Entropy-Driven Pre-tokenization for Byte Pair Encoding

Anonymous Authors[1]

Abstract

Byte Pair Encoding (BPE) has become a widely adopted subword tokenization method in modern pretrained language models due to its simplicity and strong empirical performance across downstream tasks. However, applying BPE to unsegmented languages such as Chinese presents significant challenges, as its frequency-driven merge operations are agnostic to linguistic boundaries. To address this, we propose two entropy-informed pre-tokenization strategies that guide BPE segmentation using unsupervised information-theoretic cues. The first approach uses pointwise mutual information and left/right entropy to identify coherent character spans, while the second leverages predictive entropy derived from a pretrained GPT-2 model to detect boundary uncertainty. We evaluate both methods on a subset of the PKU corpus (Emerson, 2005) and demonstrate substantial improvements in segmentation precision, recall, and F1 score compared to standard BPE. Our results suggest that entropy-guided pre-tokenization not only enhances alignment with gold-standard linguistic units but also offers a promising direction for improving tokenization quality in low-resource and multilingual settings.

1. Introduction

Modern pretrained language models often rely on BPE as a core tokenization strategy because it is simple and effective, leading to its widespread adoption. (Gage, 1994; Sennrich et al., 2016). BPE iteratively merges frequent character pairs to construct a compact vocabulary, which enables the model to capture meaningful subword units and represent a wide range of linguistic phenomena across different languages. Its success in English and many Indo-European languages has made it the default tokenizer in many large-scale models such as GPT (Radford et al., 2019a), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019).

However, the application of BPE to Chinese presents unique challenges. Unlike alphabetic languages, Chinese lacks explicit word boundaries (e.g., spaces), and each character can serve as a standalone word or part of multi-character words with varying syntactic or semantic roles (Sproat et al., 1994). When BPE is applied naively – treating each character as a base unit and relying solely on frequency-driven merging – it often fails to capture the true internal structure of Chinese words. As a result, the created token sequences may not align with linguistically meaningful units, which can degrade downstream performance and interpretability. To this end, we introduce and evaluate two distinct entropy-driven pre-tokenization strategies for BPE:

- Statistical Methods: We compute pointwise mutual information (PMI) and left/right entropy to identify potential segmentation boundaries based on local co-occurrence strength and contextual diversity.
- Auto-regressive LLM-based Methods: We use a pretrained GPT-2 model (Radford et al., 2019a) to estimate token-level predictive entropy, leveraging model uncertainty to guide boundary detection.

We examine each approach independently and analyze their effect on BPE vocabulary learning and downstream segmentation quality. We compare both entropy-informed BPE variants to a standard frequency-driven BPE baseline, highlighting differences in tokenization granularity, compression efficiency, and alignment with gold-standard Chinese word segmentation. Our findings demonstrate that incorporating entropy-based pre-tokenization can reshape BPE token structure, offering new insights into subword modeling in unsegmented scripts.

2. Related Works

2.1. Subword Tokenization via BPE

Byte-Pair Encoding (BPE) was introduced in data-compression research, then repurposed for open-vocabulary neural translation by Sennrich et al.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(2016). Its deterministic merge procedure yields compact, variable-length subword units that now underpin GPT-series models, RoBERTa, BART, among others. SentencePiece (Kudo & Richardson, 2018) generalized BPE to language-independent, raw-text training, and popularized the alternative unigram language model segmentation that performs stochastic sampling for regularization. Building on this foundation, subsequent work sought to increase flexibility in tokenization approaches. For instance, BPE-Dropout (Provilkov et al., 2021) injects controlled noise during training to expose models to alternative segmentations, while Charformer (Tay et al., 2021) learns gradient-based block selection end-to-end, eliminating the static pre-processing stage.

2.2. Pre-tokenization Constraints

Most subword algorithms, including BPE, operate on pretokenized input sequences, where initial boundaries, typically defined by whitespace, act as hard constraints on the merge process. In fact, only very recently has research investigated removing these constraints, (Schmidt et al., 2025; Liu et al., 2025). The merge space produced by BPE is thus bounded by this pre-tokenization constraint. While whitespace provides reliable word boundaries in alphabetic scripts (Manning & Schütze, 1999), Chinese lacks such delimiters. Traditional Chinese NLP pipelines often rely on heuristic rules to segment text, compensating for the absence of explicit word boundaries in written Chinese. These heuristics commonly involve dictionary-based matching, statistical modeling, or rule-based systems to determine segmentation points (Huang & Liu, 1997; Xue, 2003). While such approaches have achieved reasonable success, they struggle with ambiguities and out-of-vocabulary words, often resulting in inconsistent or fragmented tokenization. Dictionary-based methods are particularly sensitive to lexicon coverage, failing on novel terms, while statistical methods require extensively annotated corpora and may lack domain adaptability. Rule-based systems, although deterministic, often lack the flexibility to accommodate linguistic variability.

2.3. Information-Theoretic Cues

Information theory offers language-agnostic cues for identifying word boundaries. Early work suggested that statistical irregularities, such as peaks in mutual information and entropy, correlate with morphological structure (Harris, 1968). This insight was later formalized through models based on branching entropy, which measures the uncertainty of character sequences in left and right contexts to identify likely segmentation points (Tanaka-Ishii, 2005). Such entropy-based methods have enabled effective unsupervised word segmentation, particularly in languages like Chinese (Jin & Tanaka-Ishii, 2006). To the best of our knowledge, however, these techniques have not been widely adopted in modern tokenizers for large language models.

2.4. Morphological and Sub-character Tokenization

Beyond word-level segmentation, recent studies exploit the internal structure of logographic scripts. Sub-character tokenization decomposes characters into glyph components or phonetic codes, allowing parameter sharing across radicals and improving generalization on rare forms (Chen & Deng, 2020). Mega-tokenization extends the opposite axis, merging highly frequent multi-character expressions into mega units that preserve semantics over longer spans. Such approaches demand additional symbol inventories or external alignment, yet they alleviate the granularity mismatch between BPE vocabularies and Chinese morphology.

2.5. Byte-Level Tokenization Alternatives

An emerging line of research in language modeling seeks to eliminate reliance on fixed subword vocabularies by operating directly on raw byte sequences. Byte Latent Transformer (BLT) (Pagnoni et al., 2024) exemplifies this approach by dynamically segmenting input into variable-length byte patches, with boundaries guided by next-byte entropy from a lightweight language model. This enables adaptive computation and has shown performance comparable to BPE-based models at the 8B scale.

Other notable byte-level models include ByT5 (Xue et al., 2022), which processes UTF-8 byte sequences directly, removing the need for explicit tokenization. It shows strong multilingual performance and robustness to noise, particularly in low-resource settings and languages under-served by subword vocabularies. Similarly, CANINE (Clark et al., 2021) operates at the character level and introduces a downsampling mechanism to manage sequence length while preserving linguistic detail. Both models perform competitively with subword-based approaches, underscoring the potential of tokenization-free pipelines for language-agnostic applications.

2.6. Byte Pair Encoding and Pre-tokenization

BPE is a widely used tokenization algorithm designed to construct a compact subword vocabulary from unlabeled text (Sennrich et al., 2016). Given a corpus and a target vocabulary size T, BPE applies a deterministic, greedy merging procedure that identifies and merges the most frequent adjacent symbol pairs until the vocabulary size is met.

Traditional BPE pre-tokenization relies on spaces, reflecting assumptions typical of Indo-European languages. However, in unsegmented languages like Chinese, this approach often leads to suboptimal merges, as BPE operates on units that misalign with meaningful linguistic spans.

2.7. Entropy-Driven Pre-tokenization

To address this mismatch, we introduce an entropy-based pre-tokenization step prior to standard BPE. Rather than assuming fixed boundaries, we first segment the raw Chinese corpus using unsupervised entropy signals—either via statistical co-occurrence (e.g., PMI and left/right entropy) or neural predictive entropy from an auto-regressive language model. These signals produce segmentation points based on information-theoretic cues, yielding boundary candidates that reflect distributional irregularities in character sequences.

Once segmented, we apply standard whitespace-delimited BPE on this preprocessed corpus. This preserves the efficiency and scalability of BPE while biasing token construction toward linguistically meaningful units. In contrast to conventional BPE, our approach explicitly inserts structure into the initial token stream, guiding merge operations with signals grounded in either local statistics or model uncertainty.

3. Methods

Motivated by information-theoretic signals and the structural challenges of unsegmented scripts like Chinese, we investigate two pre-tokenization strategies for Chinese BPE based on entropy signals. Both aim to identify linguistically plausible token boundaries prior to subword vocabulary construction. The first method uses symbolic statistical measures, while the second leverages uncertainty estimates from an auto-regressive language model. In both cases, the resulting pre-segmented corpus is passed to a standard BPE tokenizer with whitespace-based pre-tokenization, thereby constraining merges to occur only within identified spans.

3.1. Statistical-based Pre-tokenization

The statistical method draws inspiration from unsupervised word-segmentation literature (Jiang et al., 2022). We enumerate every possible n-gram ($2 \leq n \leq n_{\max}$) in the corpus and we assign a utility score to each individual n-gram (multi-character span) occurrence. Specifically, every n-gram occurrence is treated as a candidate w, and its utility score is based on a combination of internal cohesion and contextual separability:

$$U_{\text{stat}}(w) = \min_{(c_i, c_{i+1}) \subset w} \text{PMI}(c_i, c_{i+1}) + \lambda \min\big(H_{\text{left}}(w), H_{\text{right}}(w)\big) \quad (1)$$

Internal cohesion: Pointwise Mutual Information (PMI) measures the associative strength between two adjacent characters:

$$\text{PMI}(c_i, c_{i+1}) = \log \frac{f(c_i c_{i+1})\, T}{f(c_i)\, f(c_{i+1})}, \quad (2)$$

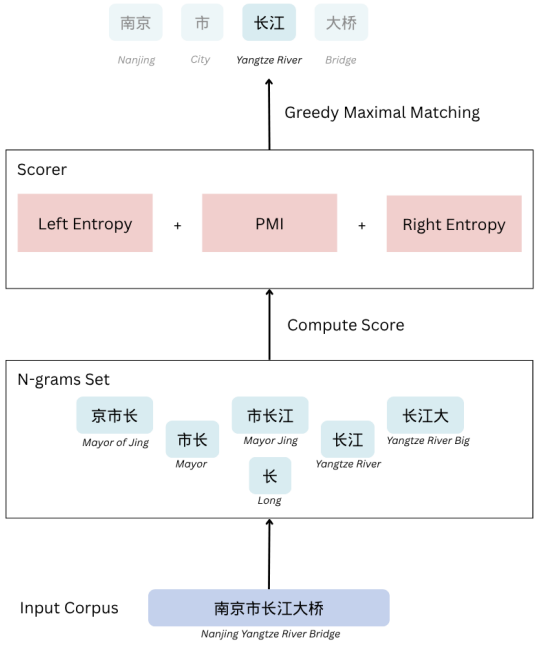

Figure 1. Overview of the statistical method. The algorithm applies greedy maximal matching based on scores from Left Entropy, PMI, and Right Entropy to select meaningful n-grams, producing the final segmentation.

where $f(\cdot)$ denotes corpus frequency and T is the total number of character tokens. A large PMI indicates that the pair co-occurs far more often than chance, suggesting that they should remain in the same token. We take the minimum PMI among all adjacent pairs inside w so that a single weak link can lower the overall cohesion score, preventing loosely connected parts from being merged.

Contextual separability: Left and right entropy quantifies how diversely a span appears with its immediate context:

$$H_{\text{left}}(w) = -\sum_{l} P(l \mid w) \log P(l \mid w), \quad (3)$$

$$H_{\text{right}}(w) = -\sum_{r} P(r \mid w) \log P(r \mid w), \quad (4)$$

with $P(l \mid w) = \frac{f(lw)}{\sum_{l'} f(l'w)}$ and $P(r \mid w) = \frac{f(wr)}{\sum_{r'} f(wr')}$. A large entropy means the span occurs with many different neighbors, signaling a plausible word boundary. We again take the minimum of ljeft and right entropies so that a single side with low diversity keeps w from being split prematurely.

Balancing the two terms: As illustrated in Figure 2, the ranges of PMI and entropy differ significantly, often by several orders of magnitude. This disparity can cause one score to dominate the utility score if left unadjusted. To address

this imbalance, we introduce a scaling hyperparameter, λ, which serves to modulate the relative contributions of both terms. By tuning λ, we can control the extent to which each term influences the overall optimization process, ensuring neither overwhelms the other.

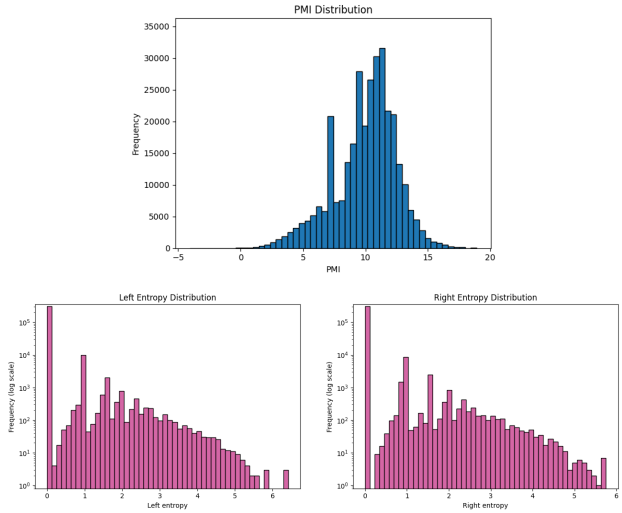

Figure 2. Distributions of statistical features from the PKU dataset. Top: PMI distribution showing a peak in the range of 9–12. Bottom: Left and Right entropy distributions, shown on a log scale, are heavily skewed toward zero, indicating that many characters consistently occur within fixed contexts.

Greedy conflict-free selection: After assigning utility scores to candidate spans, the sentence is processed in a left-to-right traversal. At each character position, the span with the highest score that begins at that location is selected. Once a span is chosen, all characters it encompasses are marked as fixed, thereby preventing subsequent spans from overlapping with it. This single-pass procedure produces a non-overlapping segmentation that optimizes local utility without requiring backtracking. Characters not included in any multi-character span are treated as singleton segments, thereby avoiding any out-of-vocabulary (OOV) tokens at inference time. The resulting segmented corpus is subsequently provided as input to a standard BPE tokenizer, which performs token merges exclusively within segments delineated by spaces.

3.2. Auto-regressive LLM-based Pre-tokenization

The second approach estimates token boundaries using predictive uncertainty derived from a pretrained auto-regressive transformer model. At each character position t, we compute the conditional entropy of the next token given the left context:

$$H(x_t \mid x_{<t}) = -\sum_{x \in \mathcal{V}} P(x_t = x \mid x_{<t}) \log P(x_t = x \mid x_{<t}) \tag{5}$$

The conditional entropy measures the model's uncertainty about the next character: when the value is low, the upcoming symbol is highly predictable from its left context, implying that the sequence is continuing a cohesive lexical unit; conversely, sharp spikes in entropy indicate a sudden drop in predictability, signaling a likely semantic shift and the onset of a new word. To obtain robust estimates, the raw entropy sequence is smoothed using a small moving window, and local maxima are identified as candidate segmentation points.

For this method, we use a GPT-2 model trained specifically for Chinese language modeling (Radford et al., 2019b; Zhao et al., 2019). While we explored larger and more recent architectures, we found that this GPT-2 model offers a good balance between model capacity and computational efficiency. Its moderate size and widespread availability make it a practical choice for analyzing the statistical properties of text. The model consists of 24 transformer decoder layers with a hidden size of 1024, yielding approximately 325 million parameters in total. During inference, we tokenize input sentences at the character level, compute the per-token entropy based on the model's output distribution, and insert segmentation boundaries at entropy peaks.

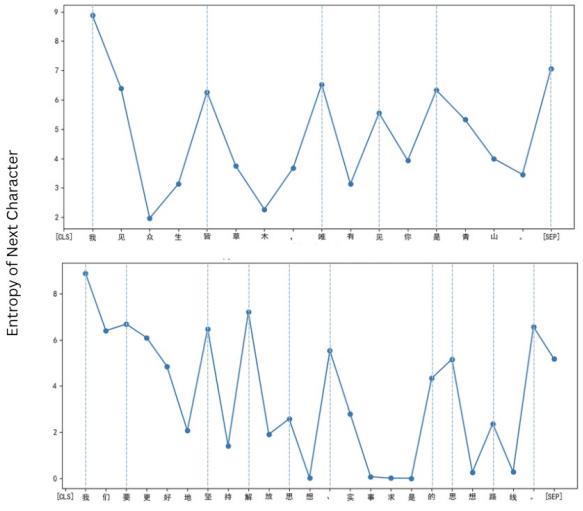

Figure 3. Next-character entropy scores for two randomly selected Chinese sentences evaluated by GPT-2. Each plot illustrates the entropy of the model's next-character prediction at each token position. Blue dashed lines denote local peaks, which serve as span boundaries. These examples are provided to illustrate how the model's uncertainty varies across different parts of a sentence.

4. Experiments

4.1. Datasets

This study uses a subset of the PKU dataset from the SIGHAN 2005 bake-off task. The PKU corpus is a widely used benchmark in segmentation research, consisting of manually annotated sentences drawn from formal written sources such as news articles and academic texts. It provides a reliable reference for evaluating the alignment of predicted segmentation boundaries with linguistically validated ground truth.

To reduce computational demands associated with entropy-based methods, particularly predictive entropy calculation, we limit our experiments to 10% of the full PKU training corpus. This subset contains 2,255 sentences, approximately 90,000 Chinese characters. It was selected to balance experimental feasibility with computational cost, particularly given that the reduced subset required approximately two days to process using the auto-regressive method. Gold-standard word boundaries are retained for evaluation. All characters are simplified Chinese. Each sentence is tokenized at the character level to create the input sequence.

While the dataset size is limited, the controlled experimental setup facilitates direct comparison between methods under consistent conditions. This design supports fine-grained qualitative and quantitative analysis of segmentation behavior, tokenization structure, and alignment with annotated word boundaries.

4.2. Baselines and Configurations

We evaluate three tokenization strategies, each based on Byte Pair Encoding (BPE). Prior work suggests that smaller vocabularies are more suitable for unsegmented languages like Chinese. For example, the GPT-2 model used in Section 3.2 has a vocabulary size of 21,128 tokens. Given the limited size of our dataset, we choose a reduced vocabulary of 12,000 tokens to balance representational efficiency and data sparsity. All methods operate on the same character-level input sequences derived from the preprocessed PKU corpus described in Section 4.1.

- **Standard BPE**: A baseline implementation of frequency-based BPE applied directly to character sequences, without any pre-tokenization. This mirrors the standard application of BPE in contexts like Chinese, where whitespace-based tokenization is not applicable. Merges are selected purely based on adjacent symbol pair frequency, without any linguistic constraints.
- **Statistically-based Entropy + BPE**: Our method that introduces a pre-tokenization step based on statistical signals. Character sequences are first segmented using a score that combines pointwise mutual information and left/right entropy (as described in Section 3.1). The resulting boundaries constrain BPE merges to occur only within identified spans.
- **Auto-regressive LLM-based Entropy + BPE**: Our method that applies pre-tokenization using next-character predictive entropy estimated from a pre-trained autoregressive language model. Entropy peaks are treated as boundary candidates. These boundaries are inserted into the sequence prior to BPE training.

All tokenizers are trained from scratch using the same segmented or unsegmented input, and each is restricted to operate over a vocabulary of 12,000 subword units. This fixed capacity allows for a direct comparison of token efficiency, segmentation quality, and vocabulary utilization across methods.

4.3. Qualitative Analysis of Segmentation Behavior

In both methods, entropy acts as an information-theoretic prior for segmentation, with no modification to the downstream BPE algorithm. This modularity ensures that comparisons with the standard BPE baseline remain valid and interpretable.

To further understand the behavior of entropy-guided pre-tokenization, we present a qualitative visualization of token boundary selection under different methods. This analysis provides insight into how statistical and model-based entropy signals influence segmentation decisions prior to BPE application.

Figure 4 illustrates the token boundaries selected by multiple pre-tokenization strategies for a representative Chinese sentence. Each row corresponds to a different method: the gold-standard segmentation from the PKU dataset, segmentation based solely on predictive entropy from GPT-2, segmentation using left/right entropy alone (entropy-only), and the statistical method with varying values of the weighting parameter λ. We explored λ values using a standard grid search. Vertical lines denote the identified segmentation boundaries.

This visualization highlights several key trends. First, the predictive entropy method identifies boundaries that align closely with semantic units, often matching human annotations. Second, the statistical method demonstrates flexible boundary control via the λ parameter; smaller values result in shorter, more fragmented tokens, while larger values emphasize contextual diversity, yielding longer and more coherent spans. Notably, the entropy-only method achieves reasonable alignment with the gold standard, suggesting that information-theoretic signals alone carry substantial linguistic relevance even without subsequent BPE processing.

These observations support the hypothesis that entropy-informed pre-tokenization can act as a lightweight yet effective proxy for unsupervised word segmentation, offering a principled mechanism to introduce structure into the BPE pipeline for unsegmented scripts.

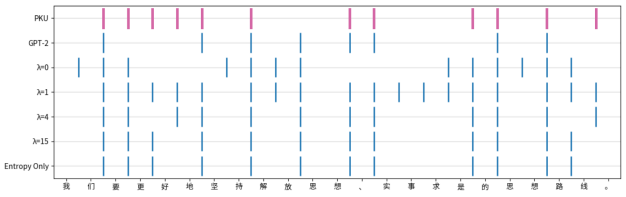

Figure 4. Comparison of pre-tokenization methods on a sample Chinese sentence. Observing the vertical lines from top to bottom, the method with $\lambda = 4$ yields token boundaries that align most closely with the gold standard in the top row. It segments nearly perfectly compared to the ground truth in the top row, with only one extra boundary inserted after the 10th character.

4.4. Intrinsic Evaluation

We assess segmentation quality using precision, recall, and F1 score by comparing predicted subword boundaries to gold-standard word segmentation in the PKU corpus. Precision reflects the proportion of predicted boundaries that align with true word boundaries, while recall measures the proportion of true boundaries that are correctly predicted. The F1 score is computed following the official SIGHAN Bakeoff evaluation script, which defines a word-level match as a segment whose start and end positions exactly align with a gold-standard word. We split our corpus into 70% training and 30% testing for model development and evaluation.

Our experimental procedure consists of four main steps:

1. **Pre-tokenization:** We apply various pre-tokenization strategies to the training data. These include entropy-based methods, GPT-2 uncertainty, and a no-pre-tokenization baseline.
2. **BPE Training:** A Byte-Pair Encoding (BPE) tokenizer is trained on the pre-tokenized training set. All methods use the same algorithm and trainer configuration to ensure consistency.
3. **Segmentation:** The trained tokenizer is then used to segment the test set at inference time. The resulting tokens are treated as predicted word boundaries.
4. **Evaluation:** We compute precision, recall, and F1 score by comparing the predicted word boundaries to the PKU gold-standard segmentation. Character-level boundary matching is used for evaluation.

Our results demonstrate that entropy-based pre-tokenization methods substantially outperform the baseline BPE approach. The best performance is achieved when using an entropy-regularized approach with $\lambda = 4$, which yields the highest F1 score of 58.73, significantly surpassing the baseline's 49.30 by 9.43 percentage points. This setting also achieves the best precision (54.21) and the second-highest recall (64.06), indicating its effectiveness in correctly identifying subword boundaries with fewer false positives.

Method	Precision	Recall	F1
Baseline	46.89	51.96	49.30
GPT-2	52.07	**64.69**	57.70
$\lambda = 0$	28.69	42.92	34.39
$\lambda = 1$	41.24	55.91	47.47
$\lambda = 4$	**54.21**	64.06	**58.73**
$\lambda = 15$	52.83	62.17	57.12
Entropy Only	51.28	60.98	55.71

Table 1. Segmentation results on the PKU dataset. All scores are computed on the unseen 30% of 2,255 sentences using character-level boundary comparison.

The GPT-2 method, which leverages language model uncertainty for boundary detection, performs competitively, achieving an F1 score of 57.70 and the highest recall (64.69), though its precision (52.07) is slightly below that of $\lambda = 4$. These results highlight the strong predictive signal of token-level entropy derived from pretrained LLMs, even without additional regularization.

The entropy-only method also shows strong performance (F1 = 55.71), suggesting that raw entropy is a useful heuristic for segmentation. However, it is outperformed by the regularized variants, indicating that combining entropy with structural constraints improves segmentation accuracy.

Varying the regularization strength λ provides insight into the trade-off between precision and recall. A low value like $\lambda = 0$ yields poor overall performance (F1 = 34.39), primarily due to low precision (28.69). In contrast, moderate values such as $\lambda = 1$ and $\lambda = 15$ show solid gains (F1 = 47.47 and 57.12, respectively), with $\lambda = 15$ emphasizing recall (62.17) more than precision.

Overall, these results show that entropy-based pre-tokenization with tuned parameters provides the most accurate and balanced subword boundary predictions, outperforming both frequency-based baselines and auto-regressive language models.

5. Discussion

The findings of our intrinsic evaluation underscore the effectiveness of entropy-informed pre-tokenization in aligning subword units with linguistically meaningful boundaries in Chinese, an unsegmented language where traditional BPE

tokenization methods often fail to capture semantic structure. These improvements in tokenization fidelity suggest potential downstream benefits across a range of natural language processing tasks.

5.1. Extrinsic Evaluation

While this work focuses on intrinsic segmentation accuracy, a critical direction for future research involves evaluating the impact of entropy-guided tokenization on extrinsic benchmarks. Tasks such as named entity recognition (NER), machine translation, and question answering are particularly sensitive to token boundary precision. Our work shows promising signs that it could be beneficial to integrate our proposed pre-tokenization strategies into standard transformer-based LLM training, primarily when modeling low-resource or unsegmented languages (due to computational limitations, we did not include such in our research). In such settings, where linguistic resources are limited or unavailable, the ability to induce meaningful segmentation from unsupervised entropy signals may offer significant improvements in generalization and cross-lingual transfer.

5.2. Toward Byte-Level Robustness

Another promising direction is adapting entropy-driven methods like ours to byte-level tokenization schemes. Byte-level BPE has gained traction in multilingual settings due to its robustness and language independence. Unlike character- or word-level tokenizers, byte-level approaches operate directly on raw input streams, avoiding assumptions about orthography or script boundaries.

However, this flexibility often comes at the cost of interpretability and token efficiency. We propose to investigate whether entropy signals—derived from character-level statistics or model-based predictive distributions—can serve as segmentation priors in byte-level frameworks. By embedding structural cues into the byte-level token stream, it may be possible to preserve the generality of byte-level models while recovering aspects of morphological awareness.

Such an approach could enhance multilingual performance in real-world conditions involving code-switching, noisy user-generated text, or non-standard encodings. Moreover, entropy-regularized byte-level tokenization could serve as a bridge between data-driven robustness and linguistically informed structure in foundation model pretraining.

6. Conclusion

This paper introduces two entropy-driven pre-tokenization methods to address the limitations of Byte Pair Encoding in unsegmented languages such as Chinese. By incorporating information-theoretic signals, specifically statistical co-occurrence metrics and predictive entropy from a pre-trained autoregressive language model, we effectively bias BPE toward more linguistically coherent token boundaries. Experimental results on the PKU segmentation benchmark confirm that both approaches significantly outperform standard frequency-based BPE, with the statistical method capable of yielding the highest F1 score with proper hyperparameter tuning. Our methods preserve the modularity and efficiency of existing BPE frameworks while improving token granularity and interpretability. We demonstrate the potential of entropy-based priors in bridging the gap between statistical tokenization and linguistic structure across diverse scripts and resource conditions. Finally, we encourage incorporating our work into LLM training, particularly for unsegmented languages like Chinese, with the aim of improving performance on downstream tasks.

Impact Statement

This paper presents work aimed at advancing the field of Natural Language Processing (NLP) by improving models' ability to understand unsegmented text, thereby contributing to more equitable language technologies that extend beyond English. Given that all tokenization decisions directly influence model behavior and downstream task performance, our proposed methods should be critically evaluated in applied contexts before deployment. While our work seeks to enhance the performance of NLP systems across diverse languages, we are not aware of any immediate societal concerns that necessitate specific discussion.

References

Chen, Y. and Deng, C. Radical-aware neural word segmentation for chinese. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5321–5330, 2020.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. Canine: Pre-training an efficient tokenization-free encoder for language representation. In *Transactions of the Association for Computational Linguistics*, volume 9, pp. 121–138, 2021. URL https://aclanthology.org/2021.tacl-1.8/.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.

Emerson, T. Overview of the fourth sighan bakeoff for chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pp. 1–10, 2005.

Gage, P. A new algorithm for data compression. *C Users Journal*, 12(2):23–38, 1994.

Harris, Z. *Mathematical Structures of Language*. Wiley, New York, NY, USA, 1968.

Huang, C.-R. and Liu, F.-H. Chinese word segmentation: A hybrid approach using dictionary and statistics. In *Proceedings of the 4th Workshop on Very Large Corpora*, 1997.

Jiang, P., Long, D., Zhang, Y., Xie, P., Zhang, M., and Zhang, M. Unsupervised boundary-aware language model pretraining for chinese sequence labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 526–537, Singapore, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.34.

Jin, Z. and Tanaka-Ishii, K. Unsupervised segmentation of chinese text by use of branching entropy. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 428–435, 2006.

Kudo, T. and Richardson, J. 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*, pp. 66–71, 2018.

Liu, A., Hayase, J., Hofmann, V., Oh, S., Smith, N. A., and Choi, Y. Superbpe: Space travel for language models, 2025. URL https://arxiv.org/abs/2503.13423.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. In *arXiv preprint arXiv:1907.11692*, 2019.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Pagnoni, A., Pasunuru, R., Rodriguez, P., Nguyen, J., Muller, B., Li, M., Zhou, C., Yu, L., Weston, J., Zettlemoyer, L., Ghosh, G., Lewis, M., Holtzman, A., and Iyer, S. Byte latent transformer: Patches scale better than tokens. 2024. URL https://arxiv.org/abs/2412.09871. arXiv preprint.

Provilkov, A., Lu, Y., Morin, L., and Grangier, D. Bpe-dropout: Simple and effective subword regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2750–2759, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019a. OpenAI technical report.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019b.

Schmidt, C. W., Reddy, V., Tanner, C., and Pinter, Y. Boundless byte pair encoding: Breaking the pre-tokenization barrier, 2025. URL https://arxiv.org/abs/2504.00178.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1715–1725, 2016.

Sproat, R., Shih, C., Gale, W., and Chang, N. A stochastic finite-state word-segmentation algorithm for chinese. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 66–73, 1994.

Tanaka-Ishii, K. Entropy as an indicator of context boundaries: An experiment using a corpus of japanese. *Journal of Quantitative Linguistics*, 12(1):65–87, 2005.

Tay, Y., Zhang, A., Bahri, D., and Metzler, D. Charformer: Fast character transformers via gradient-based subword tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4283–4298, 2021.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. Byt5: Towards a token-free future with pre-trained byte-to-byte models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. URL https://arxiv.org/abs/2105.13626.

Xue, N. Chinese word segmentation as character tagging. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pp. 133–134, 2003.

Zhao, Z., Chen, H., Zhang, J., Zhao, X., Liu, T., Lu, W., Chen, X., Deng, H., Ju, Q., and Du, X. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, pp. 241, 2019.