Evaluating Text Segmentation for Representation Quality in a Low-Resource Language

Anonymous authors
Paper under double-blind review

Abstract

Text segmentation can directly affect the quality of downstream language representations, particularly in low-resource, non-delimited languages such as Myanmar, where syllable and word boundaries do not necessarily align. We evaluate five segmentation approaches—grammar-informed rules, dictionary lattices, statistical CRFs, and two neural language models—on 1,100 verified sentences across Government, News, and Novel domains. Beyond intrinsic segmentation performance, we assess downstream effects using three multilingual subword tokenizers and three embedding models. Grammar-informed segmentation remains stable across domains (Boundary F1: 0.771–0.815), while neural language models show greater degradation on literary text, with Boundary F1 decreasing by up to 0.168. Boundary accuracy is strongly correlated with semantic embedding quality (Pearson $r = 0.895$, $p < 10^{-5}$), while poor segmentation increases subword sequence length by up to 20%. These results show that segmentation quality is an important factor in downstream representation quality and that explicit linguistic knowledge remains valuable for robust low-resource language processing.

1 Introduction

In natural language processing, text segmentation is a foundational step that determines the units from which downstream language representations are constructed (Vaswani et al., 2017). In many high-resource languages, whitespace provides an initial approximation of word boundaries, while subword tokenization methods such as Byte-Pair Encoding and WordPiece further decompose words into units suitable for neural language models (Sennrich et al., 2016; Kudo, 2018). For non-delimited writing systems, however, whitespace does not reliably indicate word boundaries, making text segmentation a non-trivial structured prediction problem. This challenge is particularly pronounced for low-resource, morphologically rich languages, where syllable boundaries do not necessarily correspond to word boundaries and limited linguistic resources constrain data-driven approaches. Segmentation errors can therefore propagate beyond boundary detection: poorly segmented words may be fragmented into additional subword units, increasing sequence length and altering the semantic representations produced by multilingual language models (Xue et al., 2022; Rust et al., 2021). Consequently, evaluating text segmentation requires considering not only intrinsic boundary accuracy but also its effects on downstream representation quality.

Contemporary multilingual foundation models offer an alternative to language-specific segmentation methods by leveraging knowledge acquired from large-scale multilingual pre-training (Touvron et al., 2023; Team et al., 2025). Their effectiveness, however, can depend on the amount and diversity of language-specific data represented during pre-training. This presents a particular challenge for low-resource languages such as Myanmar (Burmese), where linguistic variation across formal, journalistic, literary, and colloquial writing may not be equally represented in available training corpora. A segmentation method that performs well on formal text may therefore not maintain the same boundary accuracy when applied to substantially different writing styles, such as literary dialogue. Such domain sensitivity is consequential because segmentation errors can propagate to downstream subword tokenization and semantic representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This motivates a central question of this study:

> Can multilingual neural language models provide robust word segmentation across domains in a low-resource language, or can explicit linguistic knowledge provide greater stability under domain shift?

To answer this question, we use Myanmar as a testbed, where syllable boundaries do not necessarily align with word boundaries and grammatical particles and postpositions further complicate segmentation. Writing conventions also vary across formal Government, News, and literary Novel domains (Htay & Murthy, 2008; Thu et al., 2014). We compare five representative systems spanning distinct computational paradigms:

1. **Grammar-Informed Rules (`mmdt-tokenizer`)**: deterministic segmentation using explicit linguistic rules.
2. **Dictionary Lattice Optimization (`oppaWord`)**: dictionary-based candidate segmentation using lattice optimization.
3. **Supervised Sequence Modeling (`ricelang`)**: character-level CRF-based boundary prediction.
4. **Cloud Foundation Model (`Gemini 3.1 Flash Lite`)**: proprietary neural language model prompted for word segmentation.
5. **Local Language Model (`Gemma 4 E2B-it`)**: open-weight instruction-tuned model executed locally.

We benchmark these systems on 1,100 human-annotated sentences across the three domains. Beyond intrinsic segmentation performance, we evaluate downstream representation quality through subword fertility across XLM-R, Qwen 2.5, and mBERT, and semantic embedding fidelity across LaBSE, mE5, and MiniLM. This evaluation examines whether differences in word-boundary accuracy propagate to downstream subword and semantic representations. The main contributions of this work are threefold:

1. **Cross-paradigm comparison.** We provide a systematic evaluation of word segmentation approaches spanning explicit linguistic rules, dictionary-based optimization, supervised statistical modeling, and neural language models within a common experimental setting.
2. **Cross-domain robustness analysis.** We examine how these segmentation paradigms behave under shifts from formal and journalistic text to literary language, providing empirical evidence of substantial differences in their stability across writing styles.
3. **Segmentation-to-representation analysis.** We connect intrinsic segmentation performance with downstream representation quality, quantifying the relationship between boundary accuracy, subword fragmentation, and semantic embedding fidelity.

2 Related Work and Background

2.1 Myanmar Script and Segmentation Challenges

In continuous scripts without typographic whitespace (such as Chinese, Japanese, Thai, Khmer, and Myanmar), mapping raw character sequences into discrete representation units is a longstanding challenge in natural language processing (Htay & Murthy, 2008). The Myanmar writing system is composed of syllabic clusters, but a single semantic word frequently spans multiple syllables, and individual syllables rarely correspond to independent syntactic entities. Furthermore, Myanmar exhibits complex agglutinative morphology: case markers, tense-aspect-mood indicators, and honorific particles attach directly as enclitics to root nouns and verbs (Thu et al., 2014). Early computational paradigms relied on syllable-level longest matching heuristics and lexicon lookup (Thet et al., 2008; Pa & Thein, 2008), or unsupervised Bayesian models integrating dictionary priors (Thu et al., 2014; Thida et al., 2018). While effective for constrained domains, these approaches struggle with out-of-vocabulary compounding and stylistic variation across genres.

2.2 Segmentation Approaches: Rules, Statistics, and Language Models

Supervised sequence modeling approaches, particularly Conditional Random Fields (CRFs), established strong baselines for word boundary identification by modeling transitional dependencies across character sliding windows (Pa et al., 2016). Modern toolkits such as `ricelang` (San, 2026) implement character-level CRF sequence labelers without manual feature engineering. In parallel, lattice-based architectures like `oppaWord` (Thu, 2025) construct candidate Directed Acyclic Graphs (DAGs) of word spans, balancing lexicon coverage against maximum matching paths. Recently, `mmdt-tokenizer` (Thida et al., 2026) introduced a grammar-informed approach that directly encodes morphosyntactic production rules (e.g., compound nouns, clitic attachment, and numerical classifiers). These paradigms represent fundamentally different approaches: statistical CRFs learn data-driven transition distributions, lattice models optimize graph shortest paths, and grammar-informed systems impose explicit structural rules.

2.3 The Tokenizer Bottleneck in Multilingual Models

Recent evaluations demonstrate rapid progress in the multilingual capabilities of large language models on low-resource benchmarks (Aung et al., 2026; Team et al., 2025). However, foundational work by Rust et al. (2021) demonstrated that the quality of monolingual tokenization is a primary determinant of downstream representation performance in multilingual transformers. Suboptimal tokenization leads to excessive subword fertility (number of subwords per word), over-fragmentation into single characters, and degraded semantic representations. When general-purpose LLMs are prompted to perform structured in-context tokenization, their outputs reflect their internal representations of word boundaries. Yet, whether these neural representations remain robust when transitioning from curated web text to informal or literary prose remains an open empirical question.

3 Evaluation Framework

Our evaluation is guided by three research questions that examine segmentation performance, cross-domain robustness, and downstream representation quality:

1. **RQ1**: How do grammar-informed, dictionary-based, supervised statistical, and neural language model approaches compare in word segmentation accuracy for Myanmar text?
2. **RQ2**: How robust are these segmentation approaches across Government, News, and Novel domains, and how does their performance change under domain shift?
3. **RQ3**: To what extent is segmentation quality associated with downstream representation quality, as measured by subword fertility and semantic embedding fidelity?

Together, these questions move the evaluation from intrinsic segmentation accuracy (**RQ1**), to robustness across writing styles (**RQ2**), and finally to the downstream consequences of segmentation choices (**RQ3**).

3.1 Evaluation Datasets

We evaluate all systems on a manually annotated corpus of 1,100 contemporary Myanmar sentences spanning three domains:

1. **Government (Formal):** 211 sentences, 4,164 tokens, and 1,600 unique words from official textbooks and press communications, characterized by standardized phrasing and formal grammatical conventions.
2. **News (Mixed Formality):** 300 sentences, 6,167 tokens, and 2,149 unique words from online news sources, combining standard written conventions with colloquial expressions and variation in particle usage.
3. **Novels (Literary):** 589 sentences, 11,972 tokens, and 4,889 unique words from three authors, combining literary prose and character dialogue with greater lexical and stylistic variation.

3.2 Segmentation Methods

We evaluate five systems representing distinct segmentation paradigms.

Grammar-Informed Rules (`mmdt-tokenizer`): `mmdt-tokenizer` (v0.1.3) is a deterministic Myanmar segmenter based on hand-crafted linguistic rules for compounds, verbal morphology, and clitic attachment. We enable pattern protection ($protect_{pattern} = True$) for structured expressions such as dates, currencies, phone numbers, and URLs.

Dictionary Lattice Optimization (`oppaWord`): `oppaWord` constructs a dictionary-based lattice of candidate word spans over syllable sequences and selects a segmentation using Bidirectional Maximum Matching (Bi-MM). We use its dictionary-only lattice mode with a Bi-MM boost and deterministic punctuation and particle post-processing.

Supervised Statistical Modeling (`ricelang`): `ricelang` (v0.4.3) uses a Conditional Random Field (CRF) to predict word boundaries from character sequences using local character n-gram features. We apply `ricelang.tokenize(sentence, form="word")` without task-specific fine-tuning.

Neural In-Context Learning (`Gemini 3.1 Flash Lite`): We evaluate `Gemini 3.1 Flash Lite` (Team et al., 2025) as a few-shot segmentation model using five linguistic rules and four demonstrations spanning the evaluation domains. We use JSON-constrained output and temperature $T = 0.0$; the demonstrations are shown in Table 1.

Local Language Model (`Gemma 4 E2B-it`): We run `Gemma 4 E2B-it` (Team et al., 2024) locally using LiteRT-LM with the same rules, demonstrations, and $T = 0.0$ configuration as Gemini. JSON formatting is enforced through prompt instructions and parsed locally, with each sentence processed in an independent context.

Table 1: Few-shot prompt exemplars utilized for in-context structured prediction. Tokens in the reference output are delimited by a pipe (|) character.

Domain	Input Sentence and Structured Tokenization Reference
News 1	*Input:* မိတ္ထီလာမြို့ကိုဝင်နိုင်ထွက်နိုင်တဲ့လမ်းကြောင်း၅ကြောင်းထက်မနည်းရှိပါတယ်။ *Output:* မိတ္ထီလာ | မြို့ | ကို | ဝင်နိုင်ထွက်နိုင် | တဲ့ | လမ်းကြောင်း | ၅ | ကြောင်းထက်မနည်း | ရှိပါတယ် | ။
News 2	*Input:* အဲဒီကနေချင်းပြည်နယ်အိန္ဒိယနယ်စပ်အထိသွားလို့ရပါတယ်။ *Output:* အဲဒီကနေ | ချင်းပြည်နယ် | အိန္ဒိယ | နယ်စပ် | အထိ | သွား | လို့ | ရပါတယ် | ။
Gov.	*Input:* အောက်ပါအရည်အချင်းများနှင့် ပြည့်စုံသောနိုင်ငံသားများသည် ရွေးကောက်တင်မြှောက်ခံပိုင်ခွင့်ရှိသည် *Output:* အောက်ပါ | အရည်အချင်းများ | နှင့် | ပြည့်စုံသော | နိုင်ငံသားများ | သည် | ရွေးကောက်တင်မြှောက်ခံပိုင်ခွင့် | ရှိသည်
Novels	*Input:* သူ့ဘေးမှထိုင်ခုံကားလွတ်နေ၏။ *Output:* သူ့ဘေး | မှ | ထိုင်ခုံ | ကား | လွတ် | နေ | ၏ | ။

Prompt Guidelines and Rules. Both language models were provided with five explicit instructions: (1) **Compound Words:** do not split multi-syllable compound nouns forming a single semantic concept; (2) **Post-positions and Clitics:** separate grammatical case markers and cliticized post-positions from their nominal stems; (3) **Punctuation Isolation:** isolate sentence-ending punctuation marks (။, ၊); (4) **Alphanumeric Spans:** preserve loanwords, dates, numbers, and Latin characters as contiguous spans; and (5) **Output Format:** return exclusively a JSON-compliant array of token strings.

3.3 Evaluation Metrics

We evaluate segmentation performance using four complementary metrics that capture exact word recovery, boundary detection, generalization to rare words, and segmentation granularity.

Word-Level Exact Match (Word P, R, F_1): A predicted token is scored as true positive if and only if its character span and spelling match the ground-truth annotation exactly. Word F1 reflects the balance between over-segmentation and under-segmentation.

Character-Gap Boundary Accuracy (Boundary F_1): For a sentence of length L, let $\mathcal{B}$gold and $\mathcal{B}$pred denote the gold and predicted inter-character boundary sets. We compute:

$$B_P = \frac{|\mathcal{B}\text{gold} \cap \mathcal{B}\text{pred}|}{|\mathcal{B}\text{pred}|}, \quad B_R = \frac{|\mathcal{B}\text{gold} \cap \mathcal{B}\text{pred}|}{|\mathcal{B}\text{gold}|}, \quad B_{F1} = \frac{2B_P B_R}{B_P + B_R}. \tag{1}$$

Boundary F1 provides partial structural credit when a system correctly identifies some word boundaries even when the resulting token spans do not fully match the gold segmentation.

Out-of-Vocabulary Generalization (OOV Recall): We define OOV terms as lexical types occurring fewer than five times across the full corpus (7,527 unique types, accounting for 38–46% of all token occurrences). OOV Recall measures the fraction of these rare entities correctly segmented.

Granularity Diagnostic (Tokens per Sentence, Tok/Sent): The average number of tokens produced per sentence serves as an architectural diagnostic, detecting whether a model exhibits systematic over-segmentation or under-segmentation.

3.4 Experimental Setup

All five segmentation systems (`mmdt`, `oppaword`, `ricelang`, `llm_gemini`, and `llm_gemma4`) were evaluated on the same 1,100-sentence Myanmar corpus,comprising 211 Government, 300 News, and 589 Novel sentences. Table 2 summarizes the model and execution settings.

Table 2: Experimental configuration. Hardware and acceleration refer to local Gemma inference only.

Component	**Configuration**
LLM prompting & decoding	Five prompt instructions (including the JSON output requirement) and four demonstrations. Temperature was set to 0.0 for both LLMs; top-p and top-k were not explicitly set and remained at their respective inference-engine defaults.
Cloud LLM	`gemini-3.1-flash-lite`, accessed through Vertex AI with `google-genai`; JSON-constrained output.
Local LLM	Instruction-tuned `gemma-4-E2B-it.litertlm`, run with LiteRT-LM; JSON-format instructions and local response parsing. Each sentence was processed in an independent context.
Subword tokenizers	`FacebookAI/xlm-roberta-base`, `Qwen/Qwen2.5-7B`, and `google-bert/bert-base-multilingual-cased`.
Sentence encoders	`sentence-transformers/LaBSE` (768 dimensions), `intfloat/multilingual-e5-base` (768 dimensions), and `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` (384 dimensions). The `passage:` prefix was used for mE5 inputs.
Local inference device	Apple M2; 8 CPU cores (4 performance, 4 efficiency); 10 GPU cores; 8 GB unified memory.
Local inference runtime	LiteRT-LM 0.13.1 with the Apple Metal GPU backend.

The reported results are based on a single evaluation run per system, in which each sentence was processed once under greedy deterministic decoding ($T = 0.0$).

For subword analysis, each tokenizer processed the output of each segmentation system and the gold segmentation. Subword counts were compared with the corresponding gold-segmented input. For representation analysis, cosine similarity was computed between sentence embeddings of predicted and gold segmentations. The software environment comprised Python 3.14, `google-genai` 2.8.0, `litert-lm` 0.13.1, `pydantic` 2.13.4, `mmdt-tokenizer` 0.1.3, `ricelang` 0.4.3, `transformers` 5.17.0, `tokenizers` 0.23.1, `sentence-transformers` 6.0.1, and PyTorch 2.14.0.

4 Results

4.1 Segmentation Performance

Table 3 summarizes intrinsic segmentation performance across the three domains. `mmdt-tokenizer` and Gemini 3.1 Flash Lite achieve the strongest overall results. Gemini obtains the highest Word F1 and Boundary F1 on Government text (0.497 and 0.801) and the highest Boundary F1 on News (0.834), while `mmdt-tokenizer` achieves the highest Word F1 on News (0.647) and both the highest Word F1 and Boundary F1 on Novels (0.537 and 0.802).

Table 3: Intrinsic segmentation performance across domains. Best results within each domain are shown in **bold**.

System	Domain	Word P	Word R	Word F1	B-F1	OOV R	Tok/Sent
mmdt-tokenizer	Gov.	0.404	0.521	0.455	0.771	0.307	25.48
(Grammar Rules)	News	0.626	0.669	**0.647**	0.815	0.418	21.95
	Novels	0.529	0.546	**0.537**	**0.802**	0.278	20.98
oppaWord	Gov.	0.129	0.234	0.166	0.649	0.094	35.82
(Dictionary Lattice)	News	0.217	0.333	0.263	0.739	0.183	31.48
	Novels	0.213	0.330	0.259	0.736	0.169	31.53
ricelang	Gov.	0.120	0.227	0.157	0.664	0.086	37.42
(Statistical CRF)	News	0.240	0.350	0.285	0.741	0.180	29.92
	Novels	0.197	0.299	0.237	0.731	0.124	30.80
Gemini 3.1 Flash Lite	Gov.	0.440	0.571	**0.497**	**0.801**	**0.383**	25.62
(Cloud Foundation Model)	News	0.494	0.599	0.542	**0.834**	**0.447**	24.91
	Novels	0.359	0.485	0.412	0.785	**0.313**	27.49
Gemma 4 E2B-it	Gov.	0.461	0.361	0.405	0.700	0.328	15.46
(Local Edge LLM)	News	0.451	0.338	0.386	0.746	0.290	15.43
	Novels	0.413	0.295	0.344	0.578	0.265	14.50

`oppaWord` and `ricelang` achieve lower Word F1 but comparatively higher Boundary F1. Both produce 30–37 tokens per sentence, compared with 20–22 in the gold annotations, indicating finer-grained segmentation. This difference helps explain why their Boundary F1 remains relatively high despite fewer exact word-level matches. In contrast, Gemma 4 E2B-it produces only 14–15 tokens per sentence, indicating a tendency toward under-segmentation.

4.2 Cross-Domain Robustness Under Domain Shift

The results show different cross-domain performance patterns across the evaluated systems:

Grammar-Informed Consistency. `mmdt-tokenizer` shows the most consistent boundary performance across domains, with Boundary F1 ranging from 0.771 to 0.815 (a span of 0.044). Its deterministic rules remain stable across formal and literary text, achieving the highest Word F1 on News (0.647) and Novels (0.537).

Neural Sensitivity in Cloud vs. Edge Models. `Gemini 3.1 Flash Lite` maintains stable boundary detection across domains (Boundary F1 between 0.785 and 0.834), but its Word F1 drops from 0.542 on News to 0.412 on Novels as it splits colloquial clitics into morphemes. In contrast, `Gemma 4 E2B-it` suffers a direct boundary detection drop on literary text, with Boundary F1 falling from 0.746 on News to 0.578 on Novels (a decrease of 0.168) due to punctuation fusion and under-segmentation.

Dictionary and CRF Baselines. `oppaWord` and `ricelang` post their lowest Boundary F1 on Government text (0.649 and 0.664) and improve on News and Novels (0.731 to 0.741), though persistent over-segmentation keeps their Word F1 low across all three domains.

4.3 Downstream Representation Quality

To examine whether segmentation differences propagate to downstream representations, we evaluate (1) subword fragmentation across multilingual tokenizers and (2) semantic similarity across multilingual sentence encoders. Table 4 summarizes these results.

Table 4: Downstream representation quality across multilingual subword tokenizers (expansion ratio relative to gold) and semantic sentence encoders (cosine similarity to gold).

		Subword Expansion (Ratio)				**Semantic Similarity** (cos)		
Domain	**System**	**XLM-R**	**Qwen**	**mBERT**	**Single%**	**LaBSE**	**mE5**	**MiniLM**
Gov	*raw (unseg)*	0.842	0.938	0.308*	15.5%	0.252	0.994	0.366
Gov	*gold*	1.000	1.000	1.000	21.7%	1.000	1.000	1.000
Gov	**mmdt**	**1.055**	1.019	**1.037**	24.3%	0.992	0.997	0.938
Gov	Gemini	1.069	1.021	1.024	26.1%	0.995	0.998	0.901
Gov	oppaWord	1.208	1.053	1.101	31.7%	0.975	0.992	0.838
Gov	ricelang	1.205	1.058	1.103	31.3%	0.983	0.994	0.816
Gov	Gemma	0.949	0.981	0.992	21.3%	0.984	0.995	0.415
News	*raw (unseg)*	0.798	0.912	0.376*	16.7%	0.224	0.991	0.751
News	*gold*	1.000	1.000	1.000	23.5%	1.000	1.000	1.000
News	**mmdt**	1.012	1.005	0.999	23.9%	0.986	0.997	0.944
News	Gemini	1.057	1.020	1.004	26.4%	0.990	0.998	0.908
News	oppaWord	1.182	1.050	1.019	30.0%	0.972	0.993	0.501
News	ricelang	1.157	1.042	1.021	29.7%	0.975	0.994	0.516
News	Gemma	0.949	0.980	0.991	22.9%	0.983	0.997	0.858
Novels	*raw (unseg)*	0.820	0.912	0.810*	19.0%	0.836	0.990	0.369
Novels	*gold*	1.000	1.000	1.000	23.9%	1.000	1.000	1.000
Novels	**mmdt**	1.020	1.003	1.012	25.0%	0.984	0.997	0.967
Novels	Gemini	1.093	1.034	1.015	28.3%	0.987	0.996	0.938
Novels	oppaWord	1.157	1.051	1.032	31.3%	0.978	0.993	0.926
Novels	ricelang	1.142	1.047	1.031	30.2%	0.975	0.994	0.920
Novels	Gemma	0.943	0.959	0.967	23.7%	0.965	0.988	0.674

Note: Subword expansion ratios are computed relative to gold. Single% denotes single-character fallback proportion under XLM-RoBERTa. *Low raw ratios under mBERT reflect sequence collapse into `[UNK]` tokens caused by whitespace-delimited pre-tokenization failure on space-less script. Correlations with Boundary F1: Microsoft mE5 $r = 0.895$ ($p < 10^{-5}$); Google LaBSE $r = 0.738$ ($p = 0.0017$); MiniLM Spearman $\rho = 0.632$ ($p = 0.011$).

4.3.1 Subword Fragmentation and Sequence Expansion

Following standard tokenizer evaluation practice (Rust et al., 2021), we evaluate subword fragmentation across three tokenization architectures: SentencePiece Unigram in XLM-R (Conneau et al., 2020), Byte-BPE(Qwen 2.5), and WordPiece(mBERT) (Devlin et al., 2019). We report the sentence-level subword expansion ratio relative to gold, $\frac{1}{|D|}\sum_{s\in D}\frac{N_{\text{sub}}(s)}{N_{\text{gold}}(s)}$, and the single-character fallback proportion (single%) under XLM-RoBERTa as the primary indicator of lexical shattering.

As shown in Table 4, `mmdt-tokenizer` remains close to the gold reference across all three tokenizers (ratios 0.999–1.055). In contrast, `oppaWord` and `ricelang` produce more subwords in XLM-R (ratios up to 1.208) and a higher single-character fallback rate (up to 31.7%). `Gemini` also expands subwords on Novels in XLM-R (ratios up to 1.093), while `Gemma` under-segments across all domains (< 1.000).

4.3.2 Semantic Embedding Fidelity

We next examine whether segmentation quality affects continuous semantic representations using Google LaBSE (Feng et al., 2022), Microsoft `multilingual-e5-base` (Wang et al., 2024), and multilingual MiniLM. For each sentence, we compute the cosine similarity between sentence embeddings produced from the predicted and gold segmentations.

Across all three encoders, `mmdt-tokenizer` produces representations consistently close to gold (Table 4). However, unsegmented text causes substantial representation shifts, particularly for LaBSE and MiniLM, where similarity falls to 0.224 and 0.366, respectively. Under-segmentation also degrades representations: `Gemma` drops to 0.415 in MiniLM.

Boundary F1 correlates strongly with embedding similarity for all three encoders (Table 4 note), confirming that accurate word segmentation directly improves downstream semantic fidelity.

4.4 Qualitative Error Analysis

Qualitative inspection of system outputs reveals characteristic failure modes between rule-based and neural systems.

4.4.1 Rule-Based Lexical Coverage vs. Neural Contextual Drift

The two systems exhibit distinct error patterns. Errors from the grammar-informed `mmdt-tokenizer` are largely deterministic and interpretable, often reflecting lexical granularity. For example, ကလေးများ (*children*) is split into ကလေး (*child*) and the plural marker များ, while compounds such as အမျိုးသားပညာရေးစနစ် (*national education system*) are decomposed into အမျိုးသား (*national*), ပညာရေး (*education*), and စနစ် (*system*).

In contrast, `Gemini 3.1 Flash Lite` shows greater context-dependent boundary variation. It preserves the formal past-tense construction ခဲ့ပါသည် as a single token in Government and News text, but splits the colloquial ခဲ့ပါတယ် in Novels into ခဲ့, ပါ, and တယ်. This pattern illustrates greater sensitivity to writing style and is consistent with the cross-domain performance variation observed quantitatively.

4.4.2 Sentence Boundary and Punctuation Disruption in Small LLMs

`Gemma 4 E2B-it` shows substantial degradation on literary text (B-F1: 0.578), characterized by punctuation-related fusion and under-segmentation. It frequently attaches sentence-final particles to the Myanmar full stop (။), producing fused tokens such as ၏။ and တယ်။. It also averages only 14.50 tokens per sentence, approximately 30% below the gold reference, indicating systematic under-segmentation of conversational text.

5 Discussion

Our findings highlight three implications for segmentation in low-resource, non-delimited languages:

1. **Neural Models Show Cross-Domain Variation.** Gemini performs strongly on formal text but degrades on literary text, showing greater sensitivity to writing style.
2. **Linguistic Rules Provide Stability.** `mmdt-tokenizer` maintains comparatively stable performance across domains, supporting the value of explicit linguistic knowledge and motivating hybrid rule–neural approaches.

3. **Segmentation Granularity Affects Sequence Length.** `oppaWord` and `ricelang` produce 30–37 tokens per sentence versus 20–21 in the gold annotations, while `mmdt-tokenizer` remains closer at 21–25. Such fragmentation increases sequence length, which is relevant to Transformer self-attention with $\mathcal{O}(N^2)$ complexity (Vaswani et al., 2017).

Limitations and Future Directions. This study is limited to Myanmar, three domains, and a selected set of models. Future work will extend the evaluation to downstream tasks such as Named Entity Recognition and Part-of-Speech tagging, explore additional multilingual models, and investigate hybrid approaches combining linguistic constraints with neural segmentation.

6 Conclusion

This paper evaluated how text segmentation affects representation quality in a low-resource, non-delimited, and morphologically rich language, using Myanmar text across Government, News, and Novel domains as a testbed. By comparing grammar-informed rules, dictionary lattices, supervised CRF sequence modeling, and neural language models, we show that segmentation quality has important downstream consequences for both semantic representations and subword tokenization. Neural language models achieve strong boundary detection on formal text but exhibit substantial degradation on literary dialogue, whereas grammar-informed segmentation remains comparatively stable across domain shifts. Across the evaluated methods and domains, boundary accuracy is strongly correlated with semantic embedding quality ($r = 0.895, p < 10^{-5}$), while poor segmentation also produces longer subword sequences. These findings demonstrate that text segmentation is more than a preprocessing choice: it plays an important role in the quality and stability of downstream language representations. More broadly, the results show that explicit linguistic knowledge remains valuable alongside statistical and neural approaches for robust language processing in low-resource settings.

Reproducibility Statement

All evaluation resources, annotation guidelines, prompts, schemas, and inference scripts used in this study are documented in the paper and accompanying supplementary materials. Model configurations and experimental settings are provided in Section 3 and the corresponding tables to support reproduction of the reported results.

LLM and Generative AI Disclosure

Generative AI tools were used to assist with code debugging, presentation, formatting, and language editing. They were not used to generate the experimental data, formulate scientific claims, or interpret the results. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final content of the work.

References

Thura Aung, Jann Railey Montalan, Jian Gang Ngui, and Peerat Limkonchotiwat. Burmese-san: Burmese nlp benchmark for evaluating large language models, 2026. URL `https://arxiv.org/abs/2602.18788`.

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

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, 2022.

Hla Hla Htay and Kavi Narayana Murthy. Myanmar word segmentation using syllable level longest matching. In *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 41–48, 2008.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, 2018.

Win Pa Pa and Ni Lar Thein. Myanmar word segmentation using hybrid approach. In *Proceedings of 6th International Conference on Computer Applications*, pp. 166–170, 2008.

Win Pa Pa, Ye Kyaw Thu, Andrew Finch, and Eiichiro Sumita. Word boundary identification for myanmar text using conditional random fields. In *Genetic and Evolutionary Computing*, pp. 447–456. Springer International Publishing, 2016.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. 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)*, pp. 3118–3135, 2021.

Kaung Htet San. ricelang: Tiny Language Detector and Tokenizer for Southeast Asian Languages, 2026. URL `https://github.com/kaunghtetsan275/ricelang`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

Gemini Team, Rohan Anil, et al. Gemini: A family of highly capable multimodal models, 2025. URL `https://arxiv.org/abs/2312.11805`.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Tun Thura Thet, Jin-Cheon Na, and Wunna Ko Ko. Word segmentation for the myanmar language. *Journal of Information Science*, 34(5):688–704, 2008. doi: 10.1177/0165551507086258.

Aye Thida, Nway Nway Han, and Sheinn Thawtar Oo. Statistical machine translation using 5-grams word segmentation in decoding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation*, pp. 83–90, 2018.

Myo Thida, Nu Wai Thet, and Thein Kyaw Lwin. MMDT-tokenizer: Grammar-based text segmentation. `https://github.com/Myanmar-Data-Tech/mmdt-tokenizer`, 2026.

Ye Kyaw Thu. oppaWord: Hybrid DAG+Bi-MM+LM Myanmar Word Segmenter, 2025. URL `https://github.com/ye-kyaw-thu/oppaWord`. Language Understanding Lab (LU Lab), Myanmar.

Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita, and Yoshinori Sagisaka. Integrating dictionaries into an unsupervised model for myanmar word segmentation. In *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*, pp. 20–27, 2014.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.