b's categories are a subset of trainset_a's, the dataset includes 37 classification labels.

4.1 Experimental Environment

All experiments were conducted on Google Colab using NVIDIA A100 (40 GB) and T4 GPUs with mixed precision (FP16) training enabled.

4.2 Parameter Regulation

The model was trained for 4 epochs with a batch size of 8 for training and 1 for evaluation. The learning rate was set to 2×10^{-5} with a warmup ratio of 0.1 and a weight decay of 0.01 to mitigate overfitting. Gradient accumulation was performed over 2 steps, with a linear scheduler adjusting the learning rate progressively.

4.3 GujiRoBERTa

We only employed GujiRoBERTa with an additional linear classifier to evaluate the NER tagging results, without incorporating any additional components. Nevertheless, this approach achieved promising performance during training (see Table 1).

Dataset	P	R	F1
A	0.9170	0.9190	0.9180
B	0.9251	0.9221	0.9236
C	0.7744	0.8418	0.8067

Table 1: Performance of GujiRoBERTa

4.4 Cross-Comparison

Therefore, we conducted further cross-comparison experiments, drawing parallels with typical configurations in NER tasks to assess the relative contributions of different model components and potential performance improvements. In the following tables, "+" indicates the inclusion of the corresponding module, while "-" denotes its exclusion.

BiLSTM	CRF	LC	F1
-	-	-	0.9180
+	-	-	0.9016
-	+	-	0.9143
-	-	+	0.9269
+	+	-	0.8850
-	+	+	0.9213
+	-	+	0.8976
+	+	+	0.8947

Table 2: Results of Dataset A

BiLSTM	CRF	LC	F1
-	-	-	0.9236
+	-	-	0.8617
-	+	-	0.9278
-	-	+	0.9218
+	+	-	0.9100
-	+	+	0.9308
+	-	+	0.8594
+	+	+	0.9012

Table 3: Results for Dataset B

BiLSTM	CRF	LC	F1
-	-	-	0.8067
+	-	-	0.7383
-	+	-	0.8112
-	-	+	0.8262
+	+	-	0.7602
-	+	+	0.8314
+	-	+	0.7547
+	+	+	0.7804

Table 4: Results for Dataset C

Through cross-comparison of the results (see Table 2, Table 3, and Table 4), we found that CRF effectively captures sequence patterns in low-dimensional label spaces by leveraging predefined transition constraints. However, as the number of labels increases, the performance of CRF decreases by 1.3% and 0.5% on Datasets A and C, respectively. This is likely because manually designed transition matrices are less capable of covering high-dimensional state spaces.

In contrast, the Logits-Constrained (LC) framework demonstrates greater generalizability. In scenarios with six or more labels ($L \geq 6$) (Datasets A/C), our LC framework exhibits a significant advantage, achieving an average F1 improvement of 1.95%. Notably, on Dataset C, which features a complex entity distribution, the dynamic masking mechanism in LC raises the F1 score from the baseline of 0.8067 to 0.8262 (+2.95%).

Moreover, the introduction of BiLSTM leads to performance degradation across all datasets, with an average $\Delta F_1 = -3.8\%$. We speculate that this is due to the disruption of the inherent attention patterns in the pretrained model caused by the addition of BiLSTM, as well as the increased risk of the bidirectional recurrent structure's parameter updates getting trapped in local optima.

4.5 Dataset Expansion

By integrating the annotated data from Dataset A according to the specifications of Dataset B, we expand the sample size of the hybrid Dataset B from 3,434 sentences to 11,307 sentences (+229%), and conduct the same experiments (see Table 5).

Dataset	Sentences	Label Types
Dataset B	3434	13
Hybrid	11307	13

Table 5: Statistics of Datasets

BiLSTM	CRF	LC	F1
-	-	-	0.9369
+	-	-	0.8964
-	+	-	0.9465
-	-	+	0.9395
+	+	-	0.8364
-	+	+	0.9439
+	-	+	0.8957
+	+	+	0.9401

Table 6: Results for Dataset B (Hybrid)

Table 6 demonstrates a positive correlation between dataset scale and model performance in NER, with the baseline F1 score increasing by 1.33% under consistent model settings. Since the CRF's global normalization enhances long-range dependency modeling and LC's dynamic masking mitigates overfitting in sparse label scenarios, the combined application of the CRF and LC frameworks yields optimal performance, surpassing the performance of individual framework implementations.

4.6 Model Selection

As noted earlier, balancing dataset size and label complexity is crucial in sequence labeling tasks. We define the optimal model selection as a function of label cardinality L and sentence count N, yielding the following empirically optimized scaling relationship:

$$\Gamma(L,N) = \begin{cases} -\text{ (LC)} & \text{if } L \geq 20 \\ & \quad \wedge\, N > 0.16L^{2.8} \\ +\text{ (CRF+LC)} & \text{otherwise} \end{cases} \tag{3}$$

Here, the threshold $0.16L^{2.8}$ is derived via parameter tuning across various datasets, and the exponent 2.8 accurately quantifies the super-linear penalty imposed by increasing label complexity on the required amount of data.

Within this framework, we identify two primary operational regimes. When label complexity is high and data is abundant, the Logits-Constrained (LC) model effectively mitigates the overfitting risk associated with the CRF's transition matrix, leading to significant performance gains. Empirical results show that the LC model explains 82% of the performance variance in this setting. Conversely, for moderate label complexity or limited data, a CRF+LC combination leverages both components: CRF captures tag transitions, while LC acts as a regularizer. The term $L^{2.8}$ quantifies the exponential increase in data required to justify an LC-only approach as label complexity grows.

To refine model selection, we formulate the configuration problem as a constrained optimization:

$$\min_{\alpha,\beta} \sum_{i=1}^{4} \left(F1_{\text{best}}^{(i)} - F1_{\text{pred}}^{(i)} \right)^2 e^{-\alpha \frac{N_i}{L_i^{\beta}}} \tag{4}$$

This is solved via gradient descent, yielding optimal parameters $\alpha = 0.16$ and $\beta = 2.8$.

Ablation studies on BiLSTM integration show consistent performance degradation ($\Delta F1 = -2.4\% \pm 1.1\%$), with the negative impact increasing in high-label, low-data settings:

$$\deg(\text{BiLSTM}) \propto L^{1.7} N^{0.6} \tag{5}$$

This suggests that BiLSTM's detrimental effect is amplified under high label density and limited data.

Based on the above analysis, we provide the following practical guidelines. First, eliminate the BiLSTM module in all configurations. Second, use the CRF+LC model by default when $L \leq 13$ or $N \leq 0.16L^{2.8}$ to fully capture transition dependencies. Third, switch to an LC-only model when $L \geq 20$ and $N > 0.16L^{2.8}$ to avoid overfitting and leverage the benefits of abundant data.

5 Conclusion

We propose a Logits-Constrained framework with GujiRoBERTa for ancient Chinese NER. The two-stage pipeline enforces BMES constraints through dynamic logits masking, eliminating invalid transitions while maintaining simplicity. Experiments show that LC outperforms traditional CRF-based methods, improving F1 by up to 2.95% in complex label scenarios. BiLSTM integration degrades performance, while dataset expansion and hybrid CRF+LC improve robustness. A data-driven model selection criterion shows LC alone excels when label count $L \geq 20$ and data size $N > 0.16L^{2.8}$. This work offers a practical, theoretically sound solution for ancient Chinese NER.

6 Limitations

Although our framework achieves high accuracy with a compact design, several limitations remain. First, the predefined Logits-Constrained matrix M is based on manual BMES rules, which may not generalize well and is highly sensitive to the accuracy of the initial token. Second, the two-stage pipeline introduces additional inference overhead compared to end-to-end models. Third, performance depends on sentence segmentation quality, making it vulnerable to errors in unpunctuated or irregular historical texts. Future work could explore adaptive constraint learning and unified architectures to address these issues.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sijia Ge. 2022. Integration of named entity recognition and sentence segmentation on Ancient Chinese based on siku-BERT. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 167–173, Taipei, Taiwan of China. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *Preprint*, arXiv:1508.01991.

Haoming Jiang, Danqing Zhang, Tianyu Cao, Bing Yin, and Tuo Zhao. 2021. Named entity recognition with small strongly labeled and large weakly labeled data. *Preprint*, arXiv:2106.08977.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David J. MacKay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Tianwen Wei, Jianwei Qi, Shenghuan He, and Songtao Sun. 2021. Masked conditional random fields for sequence labeling. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2024–2035, Online. Association for Computational Linguistics.