Layer Freezing and Cost-Sensitive Learning for English Historical Relation Classification

Jadavpur University at CLEF-HIPE 2026

Rittik Ram[1,*], Souvik Soren[1] and Sivaji Bandyopadhyay[1]

[1]*Department of Computer Science and Engineering, Jadavpur University, Kolkata, India*

Abstract

This paper details our approach for the CLEF-HIPE 2026 shared task focusing on person-place relation classification from noisy 19th-century English archival text. Initial evaluations of our unweighted baseline BERT sequence classifier revealed a vulnerability to the dataset's moderate 73/27 class imbalance, resulting in a degenerate optimization shortcut (yielding an initial baseline macro recall of 0.0000). To stabilize the decision boundary and resolve this baseline collapse, we implement a three-tier cascaded intervention: Sentence-Pair QA Prompting, Layer Freezing, and a Cost-Sensitive Cross-Entropy loss formulation. Our proposed approach successfully breaks free from the majority-class shortcut, raising the binary physical presence (isAt) macro recall to 0.5249. We also present a hypothesis on the bottlenecks limiting the multi-class epistemic certainty (at) task, specifically regarding OCR fragmentation.

Keywords

Relation Classification, Layer Freezing, Cost-Sensitive Learning, CLEF-HIPE 2026

1. Introduction

The automatic extraction of semantic relationships from historical texts is a complex challenge, primarily due to the typographical degradation and optical character recognition (OCR) noise inherent in digitized archival documents. This paper details the system submitted by Jadavpur University (Team 16) for the CLEF-HIPE 2026 shared task, which focuses on identifying spatial relationships between person and location entities in historical corpora.

Our system is designed exclusively for the English-language track (impresso-en). Because the shared task infrastructure provides pre-annotated candidate person-location pairs, our approach is framed strictly as a relation classification task over given pairs, rather than end-to-end relation extraction.

A significant hurdle in this specific domain is the natural class distribution skew of historical linkages. We present an empirical study of a transformer-based sequence classifier, detailing how standard unweighted optimization collapses under moderately imbalanced conditions. To resolve this, we propose a regularized intervention strategy utilizing parameter-efficient layer freezing and a cost-sensitive loss formulation.

2. System Architecture and Data Flattening

The development of our relation classification engine followed an empirical, iterative cycle. The initial phase focused on constructing a data ingestion pipeline to translate nested competition schemas into flat relational structures for the baseline model.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

hrittickramspi40@gmail.com (R. Ram); souviksoren645@gmail.com (S. Soren); sivaji.cse.ju@gmail.com (S. Bandyopadhyay)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2.1. Ingestion Framework and Cross-Product Generation

The official CLEF-HIPE 2026 dataset was delivered in a hierarchical JSON Lines (.jsonl) format. Each document entity consists of a high-level text string representing an archival newspaper block, accompanied by a nested array of annotated entity mentions. To make this data readable by a sequence classification model, a deterministic data flattening pipeline was engineered in Python.

For every document D, the pipeline isolates the complete string corpus and extracts the distinct boundary mentions. The engine constructs an intra-document cross-product, systematically pairing every isolated person mention (P) against every valid location mention (L) present within that specific textual boundary:

$$\mathcal{M}_{pairs} = \{(P_i, L_j) | P_i \in D \wedge L_j \in D\} \tag{1}$$

Critically, each cross-product person-location pair forms its own independent input sample for the model. For each generated pair, the categorical ground-truth labels for physical presence (isAt) and modal certainty (at) are converted into discrete numerical tensor indexes. The spatial target isAt is mapped to binary values $\in \{0, 1\}$, while the multi-class certainty task (at) is mapped sequentially using a standard case structure:

$$\text{mapping}(at) = \begin{cases} 0 & \text{if } at = \text{"FALSE"} \\ 1 & \text{if } at = \text{"PROBABLE"} \\ 2 & \text{if } at = \text{"TRUE"} \end{cases} \tag{2}$$

Pairs containing null tokens or uninstantiated entity pointers are structurally dropped during the stream, ensuring a clean, tabular dataset format ready for model ingestion.

2.2. Baseline Architecture and Tokenization Constraints

The initial model prototype was configured as a standard sequence classification network utilizing the pre-trained bert-base-cased language engine (110M parameters). A cased model variant was selected because case markings provide essential semantic indicators for identifying historical proper nouns and geographical entities amidst highly degraded text.

To address a critical requirement of relation classification, the model must be explicitly conditioned on the specific candidate pair rather than just the background text. To achieve this, the target person and location strings are concatenated and passed as a secondary sequence to the WordPiece tokenizer alongside the primary historical text context.

To manage the long-form nature of historical newspaper passages and avoid quadratic memory spikes on standard GPU environments, sequences are processed using defensive truncation and padding limits:

Listing 1: Baseline Ingestion and Tokenization Mapping

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")

def tokenize_baseline_pairs(examples):
    # Concatenating target entities to condition the prediction on the specific pair
    target_pairs = [p + " [SEP] " + l for p, l in zip(examples["person_target"], examples["
        location_target"])]

    # Enforcing strict max_length to eliminate quadratic memory spikes
    return tokenizer(
        examples["text_context"],
        target_pairs,
        max_length=512,
        truncation=True,
        padding="max_length"
    )
```

This engineering constraint locked the token sequence length to 512, preventing memory exhaustion while ensuring the classifier explicitly evaluated the spatial relationship between the defined entities.

2.3. Empirical Analysis of the Baseline Model and Degenerate Optimization

Once the data processing pipeline was active, the baseline model was trained using standard empirical risk minimization under an unweighted Cross-Entropy loss objective. The initial validation runs exposed a critical technical vulnerability regarding the class distribution.

Statistical analysis of the flattened training matrix revealed a moderate class distribution skew: approximately 73% of all extracted entity pairs were labeled as negative relations ($isAt = 0$), leaving a 27% minority split for true presence markers ($isAt = 1$).

Table 1
Class Distribution Skew in Preprocessed Training Set

Relation Label (isAt)	Total Extracted Pairs	Percentage of Dataset
0 (FALSE / No Relation)	224	72.96%
1 (TRUE / Active Presence)	83	27.04%

As seen in Table 1, when exposed to this distribution, the unweighted Cross-Entropy loss function allowed the network to default to a degenerate shortcut. The classifier minimized overall loss simply by predicting the majority class for almost every pair. While this yielded a deceptively stable raw accuracy floor of approximately 73%, the system's macro recall for identifying true historical associations collapsed to exactly 0.0000. This initial empirical failure proved that standard out-of-the-box pipelines cannot navigate moderately skewed data environments without explicit architectural adjustments.

3. Contextual Prompt Engineering and Parameter Optimization

To break past the limitations of standard text-pair classification over a skewed data distribution, we restructured the input mechanics to explicitly guide the model's self-attention layers through a Question-Answering (QA) framework.

3.1. Sentence-Pair QA Restructuring

Rather than expecting the sequence classifier to implicitly learn entity linkages from raw unformatted text blocks, the target task was reformulated. For every extracted person-location pair, a deterministic natural language query was dynamically generated to act as a semantic constraint (e.g., "Question: Is [Person] physically located in [Location]?").

Listing 2: Sentence-Pair Sequence Formatting

```
def preprocess_qa_pairs(examples):
    # Generating targeted natural language queries for Segment A
    questions = [
        f"Question:_Is_{p}_physically_located_in_{l}?"
        for p, l in zip(examples["person_target"], examples["location_target"])
    ]

    # Passing dual sequences to automatically compute segment token_type_ids
    return tokenizer(
        questions,
        examples["text_context"],
        max_length=512,
        truncation=True,
        padding="max_length"
    )
```

This allowed the WordPiece tokenizer to handle sequence formatting natively. The encoder automatically mapped the token type identifiers, ensuring the self-attention heads could structurally differentiate the query intent from the historical background noise.

3.2. The Overfitting Bottleneck

With the sentence-pair prompt infrastructure established, an experimental run was executed to fine-tune all 110 million parameters of the base model. Optimizing the entire parameter space on a restricted, low-resource historical training set (307 rows) caused immediate overfitting. By the end of the 3rd training epoch, the training loss dropped to 0.0214, while the validation loss spiked to 1.8942. This structural failure demonstrated that full-parameter updates are non-viable for this specific low-resource task, establishing the necessity for parameter-efficient fine-tuning.

4. Proposed Architecture: Layer Freezing and Cost-Sensitive Learning

To simultaneously combat overfitting and majority-class distribution bias, the final system architecture incorporates a parameter-efficient transfer learning model optimized via a cost-sensitive objective function.

4.1. Layer Freezing

Because the historical relation extraction corpus is constrained, updating the entire weight space allows the transformer layers to overwrite their generalized linguistic representations. To resolve this, a layer freezing strategy was engineered.

The lower embedding layer and the initial ten transformer encoder layers (`layer.0` through `layer.9`) were structurally locked, preventing their weights from receiving gradient adjustments during backpropagation. Optimization was forced exclusively into the deepest semantic abstraction modules (layers 10, 11, and the classification head).

Listing 3: Layer Freezing Implementation

```
# Freeze foundational layers to prevent catastrophic forgetting
for name, param in model.bert.named_parameters():
    if not any(target in name for target in ["layer.10", "layer.11"]):
        param.requires_grad = False
```

4.2. Cost-Sensitive Loss Formulation

To override the degenerate shortcut observed in the baseline, the optimization objective was mathematically recalibrated by mapping a cost-sensitive class weight tensor directly into the loss function.

Since the negative class is roughly 2.7 times more frequent than the positive class, we established a 3:1 weighting heuristic. This acts as a direct mathematical counterbalance against the majority class. Formally, the custom weighted Cross-Entropy loss function $\mathcal{L}_{weighted}$ applies a weight vector w configured explicitly as $w_0 = 1.0$ and $w_1 = 3.0$. This heavily penalizes the network's backpropagation path whenever it misses a rare active historical presence marker.

Handling the Multi-Class (at) Target: The shared task also requires predicting the three-class epistemic certainty (at) label. This is modeled using a shared classification head. Because the at relation is only logically valid when physical presence is confirmed, the 3:1 cost-sensitive penalty is applied strictly to the binary isAt dimension to ensure minority recall at the base level. The at logits are evaluated using a standard unweighted cross-entropy loss.

5. Empirical Evaluation and Error Analysis

The proposed model was deployed on the official testing corpus for the CLEF-HIPE 2026 impresso-en track.

As shown in Table 2, the model secured an isAt macro recall of 0.5249, validating that the cost-sensitive framework successfully broke free from the 73% majority negative class skew to isolate spatial connections. However, performance dropped significantly on the multi-class epistemic certainty task, yielding an at macro recall of 0.3258, which pulled the overall tracking Profile Score down to 0.4254 (Rank 45 out of 47).

Table 2
Official Leaderboard Performance Profiles for Team 16

Submission	Profile Score	isAt Acc.	isAt Macro Recall	at Acc.	at Macro Recall
Team 16 (Run 1)	0.4254	0.5617	0.5249	0.3889	0.3258

5.1. Error Analysis Hypothesis

To understand the performance degradation on the testing distribution, we hypothesize two primary systemic bottlenecks:

1. **WordPiece Tokenization Fragmentation:** The historical newspaper corpus contains heavy typographical corruption. We hypothesize that when the standard tokenizer encounters these distorted character sequences, it aggressively fractures individual entity mentions into multiple sub-tokens. This fragmentation likely dilutes the spatial coordinates of positional embeddings.
2. **Epistemic Modality Ambiguity:** The multi-class classification target requires the network to differentiate subtle linguistic nuances between absolute historical truth and speculative documentation. The localized classification head lacks the structural capacity to decode abstract contextual hedging patterns without an external historical knowledge source.

6. Conclusion

This research successfully designed and deployed an end-to-end regularized, cost-sensitive relation classification framework for the CLEF-HIPE 2026 Shared Task [1, 2]. By executing layer freezing and implementing a 3:1 class-weighted loss formulation, the system successfully stabilized the binary classification boundary against a moderate dataset skew, raising the isAt macro recall from 0.0000 to 0.5249. However, we acknowledge that the proposed strategy did not lead to competitive performance on the global leaderboard, primarily due to struggles with the multi-class epistemic certainty target.

Future iterations will explore hybrid architectures incorporating Knowledge Graph Embeddings (e.g., Wikidata QIDs) to shield the text encoders from OCR typographical noise and provide deeper contextual reasoning.

Declaration on Generative AI

During the preparation of this work, the author(s) used Google Gemini in order to: Edit grammar, format the LaTeX manuscript to comply with CEURART guidelines, and refine academic terminology based on peer review feedback. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, S. Clematide, Overview of hipe-2026: Person-place relation extraction from multilingual historical texts, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. B.-C. no, A. G. S. de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science (LNCS), Springer, 2026.

[2] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, S. Clematide, Extended overview of hipe-2026: Evaluating accurate and efficient person-place relation extraction from multilingual historical texts, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS, 2026.