Multi-Strategy Named Entity Recognition System for Ancient Chinese

Wenxuan Dong, Meiling Liu*
Northeast Forestry University, China
{curleydong, lmling2008}@163.com

Abstract

We present a multi-strategy Named Entity Recognition (NER) system for ancient Chinese texts in EvaHan2025. Addressing dataset heterogeneity, we use a Conditional Random Field (CRF) for Tasks A and C to handle six entity types' complex dependencies, and a lightweight Softmax classifier for Task B's simpler three-entity tagset. Ablation studies on training data confirm CRF's superiority in capturing sequence dependencies and Softmax's computational advantage for simpler tasks. On blind tests, our system achieves F1-scores of 83.94%, 88.31%, and 82.15% for Test A, B, and C—outperforming baselines by 2.46%, 0.81%, and 9.75%. With an overall F1 improvement of 4.30%, it excels across historical and medical domains. This adaptability enhances knowledge extraction from ancient texts, offering a scalable NER framework for low-resource, complex languages.

1 Introduction

Named Entity Recognition (NER), a fundamental task in information extraction, identifies key entities such as person names, locations, and organizations within text. It is essential for applications like information retrieval (Fetahu et al., 2021; Wang et al., 2022; Mokhtari et al., 2019). In ancient literature, NER supports the analysis of ancient Chinese texts and the extraction of humanistic knowledge. However, this task faces challenges due to limited public datasets and the unique features of classical texts, including polysemy, continuous structure, and unpunctuated traditional Chinese characters, all of which complicate entity boundary detection.

The EvaHan2025 competition[1] tackles these challenges with a 500,000-character dataset of historical and medical classical texts, expertly curated through automated annotation and manual review. Spanning subsets from *Shiji*, *Twenty-Four Histories*, and *Traditional Chinese Medicine Classics*, it encompasses diverse entity types and linguistic styles. To tackle this complexity, we propose a multi-strategy NER framework for EvaHan2025. Our system integrates a Conditional Random Field (CRF) model to capture intricate sequence dependencies in Tasks A and C, paired with a lightweight Softmax classifier for Task B to optimize efficiency for its simpler tagset. This hybrid approach outperforms official baselines, demonstrating robustness across heterogeneous datasets and advancing NER for ancient Chinese texts.

2 Related Work

2.1 Named Entity Recognition

Deep learning has shifted NER from rule-based methods to neural networks, which automatically extract features from text, improving efficiency over manual rule design. Huang et al. (Huang et al., 2015) proposed BiLSTM-CRF, combining BiLSTM's long-distance dependency capture with CRF's sequence optimization, excelling on the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003). (Ma and Hovy, 2016) advanced this with BiLSTM-CNN-CRF, using CNNs for word-level features and CRF for refinement, boosting English NER performance (Wang et al., 2022). Transformer-based models later enhanced results with contextual embeddings (Mokhtari et al., 2019), leading to paradigms like sequence labeling (Lample et al., 2016; Devlin et al., 2019), span-based recognition (Fu et al., 2021), and text generation (Zhang et al., 2022).

While these methods excel in modern languages like English and Chinese (Mokhtari et al., 2019), ancient Chinese NER remains underexplored. The EvaHan2025 competition addresses this by providing an ancient Chinese dataset, advancing domain-specific NER research.

** Corresponding author.
[1] https://github.com/GoThereGit/EvaHan

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 151–158
May 4, 2025.

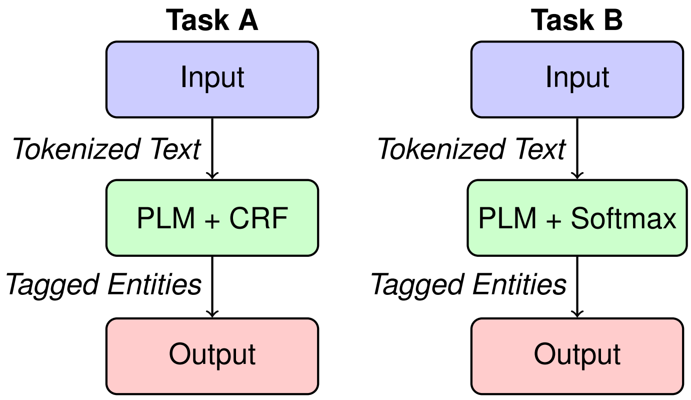

Figure 1: Architecture of the Multi-Strategy NER System. The system employs GujiRoBERTa_jian_fan as the PLM, paired with CRF for Tasks A and C (six entity types) and Softmax for Task B (three entity types).

2.2 Pre-trained Language Models

Pre-trained Language Models (PLMs) have revolutionized NLP tasks, including NER, by providing rich contextual representations. BERT (Devlin et al., 2019) pioneered this approach, with variants like RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020) enhancing efficiency. For ancient Chinese, specialized models like Siku-BERT (Wang et al., 2021) have been developed to address unique linguistic features, significantly improving performance in downstream tasks such as NER.

3 Method

3.1 Pre-processing

To avoid redundant code, we use the seqeval library for validation—even though it does not support BMES annotations. Thus, we convert BMES prefixes to BIOES during preprocessing, reducing the need for custom evaluation functions. We term this a simplified preprocessing algorithm. Secondly, in the data preprocessing stage, we process it through the custom "NERDataset" class. This class inherits from Dataset, can read text file paths and label file paths, filter out overly long sentences, and form tuples of samples and labels to meet the training requirements of the model. The EvaHan2025 dataset exhibits heterogeneity across Tasks A, B, and C, with varying entity types (six in Tasks A and C vs. three in Task B) and domain styles (*Shiji*, *Twenty-Four Histories*, and *TCM Classics*), necessitating a tailored strategy for each task.

3.2 Model

The architecture of our model is shown in Figure 1. To address the heterogeneity of the EvaHan2025 dataset, we propose a multi-strategy NER framework. We adopt GujiRoBERTa_jian_fan[2], a competition-mandated pre-trained model on ancient Chinese texts, to generate contextual representations $\mathbf{H}$ from an input sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. The model yields representations $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$:

$$\mathbf{H} = \text{GujiRoBERTa\_jian\_fan}(\mathbf{x}). \tag{1}$$

For Tasks A and C, which involve six complex entity types (Table 4), we employ a CRF layer to capture intricate label dependencies, computing the optimal sequence:

$$Y = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{H}), \tag{2}$$

where $P(\mathbf{y} \mid \mathbf{H})$ integrates transition and emission scores (Lafferty et al., 2001).

Conversely, for Task B's simpler three-entity tagset (Table 4), we use a Softmax layer to predict tags efficiently:

$$P(y_i = c \mid \mathbf{h}_i) = \frac{\exp((\mathbf{W}\mathbf{h}_i + \mathbf{b})_c)}{\sum_{c'} \exp((\mathbf{W}\mathbf{h}_i + \mathbf{b})_{c'})}, \tag{3}$$

This choice leverages Task B's reduced label transition complexity (three entities vs. six in Tasks A and C), where CRF's sequence modeling is less critical, as validated by ablation studies (Table 3), prioritizing Softmax's computational efficiency without sacrificing accuracy.

This hybrid approach leverages annotated data to bypass boundary ambiguity, with CRF ensuring accuracy for complex tasks and Softmax enhancing efficiency for simpler ones.

[2] https://huggingface.co/hsc748NLP/GujiRoBERTa_jian_fan

Subset	Task (Domain)	Labeled	Characters	Purpose
Training	A, B, C	Yes	320,000	Model Training
Validation	A, B, C	Yes	80,000	Model Selection
Blind Test	A, B, C	No	100,000	Final Evaluation

Table 1: Dataset statistics for EvaHan2025. Tasks correspond to domains: A (*Shiji*), B (*Twenty-Four Histories*), C (*Traditional Chinese Medicine Classics*). Total characters: 500,000.

Method	Test A			Test B			Test C			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline	85.90	77.50	81.48	87.09	87.92	87.50	71.84	72.95	72.40	81.41	79.82	80.61
Ours	**89.13**	**79.32**	**83.94**	**89.34**	87.30	**88.31**	**78.37**	**86.32**	**82.15**	**85.16**	**84.66**	**84.91**

Table 2: Performance Comparison (Precision, Recall, F1, as Percentages) Between Our System and the Baseline Across Test A, B, and C in EvaHan2025 Blind Tests (Close Modality).

4 Experiments

4.1 Dataset

We used the EvaHan2025 dataset, comprising 500,000 characters across three domains: Task A (*Shiji*), Task B (*Twenty-Four Histories*), and Task C (*Traditional Chinese Medicine Classics*). Statistics are detailed in Table 1, with entity tagsets in Table 4. The labeled data was split into training (80%, 320,000 characters) and validation (20%, 80,000 characters) sets for model training and validation, respectively. The unlabeled blind test set (100,000 characters) was used solely for final evaluation by the organizers, with predictions submitted post-training. This separation ensures robust and fair results.

4.2 Implementation Details

We built all models atop GujiRoBERTa_jian_fan, a pre-trained model from the Transformers library. For Tasks A and C, we added a CRF task head using the CRF library and applied a layered learning rate strategy. For Task B, we appended a Softmax layer. Models were optimized with AdamW (Loshchilov and Hutter, 2019), and performance was assessed using the seqeval library. Experiments ran on the environment in Table 5, with key hyperparameters listed in Table 10. Full details and code are available on GitHub.[3]

4.3 Metrics

In accordance with the conventions of Named Entity Recognition, we use Precision (P), Recall (R), and F1 score (F1) as evaluation metrics across all experiments. All results are reported in percentage form to ensure consistency and facilitate comparison across different models and experimental settings.

4.4 Baseline

To better evaluate our model's effectiveness, we use the official SikuRoBERTa-BiLSTM-CRF, trained on the training set without additional resources, as the baseline. Comparing our model with this baseline offers a clearer understanding of its performance and advantages.

4.5 Results

Results are presented in Table 2. Our system surpasses the baseline across all metrics for Tasks A, B, and C, achieving average F1 gains of 4.30%. This superiority stems from our multi-strategy approach: CRF effectively captures complex entity dependencies in Tasks A and C, while Softmax enhances efficiency for Task B's simpler tagset, showing strong adaptability to ancient Chinese datasets. Notably, Task C's F1 improves most (9.75%), likely due to CRF leveraging the structured patterns of *TCM Classics*, unlike Task A's diverse *Shiji* or Task B's simpler tagset (Table 4).

4.6 Ablation Study

We evaluated our multi-strategy design on EvaHan2025 using GujiRoBERTa_jian_fan as the PLM, reserving 20% of the training data as the validation set for strategy selection. Validation F1 scores are reported in Table 3 as percentages.

[3] https://github.com/wxndong/MSNER4AC

Configuration	Task A	Task B	Task C	Mean
		Single-Strategy		
PLM + CRF (All Tasks)	–	–	–	85.02
PLM + Softmax (All Tasks)	–	–	–	84.91
		Multi-Strategy		
PLM + CRF (Per Task)	**91.53**	86.79	**80.23**	86.18
PLM + Softmax (Per Task)	90.90	**86.87**	78.63	85.47
Ours (A/C: CRF, B: Softmax)	**91.53**	**86.87**	**80.23**	**86.21**

Table 3: Validation F1 scores (%). Single-strategy combines all task data; multi-strategy trains per task. '–' indicates unavailable task-specific scores for single-strategy models, as Task B's tagset (NR, NS, T) is a subset of Task A's (Table 4), causing interference that prevents isolated per-task evaluation.

4.6.1 Multi-Strategy vs. Single-Strategy

EvaHan2025 ranks submissions by mean F1 across Tasks A (*Shiji*), B (*Twenty-Four Histories*), and C (*Traditional Chinese Medicine Classics*). Single-strategy models (PLM + CRF and PLM + Softmax), trained on all tasks combined, yield mean F1s of 85.02% and 84.91%. Multi-strategy models (trained per task) reach 86.18% and 85.47%, gaining 1.16–1.27 points. This boost comes from isolating tasks: Task B's tagset (NR, NS, T) is a subset of Task A's (Table 4), causing single-strategy models to overgeneralize. Our approach avoids this interference, improving task-specific performance.

4.6.2 Task-Specific Strategy Selection

Comparing PLM + CRF (Exp. 3) and PLM + Softmax (Exp. 4) (Table 3, Appendix B), CRF excels on Tasks A (91.53% vs. 90.90%, +0.63) and C (80.23% vs. 78.63%, +1.60), handling six-entity dependencies well. Yet, in low-support labels (e.g., NB in Task A, ZZ in Task C), their differences are minor (Appendix B). For Task B, CRF (86.79%) and Softmax (86.87%) perform similarly, but Softmax cuts inference time by 63% (14.28s vs. 38.24s; Appendix 6). Our hybrid design—CRF for A and C, Softmax for B—achieves a mean F1 of 86.21%, balancing accuracy and efficiency.

4.6.3 Lightweight Analysis

For Task B, Softmax's $O(nk)$ decoding complexity (k=3) outperforms CRF's $O(nk^2)$, cutting blind test inference time by 63% (Please refer to Appendix 6) and reducing training/validation time from 202s to 86s, with F1 (86.87 vs. 86.79, +0.08). Here, n is sequence length, and k is label set size. This lightweight efficiency design optimizes efficiency for simpler tagsets without compromising accuracy.

5 Conclusion

In this paper, we propose a Multi-Strategy Named Entity Recognition (NER) system tailored for the EvaHan2025 competition. Our system demonstrates superior performance across three distinct datasets by leveraging task-specific strategies, including the use of CRF for complex sequence dependencies in Tasks A and C, and a computationally efficient Softmax classifier in Task B. Our system offers a scalable NER framework for similar low-resource, heterogeneous ancient language datasets, leveraging its multi-strategy adaptability, with potential applications in digital humanities. Future work could explore adaptive hyperparameter tuning and tagset refinement to further enhance generalization.

Limitations

Our multi-strategy NER system excels in EvaHan2025 but has limitations: inconsistent generalization and challenges with rare entities. Generalization varies across tasks. Task A's F1 drops from 91.53% to 83.94% (-7.59), likely due to overfitting to *Shiji*'s diverse data (Appendix C, Figure 2), while Task C's rises from 80.23% to 82.15% (+1.92), possibly due to a structured medical domain (Figure 3). Task B remains stable (86.87% vs. 88.31%) with a simpler tagset (Table 4). Rare entities (e.g., NB in Task A, ZZ in Task C) with low support (Appendix B) perform inconsistently. Future work could use cross-domain validation to improve generalization and data augmentation to enhance rare entity recognition.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer enhanced named entity recognition for code-mixed web queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Shekoofeh Mokhtari, Ahmad Mahmoody, Dragomir Yankov, and Ning Xie. 2019. Tagging address queries in maps search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):9547–9551.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Dongbo Wang, Chang Liu, Zihe Zhu, Jiang, Feng, Haotian Hu, Si Shen, and Bin Li. 2021. Construction and application of pre-training model of "siku quanshu" oriented to digital humanities.

Xiao Wang, Shihan Dou, Limao Xiong, Yicheng Zou, Qi Zhang, Tao Gui, Liang Qiao, Zhanzhan Cheng, and Xuanjing Huang. 2022. Miner: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. *arXiv preprint arXiv:2204.04391*.

Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. 2022. Exploring modular task decomposition in cross-domain named entity recognition. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 301–311.

A Supporting Tables in References

Tag	Meaning
	Task A (Shiji)
NR	Person name
NS	Geographical location
NB	Book title
NO	Official title
NG	Country name
T	Time expression
	Task B (Twenty-Four Histories)
NR	Person name
NS	Geographical location
T	Time expression
	Task C (TCM Classics)
ZD	TCM disease
ZZ	Syndrome
ZF	Medicinal formula
ZP	Decoction pieces
ZS	Symptom
ZA	Acupoint

Table 4: Entity tagsets for EvaHan2025 tasks.

Environment	Specification
CUDA Version	12.0
GPU	NVIDIA RTX 4090
Memory	24 GB

Table 5: Experimental environment.

B Additional Tables

This appendix provides tables supporting the experiments and ablation studies in Sections 4 and 4.6. Table 6 compares Task B runtime for PLM + Softmax and PLM + CRF, showing Softmax's efficiency (Section 4.6.2). Tables 7–9 detail per-category F1 scores for Tasks A, B, and C on the validation set, complementing Table 3 and guiding our multi-strategy NER design. Due to seqeval, F1 scores are rounded to two decimals and shown as percentages without decimals (e.g., 0.33 to 33%), not affecting comparisons.

Model	Training + Val. (s)	Blind Test (s)
PLM + Softmax	86	14.28
PLM + CRF	202	38.24

Table 6: Task B runtime comparison (seconds).

Category (Support)	F1 (CRF)	F1 (Softmax)
NB (5)	33.00	33.00
NG (731)	94.00	94.00
NO (286)	77.00	74.00
NR (2042)	95.00	95.00
NS (500)	87.00	87.00
T (193)	79.00	77.00

Table 7: Task A validation F1 scores (%).

Category (Support)	F1 (CRF)	F1 (Softmax)
NR (794)	91.00	89.00
NS (685)	84.00	83.00
T (509)	85.00	89.00

Table 8: Task B validation F1 scores (%).

Category (Support)	F1 (CRF)	F1 (Softmax)
ZA (294)	84.00	83.00
ZD (166)	73.00	73.00
ZF (197)	83.00	84.00
ZP (1083)	86.00	87.00
ZS (257)	65.00	57.00
ZZ (97)	47.00	33.00

Table 9: Task C validation F1 scores (%).

C Hyperparameters and Transition Matrix

Hyperparameter	Task A (PLM + CRF)	Task B (PLM + Softmax)	Task C (PLM + CRF)
Batch Size	32	32	32
Epochs	35	30	35
Learning Rate (PLM)	5×10^{-5}	5×10^{-5}	5×10^{-5}
Learning Rate (Head)	5×10^{-3}	5×10^{-5}	5×10^{-3}
Warmup Ratio	0.1	0.1	0.1
LR Scheduler	Cosine	Linear	Cosine
Max Gradient Norm	1.0	1.0	1.0

Table 10: Key hyperparameter settings.

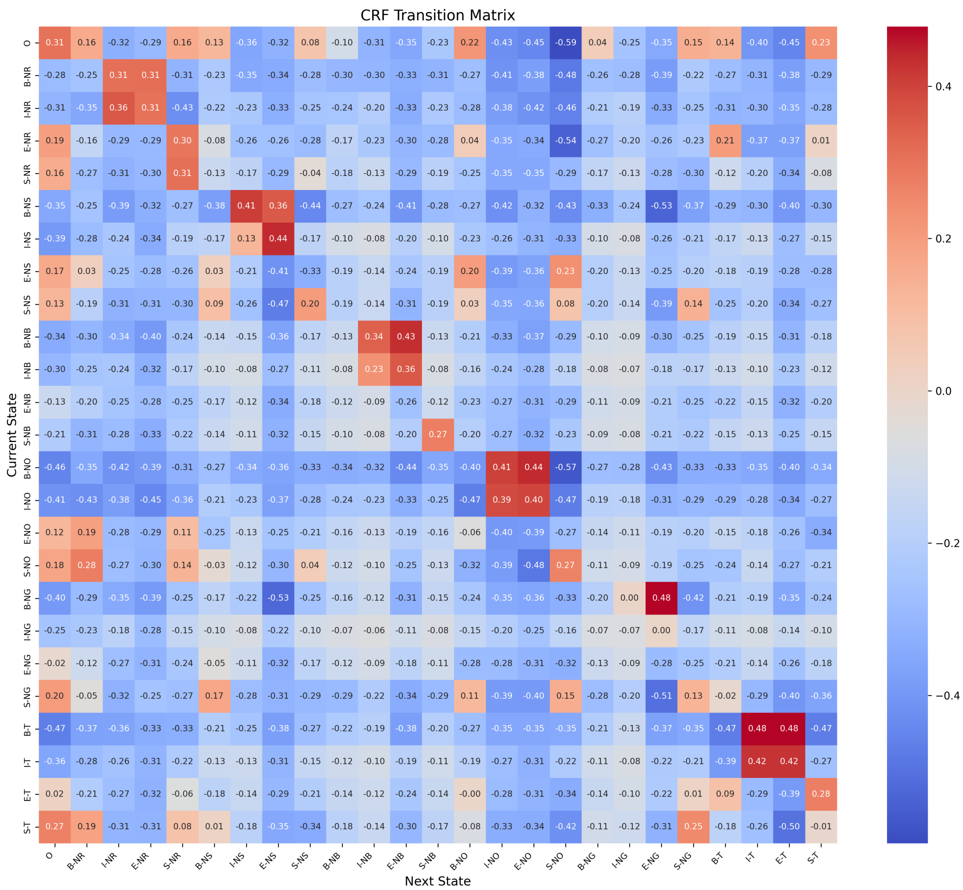

Figure 2: Task A CRF transition matrix (Exp. 3). Rows: current state; columns: next state. Color depth shows transition probability (-0.5 to 0.5).

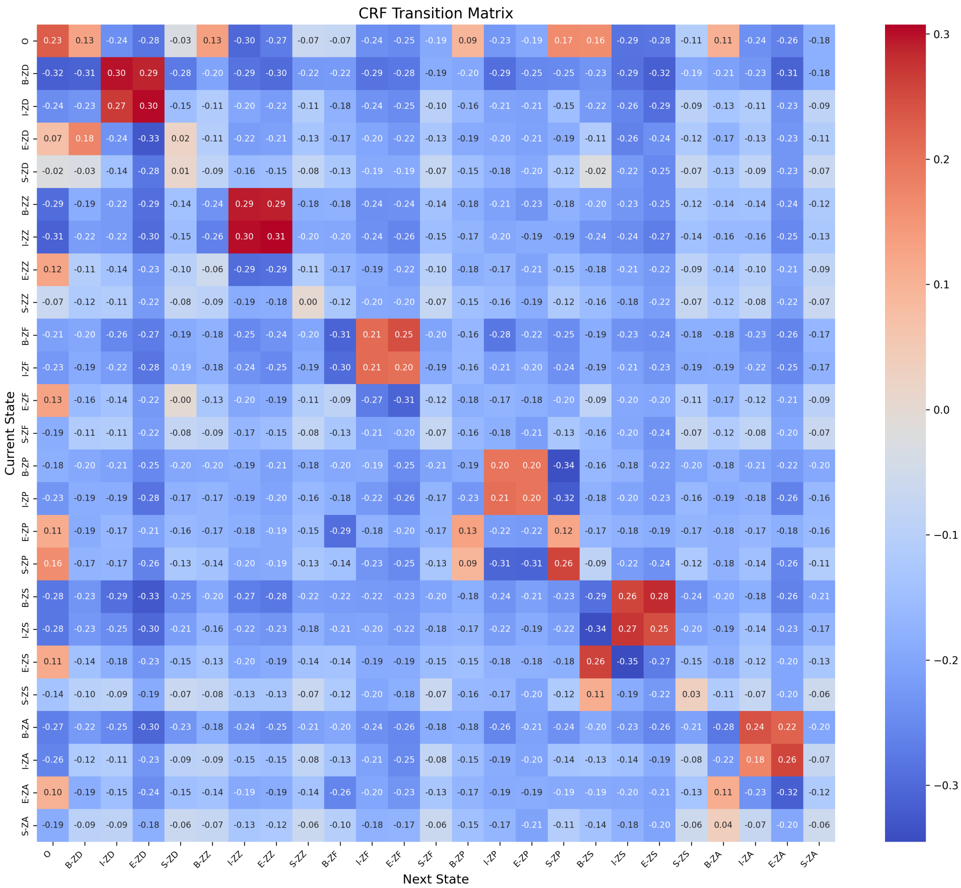

Figure 3: Task C CRF transition matrix (Exp. 3). Rows: current state; columns: next state. Color depth shows transition probability (-0.5 to 0.5).