Integrating Syllable-Level Structure and Local Rescoring for Low-Resource Tibetan Named Entity Recognition

Anonymous ACL submission

Abstract

Low-resource Tibetan named entity recognition is challenging because of limited annotated data and the mismatch between SentencePiece subwords and Tibetan syllable boundaries. In particular, when an entity-final syllable is merged with a following particle or punctuation mark, a conventional subword-level BIO model cannot recover the exact character endpoint, resulting in frequent right-boundary overextension. To address this problem, we propose SEB, a coarse-to-fine method that combines subword-level entity recognition with syllable-level local endpoint calibration. A pretrained encoder and a CRF first identify coarse entity regions and types. Contextualized subword representations are then projected onto a Tibetan syllable grid through character-overlap-weighted pooling. For each detected entity, SEB jointly scores a small set of local syllable endpoints and an explicit no-change action using span semantics, right-side context, entity type, entity length, and local BIO emission compatibility. On the Tibetan test split of CINO-Festival-NER, SEB achieves a character-offset strict span F1 of $74.86 \pm 1.01\%$, compared with $36.90 \pm 0.42\%$ for the subword CRF, where values are means and sample standard deviations over three random seeds. Right-boundary overextensions number 3–5 for SEB and 82–85 for CRF across these runs. These results support the effectiveness of syllable-level endpoint calibration in the evaluated festival-domain setting.

1 Introduction

Named entity recognition (NER) is a fundamental task for information extraction, knowledge graph construction, information retrieval, and question answering. With the rapid growth of Tibetan digital resources, accurately identifying entities such as locations, festivals, events, and cultural objects has become increasingly important for Tibetan knowledge organization and cultural-resource development (Deng and Wu, 2023; Zhou et al., 2024). However, Tibetan NER remains challenging because of limited annotated data, imbalanced entity-type distributions, and insufficient pretrained resources and basic language-processing tools (Sun et al., 2022; Zhou et al., 2024; Yang et al., 2022).

Existing approaches to low-resource NER mainly rely on pretrained representations, cross-lingual knowledge transfer, and structural supervision (Wu and Dredze, 2019; Conneau et al., 2020; Mueller et al., 2020). Although these methods improve entity-region localization and type prediction, their boundary decisions are generally constrained by the subword units produced by pretrained tokenizers. SentencePiece learns subword units directly from raw sentences, whereas Tibetan word segmentation has also been explicitly formulated as a syllable-level tagging problem (Kudo and Richardson, 2018; Liu et al., 2011). This difference is particularly important for Tibetan, whose linguistically meaningful boundaries are more closely associated with syllables than with automatically learned subword units.

A common error in Tibetan NER is right-boundary overextension. Tibetan entities may be followed by case particles, other functional elements, or punctuation, and previous Tibetan NER studies have used case-auxiliary words, punctuation, and syllable or word-position information as boundary-related features (Yu et al., 2010). When the final syllable of an entity and a following element are merged into the same subword, a conventional BIO model cannot place an entity boundary inside that subword. Consequently, the model may correctly locate the entity region but still include the following functional element in the predicted span. Such boundary ambiguity is also a recognized problem in neural NER (Zhu and Li, 2022). Our analysis shows that only approximately half of the gold entities in the test set can be represented exactly by contiguous subword spans, explaining

the substantial gap between token-level sequence performance and character-level exact-match performance.

Simply snapping predicted endpoints to syllable boundaries cannot determine which syllable is the correct endpoint. Conversely, directly adopting syllable-level BIO decoding does not necessarily guarantee accurate entity localization and boundary prediction, particularly under limited training data. Training-time boundary supervision alone is also insufficient because it does not change the final subword-level output space. These observations motivate a coarse-to-fine strategy that preserves subword-level semantic localization while introducing syllable-level structure into final boundary decoding.

To this end, we propose SEB (**Syllable-level End-Boundary Calibration**), a low-resource Tibetan NER method that combines subword-level coarse recognition with syllable-level local endpoint rescoring. A pretrained encoder and a CRF first predict coarse entity regions, types, and initial boundaries (Lafferty et al., 2001). The resulting subword representations are then mapped onto a Tibetan syllable grid, and a small set of candidate endpoints is generated around each initial endpoint. The candidates are rescored using the entity representation, the candidate-final syllable, the following syllable, the entity type, and the entity length. Unlike fixed suffix rules or training-only auxiliary objectives, the learned syllable-level boundary score directly participates in final inference-time decoding. The main contributions of this work are as follows:

1. We analyze low-resource Tibetan NER from the perspective of boundary representability and show that the mismatch between SentencePiece subwords and Tibetan syllable boundaries is a major source of right-boundary overextension.

2. We propose a cross-granularity boundary-calibration method that preserves subword-level entity-region and type predictions while performing local endpoint selection on a syllable grid. This enables the model to recover character-level boundaries that cannot be represented in the original subword BIO output space.

3. Experiments on the Tibetan test set of CINO-Festival-NER show that SEB achieves a mean character-level exact-match F1 of 74.86% over three random seeds, compared with 36.90% for CRF and 50.31% for the evaluated Syl-CRF configuration. SEB outperforms both models in each run and produces 3–5 right-boundary overextensions, compared with 82–85 for CRF. Additional seed-42 comparisons with rule-based correction, training-time structural enhancement, and syllable-boundary snapping provide diagnostic evidence for endpoint calibration.

2 Related Work

2.1 Tibetan Named Entity Recognition

Early Tibetan NER methods relied on linguistic rules, dictionaries, and statistical models such as CRFs, often incorporating case particles and syllable-level boundary cues (Yu et al., 2010). Recent approaches employ neural architectures, multi-feature fusion, and pretrained language models to improve contextual representations and entity recognition (Chen et al., 2024; Liu et al., 2022; Deng et al., 2023b,a; Yang et al., 2022).

Tibetan NER resources have expanded to general and specialized domains, including TibNER, Tibetan medical texts, and festival-related cultural corpora (Zhou et al., 2024; Zhuoma et al., 2022; Deng and Wu, 2023). Nevertheless, limited annotations and mismatches between subword tokenization and Tibetan syllable boundaries remain important challenges. Unlike methods primarily enhancing entity representations, our work explicitly addresses the representability and calibration of entity endpoints.

2.2 Low-Resource Named Entity Recognition

Low-resource NER faces limited supervision, imbalanced entity distributions, and insufficient linguistic resources (Huang et al., 2025, 2021). Existing approaches include lexicon-enhanced models, pretrained multilingual representations, cross-lingual transfer, and structural learning (Wu and Dredze, 2019; Conneau et al., 2020; Mueller et al., 2020). Boundary-aware modeling has also been explored to improve span prediction (Zhu and Li, 2022).

However, improved representations or training-time supervision do not necessarily remove the boundary constraints imposed by subword-level decoding. SEB addresses this limitation by retaining coarse entity recognition while introducing syllable-level endpoint selection during inference.

2.3 Cross-lingual Named Entity Recognition

Cross-lingual NER exploits annotations and representations from high-resource languages to support low-resource tasks (Conneau et al., 2020; Mueller et al., 2020; Fu et al., 2023). For Tibetan, multilingual pretraining and Chinese–Tibetan joint learning offer potential benefits, but differences in writing systems, tokenization, and linguistic structure complicate boundary transfer (Yang et al., 2022).

Our experiments include Chinese–Tibetan joint training as an exploratory comparison. In the evaluated setting, joint training does not improve strict character-span F1 over Tibetan-only training. Rather than drawing general conclusions about cross-lingual transfer, we focus on syllable-level calibration as a targeted solution to the observed boundary-output limitation.

3 Method

3.1 Task Definition and Framework

Given a Tibetan sentence x, named entity recognition aims to predict a set of typed character spans

$$\mathcal{E} = \left\{ (c_s^{(i)}, c_e^{(i)}, t^{(i)}) \right\}_{i=1}^{N},$$

where $[c_s^{(i)}, c_e^{(i)})$ is the half-open character interval of the i-th entity and $t^{(i)}$ is its type. We consider four entity types: FESTIVAL, EVENT, ITEM, and LOCATION. A prediction is correct only when its start offset, end offset, and type exactly match the gold annotation.

SEB follows a coarse-to-fine framework, as shown in Figure 1. A pretrained encoder and a linear-chain CRF first identify coarse entity regions and types on the SentencePiece subword grid. The resulting contextualized subword representations are then projected onto Tibetan syllables. For each detected entity, SEB scores several nearby syllable endpoints together with an explicit no-change action and calibrates only the right endpoint. The predicted start position and entity type remain unchanged.

Let

$$U = (u_1, \ldots, u_L)$$

be the subword sequence and

$$H = (\mathbf{h}_1, \ldots, \mathbf{h}_L), \qquad \mathbf{h}_i \in \mathbb{R}^d,$$

be the contextualized encoder representations. A linear layer produces BIO emission scores $\mathbf{o}_i$, and the CRF is trained using the negative log-likelihood

$$\mathcal{L}_{\mathrm{CRF}}.$$

Viterbi decoding and tokenizer offsets recover the coarse entity set

$$\widehat{\mathcal{E}} = \left\{ (\widehat{c}_s^{(i)}, \widehat{c}_e^{(i)}, \widehat{t}^{(i)}) \right\}_{i=1}^{\widehat{N}}.$$

3.2 Subword-to-Syllable Projection

Before tokenization, each sentence is normalized using Unicode NFC. All tokenizer offsets, syllable boundaries, annotations, and model predictions are computed on the normalized text. A monotonic boundary mapping is retained so that predictions can be converted back to the original text offsets when required.

We segment the normalized sentence into Tibetan orthographic syllables

$$\mathcal{S} = \{ s_j = [a_j, b_j) \}_{j=1}^{M}.$$

A syllable is defined as a maximal nonempty Tibetan content span between tsheg marks, whitespace, or punctuation. Tsheg marks and whitespace are treated as delimiters. Punctuation is retained as right-side context but is not considered a valid syllable endpoint. Special tokens and tokens with empty offset spans are excluded.

Let $[p_i, q_i)$ be the character interval of subword u_i. The overlap weight between syllable s_j and subword u_i is

$$w_{ji} = |[a_j, b_j) \cap [p_i, q_i)|,$$

where $|\cdot|$ denotes the number of overlapping Unicode code points. The syllable representation is obtained by character-overlap-weighted pooling:

$$\mathbf{z}_j = \frac{\sum_{i=1}^{L} w_{ji} \mathbf{h}_i}{\sum_{i=1}^{L} w_{ji}}.$$

This projection handles both tokenization mismatches. If one subword covers multiple syllables, its representation contributes to each syllable according to the corresponding overlap length. If one syllable is split into multiple subwords, all overlapping subword representations are aggregated.

For an endpoint c, the right-context representation $\mathbf{r}(c)$ is obtained from the next punctuation span or the next syllable after skipping tsheg marks and whitespace. A learnable end-of-sentence vector is used when no right context remains.

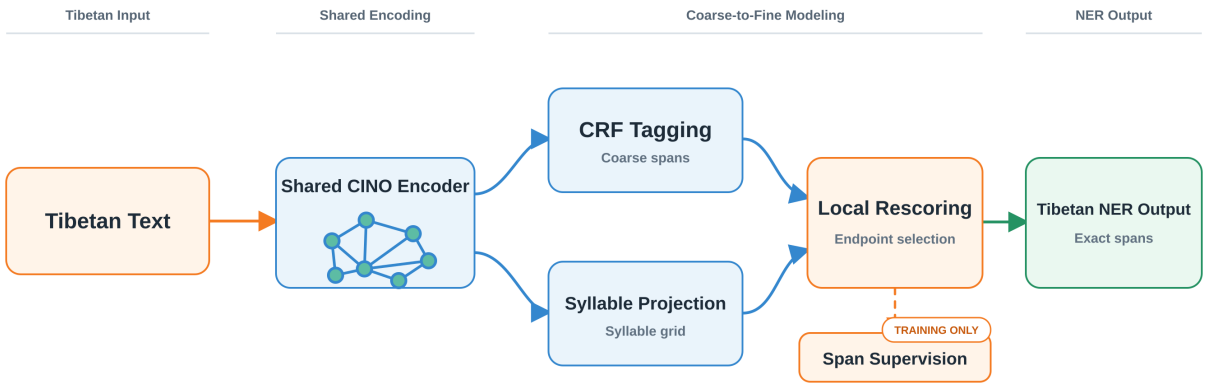


Figure 1: The framework of proposed method.

3.3 Local Endpoint Calibration

Consider a coarse entity

$$\widehat{g} = (\widehat{c}_s, \widehat{c}_e, \widehat{t}).$$

Let k_0 and e_0 be the first and last syllables overlapping the coarse entity. SEB constructs a local endpoint set

$$\mathcal{C}(\widehat{g}) = \{e \mid \max(k_0, e_0 - r) \le e \le \min(M, e_0 + r)\},$$

where $r = 2$. A syllable candidate e corresponds to the character endpoint

$$c(e) = b_e.$$

We additionally introduce a no-change action KEEP, whose endpoint is

$$c(\mathrm{KEEP}) = \widehat{c}_e.$$

The complete action set is

$$\mathcal{A}(\widehat{g}) = \mathcal{C}(\widehat{g}) \cup \{\mathrm{KEEP}\}.$$

Duplicate actions with identical endpoints are removed.

For action a, let $\eta(a)$ be the last syllable overlapping the candidate span

$$[\widehat{c}_s, c(a)).$$

Its mean-pooled span representation is

$$\overline{\mathbf{z}}_a = \frac{1}{\eta(a) - k_0 + 1} \sum_{j=k_0}^{\eta(a)} \mathbf{z}_j.$$

Let

$$\mathbf{v}_{\widehat{t}} \in \mathbb{R}^d$$

be the predicted entity-type embedding and

$$\mathbf{w}_{\ell_a} \in \mathbb{R}^d$$

be the length embedding, where

$$\ell_a = \min(\eta(a) - k_0 + 1, 64).$$

The candidate representation is

$$\mathbf{q}_a = \left[\mathbf{z}_{k_0}; \mathbf{z}_{\eta(a)}; \mathbf{r}(c(a)); \overline{\mathbf{z}}_a; \mathbf{v}_{\widehat{t}}; \mathbf{w}_{\ell_a}\right] \in \mathbb{R}^{6d}.$$

A two-hidden-layer MLP assigns the boundary score

$$S_{\mathrm{bd}}(a) = \mathbf{w}_3^\top \sigma\big(W_2 \sigma(W_1 \mathbf{q}_a + \mathbf{b}_1) + \mathbf{b}_2\big) + b_3 + \mathbb{I}[a = \mathrm{KEEP}] b_{\mathrm{keep}},$$

where

$$W_1 \in \mathbb{R}^{d \times 6d}, \qquad W_2 \in \mathbb{R}^{(d/2) \times d},$$

and σ denotes the GELU activation. Thus, the MLP dimensions are

$$6d \to d \to d/2 \to 1.$$

We also measure whether the subword emission layer supports the candidate span. Let

$$\Gamma(a) = (i_1, \ldots, i_n)$$

be the subwords overlapping $[\widehat{c}_s, c(a))$, and let

$$\ell_{i,y} = \log \mathrm{softmax}(\mathbf{o}_i)_y.$$

The local BIO emission compatibility score is

$$S_{\mathrm{em}}(a) = \frac{1}{n} \left(\ell_{i_1, \mathrm{B}\text{-}\widehat{t}} + \sum_{j=2}^{n} \ell_{i_j, \mathrm{I}\text{-}\widehat{t}} \right).$$

This term is based only on the emission layer and is therefore not referred to as a CRF score.

The final candidate score is

$$S(a) = S_{\mathrm{em}}(a) + \alpha S_{\mathrm{bd}}(a),$$

where α is selected on the validation set. The same combined score is used during both training and inference.

3.4 Training and Inference

The first epoch optimizes only $\mathcal{L}_{\mathrm{CRF}}$. From the second epoch onward, detached Viterbi predictions from the current model are used to construct boundary-training instances. Predicted and gold entities are matched one-to-one when they have the same type and a positive character-span overlap. Maximum character-level intersection over union is used when multiple matches are possible.

Importantly, the local candidate set is centered on the endpoint predicted by the current CRF rather than on the gold endpoint. For a matched predicted–gold pair, the target action is

$$a^\star = \begin{cases} \mathrm{KEEP}, & c_e^g = \widehat{c}_e, \\ e, & c_e^g = b_e \text{ and } e \in \mathcal{C}(\widehat{g}), \\ \varnothing, & \text{otherwise.} \end{cases}$$

Pairs with $a^\star = \varnothing$ are excluded from the boundary loss. This includes gold endpoints outside the local window and non-syllabic endpoints that are not already represented by the no-change action. Such entities still contribute to $\mathcal{L}_{\mathrm{CRF}}$.

For the remaining matched pairs, the boundary loss is

$$\mathcal{L}_{\mathrm{bd}} = -\sum \log \frac{\exp S(a^\star)}{\sum_{a \in \mathcal{A}(\widehat{g})} \exp S(a)}.$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CRF}} + \lambda_{\mathrm{bd}} \mathcal{L}_{\mathrm{bd}},$$

where

$$\lambda_{\mathrm{bd}} = 0.5.$$

At inference time, candidates extending beyond the start of the next coarse entity are removed to prevent entity overlap. The best syllable candidate $\widetilde{a}$ replaces the original endpoint only when

$$S(\widetilde{a}) - S(\mathrm{KEEP}) > \tau.$$

Otherwise, the original endpoint is retained. Let $\widehat{a}$ denote the selected inference action: $\widehat{a} = \widetilde{a}$ if the above condition holds, and $\widehat{a} = \mathrm{KEEP}$ otherwise. We use

$$\alpha = 1.0, \qquad \tau = 0.$$

The calibrated entity is

$$\widehat{g}^\star = \left(\widehat{c}_s, c(\widehat{a}), \widehat{t}\right).$$

Exact duplicate outputs are resolved by retaining the prediction with the higher action score.

For the ablation variants, TRAINONLY uses the joint training objective but returns the original CRF spans at inference time. INFERONLY first trains the CRF model to convergence, freezes the encoder, emission layer, and CRF, and then trains only the boundary scorer and its embeddings on matched CRF predictions.

To avoid repeatedly pooling long spans, we precompute subword–syllable overlaps and prefix sums of syllable representations and BIO emission scores. Candidate span representations and emission compatibility scores can then be obtained in constant time with respect to span length. Each entity has at most $2r+2$ actions, including KEEP, so the additional decoding cost is linear in the number of local candidates when the hidden dimensionality is treated as fixed.

4 Experiments

4.1 Experimental Setup

We conduct experiments on the CINO-Festival-NER dataset. The Chinese and Tibetan corpora share the same domain and entity-type inventory but are not sentence-aligned. The Chinese corpus contains 5,100 training sentences and 637 validation sentences and is used only for joint-training experiments. The Tibetan corpus contains 6,132 training, 766 validation, and 767 test sentences, with 1,854, 229, and 208 annotated entities, respectively. The task includes four entity types: FESTIVAL, EVENT, ITEM, and LOCATION. Their counts in the Tibetan test set are 130, 22, 30, and 26, respectively.

Entities are annotated using character offsets and mapped to SentencePiece subwords through tokenizer offsets. The first covered subword is assigned a `B-type` label and the remaining covered subwords are assigned `I-type` labels. Special tokens and padding positions are excluded from loss computation. All discriminative models use the same entity-type inventory and are evaluated against the same gold character spans.

Table 1: Dataset statistics.

Language	Split	Sentences	Entities
Chinese	Training	5,100	–
Chinese	Validation	637	–
Tibetan	Training	6,132	1,854
Tibetan	Validation	766	229
Tibetan	Test	767	208

Model-specific outputs are converted to character offsets using tokenizer offsets or syllable-to-character mappings, as appropriate.

Character-level exact-match F1 is used as the primary metric. A prediction is considered correct only when its sentence index, start offset, end offset, and entity type exactly match the gold annotation. We also report BIO sequence F1, same-type overlap recall, the number of exact matches, and the numbers of right-boundary overextensions and leftward endpoint contractions.

All subword-level discriminative models use CINO-large-v2 [1] with a linear-chain CRF. SEB is trained for 15 epochs with a maximum sequence length of 256, a batch size of 8, a dropout rate of 0.1, and a learning rate of 2×10^{-5}. The candidate radius is set to $r = 2$, and the hyperparameters are

$$\lambda_{\mathrm{bd}} = 0.5, \qquad \alpha = 1.0, \qquad \tau = 0.$$

Only the CRF objective is optimized in the first epoch, and the boundary loss is introduced from the second epoch onward. The suffix lexicon used by the rule-based baseline is extracted from the training set with a minimum frequency of 3. The default decoding parameters $r = 2$ and $\tau = 0$ are further examined through a fixed-checkpoint, inference-only sensitivity analysis on the Tibetan development split (bo_dev), as reported in Section 4.3. This analysis does not retrain the model with alternative candidate radii. Tables 2, 4, and 6 report the original seed-42 results. We additionally evaluate CRF, Syl-CRF, and SEB with seeds 42, 123, and 2026 on the same Tibetan test split. Table 3 reports the individual Char F1 scores and their mean and sample standard deviation, computed with denominator $n - 1$, where $n = 3$. Unless explicitly stated otherwise, the remaining model-specific results, ablations, and prediction-level error counts refer to the original seed-42 runs.

[1] https://huggingface.co/hfl/cino-large-v2

We compare SEB with subword-level CRF models, training-time structural enhancement methods, rule-based and syllable-based boundary correction methods, and Hunyuan-MT models under 5-shot prompting and LoRA fine-tuning. For the generative baselines, HMT-S, HMT-M, and HMT-L correspond to HY-MT1.5-1.8B, HY-MT1.5-7B, and Hy-MT2-30B-A3B, respectively. The 5-shot experiments use frozen base models with Chinese-language instructions and five in-context examples drawn from the training split. The fine-tuned variants use parameter-efficient LoRA adaptation and are evaluated without in-context demonstrations. The exact checkpoint identities and LoRA configurations are documented in Appendix A.

4.2 Comparison Methods and Overall Results

Table 2 compares generative large language models, subword-level sequence models, training-time structural enhancement methods, rule-based correction, a full syllable-level sequence model, and the proposed method in the original seed-42 setting. The detailed comparisons below refer to this setting; results across three seeds are reported separately in Table 3. Since the generative models do not produce BIO label sequences, Seq F1 is not reported for these models. For Rule, Snap, and SEB, Seq F1 is measured on the subword-level CRF branch before boundary calibration.

Under the 5-shot prompting setting, the generative large language models achieve generally poor character-level exact-match performance. HMT-S and HMT-M each correctly identify only four entities. HMT-L increases the number of exact matches to 42, but produces 734 entity predictions, approximately 3.5 times the number of gold entities. Its overlap recall is substantially higher than its exact-match recall, suggesting that few-shot large language models can locate some relevant regions but have difficulty simultaneously controlling the number of entities, entity types, and character-level boundaries. Supervised fine-tuning with LoRA substantially improves task adaptation. HMT-S-FT achieves a Char F1 of 69.57%. Under the reported adaptation settings, HMT-M-FT and HMT-L-FT obtain Char F1 scores of 63.57% and 56.98%, respectively. However, these comparisons involve differences in model generation, architecture, LoRA configuration, training duration, and input-length budget. Therefore, the observed performance differences should not be interpreted as a controlled analysis of model scaling.

Table 2: Original seed-42 results on the Tibetan test set. F1 and overlap-recall values are percentages. *Right Overext.* and *Left Contr.* denote the numbers of right-boundary overextensions and leftward endpoint contractions, respectively. Three-seed Char F1 results for CRF, Syl-CRF, and SEB are reported separately in Table 3.

Category	Model	Seq F1	Char F1	Overlap Recall	Right Overext.	Left Contr.	Exact Matches
5-shot LLM	HMT-S	–	0.84	–	–	–	4
5-shot LLM	HMT-M	–	0.82	67.31	–	–	4
5-shot LLM	HMT-L	–	8.92	67.79	–	–	42
LoRA fine-tuning	HMT-S-FT	–	69.57	–	–	–	144
LoRA fine-tuning	HMT-M-FT	–	63.57	–	–	–	–
LoRA fine-tuning	HMT-L-FT	–	56.98	–	–	–	102
Subword sequence model	CRF	77.14	37.14	83.17	82	4	78
Training-time structural learning	Span	77.36	35.93	83.20	86	–	76
Training-time structural learning	Path	**78.66**	36.93	82.20	85	–	77
Dynamic hard negatives	Dyn	75.06	36.10	82.70	82	–	76
Rule-based correction	Rule	77.14	63.81	83.17	14	16	134
Syllable-boundary snapping	Snap	77.14	38.10	83.17	82	2	80
Syllable sequence model	Syl-CRF	–	44.39	78.85	**0**	69	91
Ours	**SEB**	77.93	**75.59**	**84.13**	3	**1**	**161**

The base CRF achieves a Seq F1 of 77.14% and an overlap recall of 83.17%, but its Char F1 is only 37.14%. This discrepancy indicates that the model can localize many entity regions but has difficulty recovering their exact character-level endpoints. Span, Path, and Dyn strengthen structural supervision through span learning, path ranking, and dynamic hard-negative sampling, respectively, but their Char F1 scores remain around 36%. In particular, although Path obtains the highest Seq F1, it still produces 85 right-boundary overextensions. This result shows that introducing structural or boundary information only during training does not guarantee that the learned knowledge will alter the final CRF outputs.

Snap improves Char F1 only slightly, from 37.14% to 38.10%, while the number of right-boundary overextensions remains unchanged at 82. This demonstrates that mechanically snapping a prediction to a valid syllable boundary does not resolve the more important question of which syllable should serve as the true entity endpoint. Syl-CRF performs BIO decoding directly on the syllable grid. In the seed-42 run, it obtains a Char F1 of 44.39% and an overlap recall of 78.85%, with no counted right-boundary overextensions but 69 leftward endpoint contractions. Across seeds 42, 123, and 2026, its Char F1 scores are 44.39%, 42.65%, and 63.89%, while the corresponding contraction counts are 69, 75, and 32. No right-boundary overextensions are counted in any of these runs. Thus, the evaluated Syl-CRF configuration shows substantial variation in exact-match performance, and its low seed-42 score should not be interpreted as an intrinsic limitation of syllable-level sequence modeling.

Rule increases Char F1 to 63.81% and reduces right-boundary overextensions from 82 to 14, confirming that frequently occurring functional elements following entities exhibit a certain degree of regularity. However, Rule also produces 16 leftward endpoint contractions, indicating that a fixed suffix lexicon may incorrectly remove valid components of an entity. In comparison, SEB achieves a Char F1 of 75.59%, with 161 exact-match entities, only three right-boundary overextensions, and only one leftward contraction. Relative to CRF, SEB improves overlap recall by only 0.96 percentage points, whereas exact-match recall increases from 37.50% to 77.40%. Its main effect is therefore to convert predictions that have already localized the correct entity region into exact entity spans, rather than merely increasing entity-region coverage.

In the seed-42 comparison, SEB exceeds Rule by 11.78 percentage points in Char F1 and HMT-S-FT by 6.02 percentage points. These differences are single-run comparisons, rather than estimates of gains averaged over seeds. Together with the training/inference ablations, they support the usefulness of combining subword-level coarse recognition with local syllable-level endpoint calibration.

Results across random seeds. Table 3 reports repeated runs for CRF, Syl-CRF, and SEB using seeds 42, 123, and 2026. SEB achieves a mean Char F1 of $74.86 \pm 1.01\%$, compared with $36.90 \pm 0.42\%$ for CRF and $50.31 \pm 11.79\%$ for Syl-CRF, where the dispersion values are sample standard deviations. SEB obtains higher Char F1

Table 3: Strict character-span F1 on the Tibetan test set across three random seeds. All values are percentages. The final column reports the mean and sample standard deviation over the three displayed runs.

Model	Seed 42	Seed 123	Seed 2026	Mean ± SD
CRF	37.14	36.41	37.14	36.90 ± 0.42
Syl-CRF	44.39	42.65	63.89	50.31 ± 11.79
SEB	**75.59**	**75.29**	**73.71**	$\mathbf{74.86 \pm 1.01}$

Table 4: Main ablation results for SEB in the original seed-42 setting. All F1 values are percentages.

Model	Joint Train	Inference Rescore	Seq F1	Char F1	Right Overext.
CRF	×	×	77.14	37.14	82
TrainOnly	✓	×	77.93	37.09	85
InferOnly	×	✓	77.14	72.38	9
SEB	✓	✓	**77.93**	**75.59**	**3**

than both models at each of the three seeds. The mean of the seed-wise SEB–CRF differences is 37.97 percentage points. Right-boundary overextensions number 82, 85, and 82 for CRF and 3, 4, and 5 for SEB, respectively. These observations show that the large gain over the subword CRF recurs across the three evaluated seeds. SEB also exhibits less observed variation than the evaluated Syl-CRF configuration. The repeated runs characterize training variability on a fixed test split; they do not establish statistical significance or generalization to other datasets.

4.3 Ablation and Training-Configuration Analysis

To distinguish the effects of the training-time auxiliary boundary loss and inference-time local rescoring, we compare CRF, TrainOnly, InferOnly, and the complete SEB model using the same base architecture in the original seed-42 setting. TrainOnly uses auxiliary boundary supervision during training but directly outputs the original CRF predictions at inference time. InferOnly leaves CRF training unchanged and applies syllable-level local rescoring only during inference.

TrainOnly increases Seq F1 from 77.14% to 77.93%, but slightly decreases Char F1 to 37.09% and increases the number of right-boundary overextensions to 85. This indicates that the auxiliary boundary loss can influence the learned subword representations, but cannot readily produce exact character-level boundaries within the original BIO output space. InferOnly freezes a separately trained base model and trains only the endpoint scorer; endpoint calibration is subsequently applied at inference. By introducing syllable-level candidates and local rescoring only during inference, it already increases Char F1 to 72.38% and reduces right-boundary overextensions to nine. The complete SEB model further improves Char F1 by 3.21 percentage points and reduces right-boundary overextensions to three. These results indicate that inference-time syllable-level endpoint calibration, which combines an expanded boundary candidate space with local rescoring, accounts for most of the observed improvement over the original CRF baseline. Auxiliary boundary training is nearly ineffective when used alone, but provides more suitable boundary representations for inference-time candidate scoring.

Restricting local endpoint rescoring to the SentencePiece subword grid cannot resolve boundaries that fall inside individual subwords. Even when the scorer identifies an inappropriate endpoint, the correct boundary may remain unavailable in the candidate space. In contrast, syllable-level calibration expands the representable boundary space, enabling the recovery of entity spans that cannot be expressed by conventional subword-level BIO decoding. The syllable grid therefore not only introduces target-language structural information, but also expands the representable space of final entity boundaries.

In a separate seed-42 comparison, we evaluate Tibetan-only training and Chinese–Tibetan joint training. Under Tibetan-only training, the Char F1 scores of CRF and SEB are 37.14% and 75.59%, respectively. Under Chinese–Tibetan joint training, the corresponding scores are 36.15% and 74.58%. The improvements of SEB over CRF under the two settings are therefore 38.45 and 38.43 percentage points, respectively. Their descriptive interaction effect is -0.02.

In this seed-42 comparison, SEB yields similar improvements over CRF with Tibetan-only and Chinese–Tibetan joint training, whereas Chinese auxiliary supervision does not improve Char F1 for either model. This descriptive result is consis-

Table 5: Inference-time sensitivity of the candidate radius r and KEEP margin τ on the Tibetan development split. All runs use a fixed SEB checkpoint without retraining.

r	τ	Dev Char F1
1	0.0	71.12
2	**0.0**	**75.00**
3	0.0	75.00
2	0.5	75.00
2	1.0	75.00

Table 6: Seed-42 exact-match results on subsets defined by subword-boundary representability. Subset sizes depend on the dataset and tokenizer; the CRF and SEB exact-match counts are specific to these runs.

Subset	Entities	CRF Exact	SEB Exact
Subword-representable	105	78	78
Subword-unrepresentable	103	0	**83**
Overall	208	78	**161**

tent with endpoint calibration addressing a Tibetan boundary-output limitation that remains under both training settings. The joint-training experiment does not establish the stability of cross-lingual effects across random seeds.

Sensitivity to decoding parameters. To examine the choice of the candidate radius r and the KEEP margin τ, we perform an inference-only sensitivity analysis using a fixed trained SEB checkpoint (S1). Encoder outputs are cached, and only the inference-time candidate radius or decision margin is varied, without retraining the model.

We evaluate $r \in \{1, 2, 3\}$ with $\tau = 0$ and $\tau \in \{0, 0.5, 1.0\}$ with $r = 2$. All parameter comparisons are conducted on the Tibetan development split (`bo_dev`).

As shown in Table 5, increasing the candidate radius from one to two improves development Char F1 from 71.12% to 75.00%. Increasing the radius further to three produces no additional F1 improvement at the reported precision. We therefore retain the smaller radius $r = 2$.

When $r = 2$, all three evaluated values of τ yield the same reported development Char F1. Accordingly, we retain the default $\tau = 0$.

These findings support the selected decoding configuration for the evaluated checkpoint. Since the experiments vary inference-time parameters without retraining, they do not establish optimality across training configurations or random seeds.

4.4 Boundary Representability and Error Analysis

Boundary representability is determined by the fixed dataset, normalization procedure, and tokenizer. The model-specific exact-match counts and error-attribution results in this subsection refer to the original seed-42 runs; the additional runs are summarized in Table 3.

To explain the substantial gap between Seq F1 and Char F1, we analyze the representability of gold entity boundaries on the subword and syllable grids. In the test set, 51.4% of entity endpoints coincide with SentencePiece subword boundaries, whereas 99.0% coincide with Tibetan syllable boundaries. When both entity start and end positions are considered, only 105 of the 208 entities can be represented exactly by contiguous subword spans, accounting for 50.5% of the test entities. The remaining 103 entities cannot be recovered as exact character-level spans within the original subword-level BIO output space.

Among the 105 entities that can be represented exactly by contiguous subword spans, both CRF and SEB correctly recognize 78. These equal counts do not necessarily indicate identical correctly recognized entities, as individual predictions may differ between the two systems. Among the 103 subword-unrepresentable entities, CRF cannot produce any exact-match span, whereas SEB recovers 83. The net increase of 83 exact-match entities is largely driven by the subword-unrepresentable subset, where SEB successfully recovered 83 entities. This finding shows that SEB does not merely produce a general improvement in entity-type recognition. Instead, by transferring the local endpoint decision to the syllable grid, it overcomes the boundary constraints imposed by the subword-level output space.

Boundary representability is, however, only a necessary condition for exact recognition, not a sufficient one. Although Snap restricts its outputs to syllable boundaries, it cannot determine which syllable corresponds to the correct endpoint. In the seed-42 comparison, Syl-CRF exhibits more leftward endpoint contractions and lower entity-region overlap recall than the subword-level CRF baseline. Its additional runs show substantial variation in strict span F1, as reported in Table 3. SEB searches over only a small number of candidates around each coarse entity produced by the subword-level CRF. It can therefore preserve the semantic local-

ization ability of the subword representations while exploiting the syllable grid for exact boundary recovery.

The base CRF obtains an overlap recall of 83.17%, corresponding to approximately 173 gold entities, but only 78 predictions are exact matches. Thus, approximately 95 entities have already been localized to the relevant region but do not receive exact boundaries. Among these failures, 82 are right-boundary overextensions, accounting for 86.3%. Such overextensions are concentrated primarily among FESTIVAL entities followed by case particles, demonstrative elements, connective elements, or punctuation. In some cases, SentencePiece merges the final syllable of an entity and the following element into the same subword, preventing a conventional BIO model from placing an endpoint between them.

Rule repairs 68 right-boundary overextensions, 67 of which become exact-match entities. However, almost all of these corrections involve suffixes included in the fixed lexicon. The method additionally introduces 12 leftward endpoint contractions and corrupts 11 entities that were originally predicted correctly. In the paired comparison between the CRF baseline and the complete SEB model, 79 entities exhibiting right-boundary overextension under CRF are correctly recognized by SEB. SEB also introduces one new leftward contraction and changes three originally correct predictions into incorrect ones. The 79 recovered entities correspond to 95.2% of the net increase of 83 exact matches. This comparison characterizes the changes between the two systems but does not isolate the contribution of inference-time rescoring from joint training.

In addition, the base CRF produces 61 right-boundary overextensions for FESTIVAL entities, whereas SEB reduces this number to two. This result demonstrates that the proposed method is particularly effective at separating festival names from following functional elements. The analysis also confirms that SEB is not merely a more complicated implementation of fixed particle stripping. Instead, it makes context-dependent endpoint decisions based on the internal entity semantics, entity type, and right-side context.

Overall, the subword-level sequence model performs relatively well in coarse-grained entity-region localization and entity-type prediction, but many gold boundaries cannot be expressed on its output grid. Simple syllable-boundary snapping cannot determine the correct endpoint, while the evaluated Syl-CRF baseline shows lower region-localization performance in the seed-42 comparison. By retaining the subword-level coarse-recognition results and refining only the entity endpoints on a local syllable grid, SEB substantially improves character-level entity integrity while preserving overlap recall. Its main limitation is that it calibrates only the right endpoints of entities that have already been detected. It therefore cannot directly address completely missed entities, entity-type errors, or left-boundary errors.

5 Conclusion

This work investigates right-boundary overextension in low-resource Tibetan named entity recognition, which arises from the mismatch between subword-level output boundaries and Tibetan syllable structure. To address the difficulty faced by conventional subword-level sequence models in recovering exact character-level boundaries, we propose a coarse-to-fine method that combines subword-level coarse recognition with syllable-level local endpoint rescoring. The proposed method preserves the entity-region localization and entity-type prediction capabilities of the subword model while allowing syllable-level structure to participate directly in final boundary decisions. Across three random seeds on the Tibetan test split of CINO-Festival-NER, SEB achieves a character-level exact-match F1 of $74.86 \pm 1.01\%$, compared with $36.90 \pm 0.42\%$ for the subword CRF and $50.31 \pm 11.79\%$ for the evaluated Syl-CRF configuration. SEB obtains higher Char F1 than both models in each evaluated run, while reducing right-boundary overextensions from 82–85 for CRF to 3–5. The fixed-checkpoint development-set sensitivity analysis also supports retaining the default decoding parameters $r = 2$ and $\tau = 0$. Together with the boundary representability analysis, these results support syllable-level local endpoint calibration as an effective approach for improving strict entity-span recognition in the evaluated festival-domain setting. In future work, we will investigate the joint calibration of entity start and end boundaries, improve the handling of missed entities, entity-type errors, and complex multi-syllable entities, and further evaluate the robustness and generalizability of the proposed method on additional Tibetan-domain datasets and other low-resource languages with similar boundary structures.

To ensure the reproducibility of our experiments,

we have detailed the model architectures, hyperparameters, and evaluation metrics in Section 4.1.

Limitations

This study focuses on right-boundary calibration in low-resource Tibetan named entity recognition, where mismatches between subword tokenization and syllable structure frequently affect exact span prediction. The proposed method is designed to refine entity endpoints while preserving the coarse entity regions and types identified by the underlying sequence model.

Our experiments are conducted on the Tibetan split of CINO-Festival-NER, with repeated evaluations across three random seeds for the main models. Although the results consistently demonstrate the effectiveness of local syllable-level rescoring in this setting, further evaluation on datasets covering broader domains and entity distributions would provide additional evidence of generalizability. The proposed framework also provides a foundation for exploring more flexible boundary calibration strategies.

References

E J Chen, M Zhou, B Chen, et al. 2024. Research on tibetan named entity recognition using multi-feature fusion representation. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 340–351.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Junjie Deng, Long Chen, Yan Zhang, Yuan Sun, and Xiaobin Zhao. 2023a. Tikem: Knowledge enhanced tibetan pre-trained language model. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023b. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Yuyang Deng and Dan Wu. 2023. Chinese-tibetan bilingual named entity recognition for traditional tibetan festivals. *Data Analysis and Knowledge Discovery*, 7(7):125–135.

Yingwen Fu, Nankai Lin, Boyu Chen, Ziyu Yang, and Shengyi Jiang. 2023. Cross-lingual named entity recognition for heterogenous languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:371–382.

Cheng Huang, Nyima Tashi, Fan Gao, Yutong Liu, Jiahao Li, Hao Tian, Siyang Jiang, Thupten Tsering, Ban Ma-bao, Renzeg Duojie, et al. 2025. Tibetan language and ai: A comprehensive survey of resources, methods and challenges. *arXiv preprint arXiv:2510.19144*.

J Huang, C Li, K Subudhi, et al. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 10408–10423.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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 18th International Conference on Machine Learning*, pages 282–289, Williamstown, MA, USA. Morgan Kaufmann.

Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu, and Yeping He. 2011. Tibetan word segmentation as syllable tagging using conditional random field. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 168–177, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of transfer in multilingual named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8093–8104, Online. Association for Computational Linguistics.

Yuan Sun, Sisi Liu, Junjie Deng, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. *arXiv preprint arXiv:2205.07303*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Computational Linguistics*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hongzhi Yu, Tao Jiang, and Ning Ma. 2010. Named entity recognition for tibetan texts using case-auxiliary grammars. In *Proceedings of the International MultiConference of Engineers and Computer Scientists 2010*, volume 1, pages 601–604, Hong Kong. International Association of Engineers.

Maoke Zhou, Cairang Ejian, Caidan Daoji, Xiaoke Qi, and Xiaobing Zhao. 2024. A dataset of tibetan named entity recognition—tibner. *China Scientific Data*, 9(4):1–11.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

Cuo Zhuoma, Jia Cairang, Duanzhu Sangjie, Zhuoma Yangmao, and Zhaxi Zhuoma. 2022. Tibetan medical named entity recognition study for tibetan clinical electronic medical records. In *Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)*, volume 12506, pages 1715–1722. SPIE.

A Implementation and Reproducibility Details

A.1 Hunyuan-MT Checkpoints and LoRA Adaptation

The generative baselines use three checkpoints from the Tencent Hunyuan-MT model families:

- HMT-S: `HY-MT1.5-1.8B`.
- HMT-M: `HY-MT1.5-7B`.
- HMT-L: `Hy-MT2-30B-A3B`.

The 5-shot variants use frozen base models. The LoRA fine-tuned variants use PEFT 0.19.1 with `task_type=CAUSAL_LM`, `bias=none`, and `lora_dropout=0.05`.

For HMT-S-FT and HMT-M-FT, LoRA is applied to the query, key, value, and output projection modules, together with the gate, up, and down projection modules.

For HMT-L-FT, LoRA is applied only to the query, key, value, and output projection modules.

All fine-tuning runs use only the 6,132 Tibetan training sentences from `bo_train`. Checkpoints are saved after each epoch, and the final adapter is used for evaluation without development-based early stopping. The fine-tuned variants are evaluated without in-context demonstrations.

Table 7: LoRA configurations for Hunyuan-MT baselines.

Setting	S-FT	M-FT	L-FT
Rank	16	16	8
Alpha	32	32	16
LR	2×10^{-4}	10^{-4}	10^{-4}
Epochs	2	2	1
Batch	4/4	1/16	1/8
Max length	1024	1024	512