Sahara Tokenizers at MWE-2026 PARSEME 2.0 Subtask 1: Combining Contextual Embeddings with Structural Decoding for Multi-Word Expression Detection

Yunus Karatepe[1], Mert Sülük[1,2], Zeynep Tuğçe Kırımlı[1,3], Begüm Özbay[1,4]

[1]Istanbul Technical University
[2]Istanbul University
[3]Istanbul University-Cerrahpasa
[4]Yildiz Technical University
{karatepe22, suluk20, ozbaybe21}@itu.edu.tr, zeynep.kirimli@iuc.edu.tr

Abstract

Multi-Word Expressions (MWEs) pose a significant challenge for natural language processing systems due to their idiosyncratic semantic and syntactic properties. This paper describes our system for the PARSEME 2.0 Shared Task on automatic identification of verbal MWEs across 17 typologically diverse languages. Our approach combines multilingual BERT with explicit Part-of-Speech (POS) feature injection through a dual-head architecture that jointly performs BIO-based identification and category classification. We further investigate extensions, including Conditional Random Field (CRF) decoding for structured prediction, focal loss for addressing class imbalance, and model ensembling for improving discontinuous MWE detection. Our official submission achieves a global MWE-based F1 score of 48.39%, securing second place in the shared task. Ablation studies reveal a strong synergy between POS features and CRF decoding, with the combined approach yielding the best single-model performance. Furthermore, ensembling models trained with different objectives improves both overall F1 score and discontinuous MWE scores, demonstrating the importance of training diversity for capturing non-adjacent syntactic patterns.

1 Introduction

Multi-Word Expressions (MWEs) are idiosyncratic lexical units whose automatic identification is crucial for downstream NLP tasks (Baldwin and Kim, 2010). The PARSEME shared task series (Savary et al., 2017; Ramisch et al., 2020) addresses the core challenges of MWE detection: handling discontinuous surface realizations, managing severe class imbalance, and generalizing across typologically diverse languages.

In this work, we present a system for PARSEME 2.0 (Subtask 1) (Scholivet et al., 2026), which addresses the joint identification and classification of verbal MWEs. Our system is based on multilingual BERT (Devlin et al., 2019) augmented with explicit Part-of-Speech (POS) features. This approach employs a multi-task formulation with dual prediction heads for joint boundary identification (BIO) and category classification. Beyond the official submission, we investigate architectural extensions including Conditional Random Fields (CRF) for structured decoding, Focal Loss (Lin et al., 2017) for mitigating imbalance, and diverse ensembling strategies.

Our main contributions are as follows:

- We demonstrate that POS feature injection improves recall by 2.78 points, with strong gains on seen expressions (+1.64% F1).
- We identify a critical synergy between POS features and CRF decoding: while POS injection alone yields marginal gains, coupling it with structural constraints produces our best single-model result (70.60% F1).
- We show that ensembling models trained with diverse objectives (Cross-Entropy and Focal Loss) improves discontinuous MWE detection (+2.33 F1 in French).
- Our official submission achieves 48.39% global F1, ranking second in the Shared Task.

2 Related Work

Automatic identification of Multi-Word Expressions (MWEs) remains a core challenge in multilingual NLP due to their idiomaticity, non-

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 154–159
March 28, 2026 ©2026 Association for Computational Linguistics

compositional semantics, discontinuity, and annotation sparsity. Foundational linguistic characterisation and formal language perspectives on alternating sequence computation and structured labeling complexity trace back to early formal studies such as alternation theory and lexical systematisation, which later informed NLP-oriented taxonomies for MWEs and decoding principles (Chandra et al., 1981; Baldwin and Kim, 2010). A broad survey of MWE processing highlights persistent issues, including sparse observations, idiomaticity, discontinuity, and cross-lingual variation, motivating architectures that combine contextual representations with explicit linguistic inductive biases.

A large body of work formulates MWE identification as structured sequence labeling, adopting token-level tagging schemes (e.g., BIO/BILOU) and log-linear structured decoders. BiLSTM-CRF models established strong baselines for enforcing tag consistency and segment boundaries in linear-chain CRF formulations with Viterbi/Viterbi-style inference (Lample et al., 2016; Ma and Hovy, 2016; Lafferty et al., 2001). Contextualized Transformer encoders, especially BERT, have since become the dominant representation backbone for sequence labeling tasks, including MWEs (Devlin et al., 2019). However, even with contextual encoders, tagging systems remain sensitive to (i) extreme class imbalance (most tokens are `O`), (ii) discontinuous MWEs, and (iii) recall and generalization to unseen expressions, particularly in multilingual blind-test scenarios.

The PARSEME shared task series created standardized multilingual corpora, annotation guidelines, and evaluation protocols for verbal MWEs (VMWEs), foregrounding both continuous and discontinuous expressions and enabling systematic cross-lingual evaluation (Savary et al., 2017). Later task editions emphasised generalization to unseen VMWEs through carefully constructed blind-test splits containing expressions not observed during training (Ramisch et al., 2020). Competitive neural systems commonly augment taggers with linguistic or structural signals: ERMI injects POS and dependency features in a BiLSTM–CRF architecture (Yirmibeşoğlu and Güngör, 2020), while MTLB-STRUCT frames VMWE identification under a multi-task paradigm by incorporating auxiliary syntactic structure and using a dual tagging head on multilingual BERT with CRF decoding, analyzing imbalance-aware objectives such as focal loss and the role of structured decoders for improved discontinuity resolution and unseen-expression recall (Taslimipoor et al., 2020). Explicit long-range relation modelling for bridging syntactic gaps was studied in discontinuity-focused settings (Rohanian et al., 2019), motivating dependency-based syntactic path reasoning to connect separated MWE components. Relational syntactic inference in multilingual pipelines was further shaped by deterministic and biaffine dependency parsers and benchmarks, along with trainable parsing frameworks such as UDPipe (Nivre, 2008; Dozat and Manning, 2017; Straka et al., 2016). Beyond syntactic biasing, contrastive and self-supervised sequence objectives such as those introduced by (Jaiswal et al., 2021; Gao et al., 2021) strengthened general sequence representations and multi-task optimisation dynamics were later systematized in predictive structure learning and representation surveys (Ando and Zhang, 2005; Ruder, 2017).

Our system aligns with the Transformer-based sequence labeling framework, enhanced by POS injection for syntactic bias, CRF decoding for structural consistency, and Focal Loss to mitigate class imbalance. Additionally, we leverage ensembling to capture diverse error profiles, which is crucial for robust discontinuous and unseen MWE detection.

3 Methodology

We frame MWE identification as a sequence labeling problem requiring the detection of continuous and discontinuous expressions under heavy class imbalance. Our approach utilizes a multi-task architecture predicting both identification (BIO) and classification (Category) labels.

3.1 Submitted System

Our official submission employs `bert-base-multilingual-cased` as a shared encoder. We derive two aligned supervision signals: (1) **BIO tags** for identification (`B`, `I`, `O`), and (2) **MWE categories** for classification.

POS Injection and Dual-Head. We augment contextual embeddings by concatenating them with learned POS tag embeddings. As shown in Figure 1, this fused representation $\tilde{h}_i$ feeds into two parallel linear heads. The BIO head predicts identification tags, while the Category head assigns MWE types.

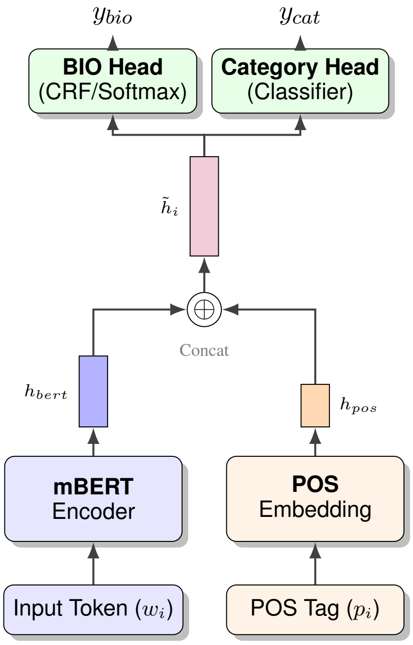

Figure 1: Architecture at a single time-step i. POS embeddings are concatenated with mBERT output to feed dual prediction heads.

3.2 Extensions

To improve robustness against multilingual interference and label imbalance, we investigate the following extensions.

3.2.1 CRF Decoding

To enforce valid label transitions, we replace the token-wise softmax with a Linear-Chain CRF. The model maximizes the score of the correct BIO sequence y, as defined in Equation 1.

$$\mathrm{Score}(y) = \sum_{i=1}^{n} \left(A_{y_{i-1},y_i} + s_{i,y_i}^{bio} \right), \qquad (1)$$

where A represents transition parameters. Inference is performed via Viterbi decoding.

3.2.2 Consistency via Masking

To align the two heads, we compute category loss only for tokens predicted as MWEs by the identification head. We apply a mask $m_i = \mathbb{I}(\hat{y}_i^{bio} \neq \mathtt{O})$ to the category loss, ensuring that the classifier focuses only on valid MWE candidates and ignores the majority O class.

3.2.3 Focal Loss

To address the dominance of non-MWE tokens, we employ Focal Loss (FL) to down-weight easy negatives and focus training on hard examples, utilizing the formulation in Equation 2.

$$\mathrm{FL}(p_t) = -\alpha(1-p_t)^{\gamma} \log(p_t). \qquad (2)$$

We compare FL against standard Cross-Entropy (CE) to evaluate its impact on recall.

3.2.4 Ensemble Strategy

We construct ensembles by aggregating predictions from K diverse models (Base, POS-injected, and Focal Loss variants). For **category classification**, we simply average the probability distributions according to Equation 3.

$$\bar{p}_i^{cat} = \frac{1}{K} \sum_{k=1}^{K} p_{i,k}^{cat}. \qquad (3)$$

For **BIO identification**, we use a hybrid scheme: standard softmax models use probability averaging, while CRF-based models use **majority voting** on Viterbi sequences to preserve structural validity. This strategy combines the high recall of Focal Loss with the precision of standard baselines.

4 Experimental Results

We evaluate our proposed system on the PARSEME 2.0 blind test set, analyzing global metrics, language-specific performance, and error categories.

4.1 Submitted System Results

Table 1 compares our system against the multilingual BERT baseline.

Impact of POS Injection. Injecting POS features improved Global F1 to **48.39%**, driven by a gain in **Recall (+2.78 points)**. This indicates that explicit morphosyntactic information aids in recognizing valid candidates missed by the baseline.

Generalization & Memorization. We observe a performance disparity between *Seen* and *Unseen* MWEs. The POS-enhanced model excelled at **Seen MWEs** (73.70% F1, +1.64% gain), suggesting that POS tags reinforce confidence in learned syntactic templates (e.g., *Verb+Noun*). However, performance on **Unseen MWEs** remained low (∼20%), indicating that while explicit syntax aids pattern matching for known expressions, it offers limited benefit for zero-shot generalization.

Configuration	Global MWE-based			Global Token-based			Generalization (F1)	
	Prec.	Rec.	F1	Prec.	Rec.	F1	Seen	Unseen
Base BERT	**46.49**	48.55	47.50	**61.95**	53.83	57.61	72.06	**20.23**
POS Features + BERT	45.77	**51.33**	**48.39**	61.53	**57.12**	**59.24**	**73.70**	19.98

Table 1: Blind test set results. POS injection improves Global F1 and Recall, particularly for Seen MWEs.

Lang	Family	P	R	F1
High Performance				
FA	Indo-Iranian	70.67	77.29	73.83
JA	Japonic	75.92	70.00	72.84
RO	Romance	61.98	71.12	66.23
Mid Performance				
PL	Slavic	51.84	64.80	57.60
HE	Semitic	52.89	60.28	56.34
FR	Romance	52.60	50.50	51.53
Low Performance				
KA	Kartvelian	26.17	69.40	38.01
EGY	Semitic	33.67	13.20	18.97
GRC	Hellenic	8.81	6.01	7.14

Table 2: MWE-based F1 scores for representative languages.

4.1.1 Language-Specific Analysis

Table 2 details performance across diverse language families. High-resource languages with distinct syntactic markers (FA, JA) achieved the highest scores ($>72\%$ F1). Conversely, low-resource or ancient languages (GRC, EGY) suffered from data sparsity. Notably, Georgian (KA) exhibited high recall but low precision (26.17%), suggesting systematic over-prediction likely driven by severe class imbalance and language-specific noise.

4.2 Ablation Studies: Extensions

We analyze extensions on the development set using a 90/10 split, focusing on two setups: **Monolingual** (French only) and **Multi-5** (FR, SV, EL, FA, JA).

4.2.1 Monolingual Case Study (French)

Table 3 highlights the interaction between linguistic features and decoding strategies.

Configuration	Global F1	Disc. F1
Single Models		
Base (mBERT)	67.59	55.65
pos	67.43	51.98
pos_crf	**69.28**	55.32
crf	68.06	52.94
crf_focal	67.61	53.28
pos_crf_focal	69.16	**56.41**
Ensemble Models		
base+pos_crf+crf_focal	69.50	54.55
base+pos_crf+pos_crf_focal	**69.92**	**57.98**

Table 3: Ablation on French (FR) development set.

Configuration	Global F1	Disc. F1
Single Models		
Base (mBERT)	69.84	47.16
pos	69.22	47.16
pos_crf	**70.60**	**47.67**
crf	69.45	45.29
crf_focal	69.39	46.19
pos_crf_focal	69.18	46.37
Ensemble Models		
base+pos_crf+crf_focal	**70.73**	**48.88**
base+pos_crf+pos_crf_focal	70.22	48.06

Table 4: Ablation on Multi-5 (FR, SV, EL, FA, JA) set.

POS-CRF Synergy. Injecting POS tags alone (pos) slightly degraded performance. However, coupling POS with CRF decoding (pos_crf) yielded the best single-model result (69.28%). This indicates that while POS tags provide valuable signals, the model requires the structured transition constraints of a CRF to utilize them effectively without overfitting.

Discontinuity via Ensembling. Single models struggled with discontinuous MWEs. However, the ensemble approach achieved a Discontinuous F1 of **57.98%** (+2.33 points over baseline). Averaging probability distributions from diverse models (Base + POS + Focal) effectively bridges syntactic gaps that single architectures miss.

4.2.2 Multilingual Analysis (Multi-5)

Table 4 summarizes the multilingual ablation results.

Synergy Consistency. Consistent with monolingual findings, pos_crf achieved the highest single-model F1 (70.60%), reversing the degradation seen with POS alone. This confirms that CRF constraints are essential for leveraging morphosyntactic cues across diverse languages.

Focal Loss & Diversity. While Focal Loss models underperformed in isolation, they were critical for the ensemble. The best system (Base + POS + Focal) reached **70.73% Global F1** and the highest **Discontinuous F1 (48.88%)**, confirming that diverse training objectives capture "hard" examples that standard models fail to detect.

5 Conclusion

We presented a multilingual MWE identification architecture combining mBERT with explicit POS features, which ranked second in the PARSEME 2.0 Shared Task (48.39% F1). Our experiments demonstrate a critical synergy between morphosyntactic features and structured decoding: while POS injection alone yields marginal gains, coupling it with a CRF layer effectively constrains the output space, achieving our best single-model performance. Furthermore, addressing the challenge of discontinuity, we showed that ensembling models trained with Focal Loss improves recall on non-adjacent expressions. Future work will further explore the integration of linguistic constraints into end-to-end training.

Limitations

Our study has several limitations that are important for interpreting the results and for guiding future improvements.

Reliance on POS quality. POS feature injection is beneficial when tags are accurate and consistent across languages; however, in low-resource or morphologically complex languages, tagging errors may propagate into MWE predictions and lead to unstable precision/recall trade-offs.

Unseen MWE generalization remains difficult. While our approach improves recall for seen expressions, performance on unseen MWEs remains a major bottleneck, suggesting the model still relies on distributional regularities observed during training rather than type-level constraints or composition-aware cues.

Token-level formulation and discontinuity. We use token-level BIO supervision and CRF decoding, which enforces local label consistency, but we do not explicitly model expression-level completeness or gap-aware structure. This can yield fragmented boundaries, particularly for discontinuous MWEs, where explicit gap modeling or syntactic integration may be necessary.

Compute and deployment cost. Ensembling improves robustness and discontinuous detection, yet increases inference time and memory. Distillation or lightweight diversity-preserving alternatives could make the approach more deployable.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing*, pages 267–292. CRC Press.

Ashok K. Chandra and 1 others. 1981. Alternation and structured sequence complexity. *Journal of the ACM*, 28:114–133.

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

Timothy Dozat and Christopher Manning. 2017. Deep biaffine neural dependency parser. In *ICLR*, pages 1–12. OpenReview.

Tianyu Gao and 1 others. 2021. Simcse: Simple contrastive learning of sentence embeddings. *EMNLP*, arXiv. Contrastive sequence representation learning.

Ashish Jaiswal and 1 others. 2021. Survey on contrastive learning. *Journal of AI Research*, Survey. Systematization of self-supervised contrastive objectives.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*, arXiv. Linear-chain CRF with Viterbi inference.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Italy.

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

Joakim Nivre. 2008. Deterministic dependency parsing algorithms. In *ACL*, pages 513–520. Association for Computational Linguistics.

Carlos Ramisch and 1 others. 2020. Edition 1.2 of the parseme shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118. Association for Computational Linguistics.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2692–2698.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv*. Survey on multi-task learning and training diversity.

Agata Savary and 1 others. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 31–47. Association for Computational Linguistics.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Milan Straka and 1 others. 2016. Udpipe: Trainable pipeline for multilingual dependency parsing. In *LREC*, pages 4290–4297. European Language Resources Association.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. Ermi at parseme shared task 2020: A sequence labeling approach. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 140–144.