HIPE-2026: Person–Place Relation Extraction from Multilingual Historical Texts

Project Report

Fomin Bogdan
Lungu Andrei

January 15, 2026

Abstract

HIPE-2026 is a CLEF 2026 evaluation lab focused on extracting and qualifying person–place relations in multilingual historical documents. Given a document with (deduplicated) person and place entities and noisy OCR text, systems must classify each possible (person, place) pair for two relation types: `at` (the person ever visited/resided in the place before publication) and `isAt` (the person is located at the place in the immediate temporal context of the document), each with evidence labels `TRUE`, `PROBABLE`, or `FALSE`. The lab emphasizes both accuracy (macro-averaged recall / balanced accuracy) and efficiency (model size and cost), plus a surprise out-of-domain generalization profile. We present two implemented transformer-based approaches: (i) a prompt-style dual-head classifier built on XLM-RoBERTa with data augmentation and oversampling, and (ii) a Multiple Instance Learning (MIL) dual-head model that aggregates evidence across multiple context windows for each person–place pair. Experiments were conducted on the HIPE-2026 *sandbox* dataset, and we report exploratory data analysis, training settings, and development-set results, followed by a discussion of findings and next steps.

Contents

1 Problem Description

1.1 Task definition

HIPE-2026 evaluates **person–place relation extraction** from multilingual historical texts. For each document, organizers provide: (i) the full OCR text (with possible errors), (ii) a list of person entities, (iii) a list of place entities, and (iv) metadata such as language and publication date. Participants must return predictions for *all possible* triples (p, ℓ, r) where p is a person, ℓ a place, and $r \in \{\texttt{at}, \texttt{isAt}\}$. Each triple is labeled independently with: `TRUE` (strong evidence), `PROBABLE` (plausible inference), or `FALSE` (no evidence or contradiction).

The relation types are:

- `at`: Did the person ever reside in or visit the place prior to the document's publication?
- `isAt`: Is the person located at the place in the **immediate temporal context** of the document?

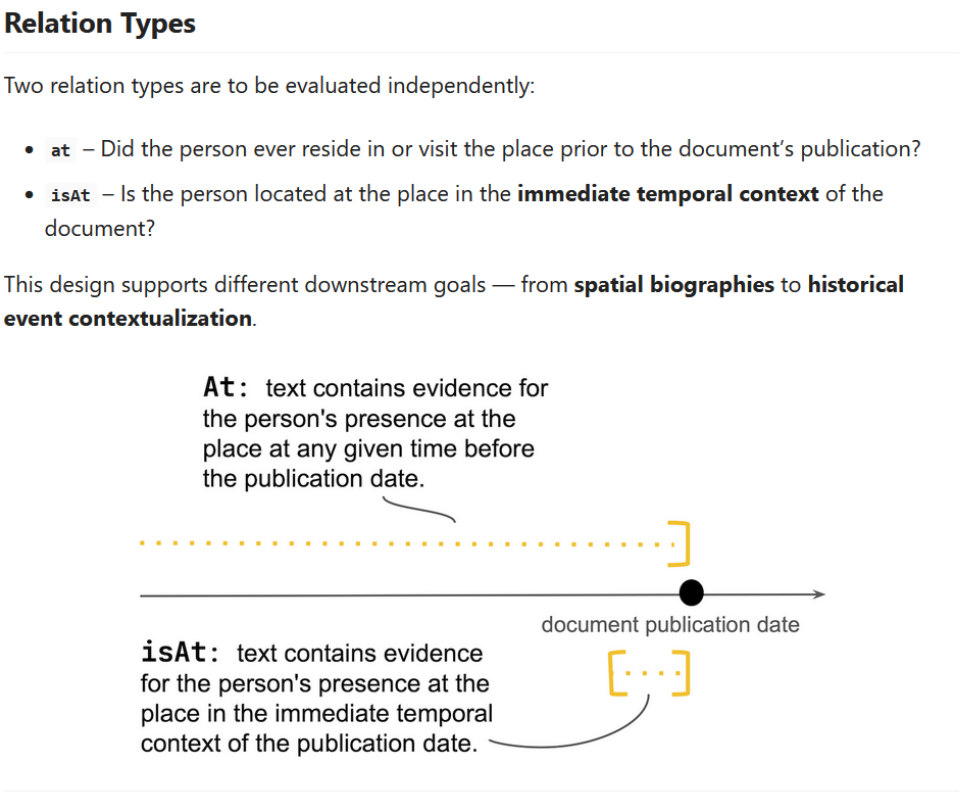

Figure 1: HIPE-2026 relation types. `at` refers to evidence of presence at any time prior to publication, while `isAt` refers to evidence in the immediate temporal context of the publication date.

1.2 Why simple co-occurrence is insufficient

Historical documents (especially OCRed newspapers) contain dense mentions of people and places, quotations, indirect references, and temporally shifted narratives (e.g., "was born in ...", "arrived yesterday from ..."). A person and a place can co-occur without a valid relation. HIPE-2026 thus requires semantic and temporal reasoning beyond co-occurrence.

1.3 Evaluation profiles

HIPE-2026 defines multiple profiles:

- **Accuracy profile:** macro-averaged recall (balanced accuracy) per relation type; systems are ranked accordingly.
- **Efficiency profile:** considers accuracy jointly with cost indicators such as model size, inference time, and hardware usage.
- **Generalization profile:** a surprise out-of-domain test set (Test Set B) evaluates robustness beyond historical news.

1.4 Datasets (official description)

Two official test settings are described by the organizers:

- **Development & Test Set A:** derived mostly from HIPE-2022 datasets; languages include French, German, English, and Luxembourgish; relations are validated for pre-annotated person/place entities.
- **Surprise Test Set B:** a 16th–18th century French literary corpus with relation labels restricted to `at`, used to test domain generalization.

2 Dataset Format Example

To make the task concrete, we include a compact real example showing (i) document metadata, (ii) a short OCR evidence snippet, and (iii) labeled person–place pairs. We only keep the fields relevant to relation modeling: *locations, evidence text, and* `at/isAt` *labels.*

2.1 Document metadata (minimal)

- **document_id:** `sn83026170-1820-05-05-a-i0004`
- **language:** `en`
- **publication date:** `1820-05-05`
- **publication:** *Alexandria Gazette & Daily Advertiser* (Newspaper; time period 1817–1822)

2.2 Evidence snippet (OCR text, shortened)

The strongest direct evidence linking a person to a place is the signature line near the end of the document:

```
...  Actuated by these views, the subscriber proposes to publish A SPLENDID EDITION
...
JOHN BINNS, Chesnut-street, Philadelphia.
```

2.3 Relevant labeled pairs (locations + labels)

Table 1 summarizes a subset of the provided sampled pairs, focusing only on the relation labels and locations.

Table 1: Minimal subset of labeled pairs for `sn83026170-1820-05-05-a-i0004` (task-relevant fields only).

Person (mention)	Location (mention)	`at`	`isAt`
JOHN BINNS	Philadelphia	TRUE	TRUE
JOHN BINNS	Chesnut-street	TRUE	TRUE
King John	Great Britain	PROBABLE	FALSE
King Henry	Great Britain	PROBABLE	FALSE
Sir Wiiliam Blackstono	Great Britain	PROBABLE	FALSE
King John	Philadelphia	FALSE	FALSE

Interpretation. The signature line explicitly associates `JOHN BINNS` with `Chesnut-street, Philadelphia`, supporting both a broad pre-publication relation (`at`) and an immediate-context relation (`isAt`) (Figure 1). In contrast, historical figures (`King John`, `King Henry`) and `Great Britain` appear in a comparative discussion; this can justify a weaker `at=PROBABLE` association while remaining `isAt=FALSE`. Many other co-occurring pairs are labeled `FALSE`.

3 Reference Papers Mentioned by the Task

This section summarizes key papers referenced by HIPE-2026 and how they motivate the current task.

3.1 Ehrmann et al. (2020): Extended Overview of CLEF HIPE 2020

HIPE-2020 introduced a systematic evaluation of named entity recognition/classification (NERC) and entity linking (EL) on **multilingual, diachronic historical newspapers**. It emphasizes challenges such as OCR noise, historical spelling variation, entity drift (places that change names/borders), and limited knowledge-base coverage—factors that also make relation extraction challenging and motivate specialized evaluation resources.

3.2 Ehrmann et al. (2022): Extended Overview of HIPE-2022

HIPE-2022 broadened HIPE with more languages, heterogeneous schemes, and additional document types. A central message is that performance often degrades under changes in language, genre, OCR quality, and annotation guidelines. HIPE-2026 extends the scope from entity-level tasks to **relation-level** reasoning between entities.

3.3 Ehrmann et al. (2021): Survey on NER for historical documents

This survey consolidates the major issues in historical NLP: heterogeneity, noise, scarcity of labeled data, and temporal language change. It reviews strategies like normalization, domain-adaptive pretraining, and multilingual transfer, which remain relevant in HIPE-2026 because relation models depend on robust representations under OCR and diachronic variation.

3.4 Ehrmann et al. (2016): Diachronic evaluation of NER on old newspapers

This work shows that NER accuracy varies across historical time slices and motivates **diachronic robustness**. For HIPE-2026, robustness is required not only across decades but also across narrative time within the same document (`isAt` vs. `at`).

3.5 Hamdi et al. (2021): NewsEye multilingual dataset

Hamdi et al. present a multilingual dataset for historical newspapers and stress reproducible benchmarks in cultural-heritage contexts. HIPE-2026 fits into this ecosystem by creating supervised person–place relation labels on top of multilingual historical texts.

4 State of the Art

4.1 From sentence-level to document-level relation extraction

Transformer-based relation extraction typically classifies a relation r between two entities e_1, e_2 using contextual representations. In sentence-level RE, a single sentence contains both entities, and a pooled embedding (e.g., [CLS]) is used for classification. HIPE-2026 is closer to **document-level RE** because: (i) evidence can appear across sentences, and (ii) multiple mentions of an entity may exist with differing relevance.

4.2 DocRED (Yao et al., 2019): document-level RE with entity aggregation

Core contribution. DocRED introduced a benchmark where relations may require **multi-sentence reasoning**. The key modeling pattern is: encode the whole document, aggregate representations for each entity from all its mentions, and score each entity pair.

Typical architecture.

1. Encode the document with a transformer.
2. For each entity, pool mention span embeddings into an **entity vector** (e.g., max/attention pooling).
3. For each entity pair, compute a pair representation and apply a classifier (often bilinear + MLP).
4. Optionally add structured reasoning (graphs over mentions/sentences).

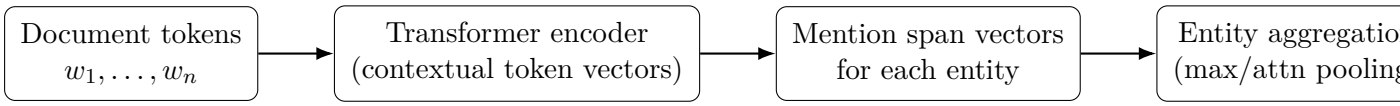

Figure 2: DocRED-style document-level RE: encode document, aggregate entity mentions, classify entity pairs.

Relevance to HIPE-2026. HIPE-2026 similarly scores *many* entity pairs per document. DocRED motivates two practical ideas:

- **Mention aggregation:** use all occurrences of the person/place rather than a single local window.
- **Shared encoding:** encode a document once and reuse representations for many pairs (important for efficiency).

4.3 Entity marker methods (Baldini Soares et al., 2019): "Matching the Blanks"

Core contribution. Entity marker methods show that inserting explicit boundary tokens around the arguments often yields strong gains in RE because attention can directly focus on the two entity spans.

Typical architecture.

- Insert markers around entity mentions.
- Encode with a transformer.
- Represent the pair using marker token embeddings (e.g., concatenate contextual vectors at marker positions).
- Classify with an MLP.

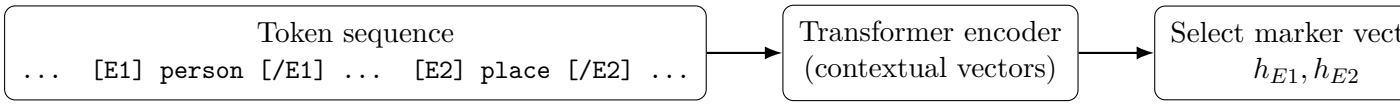

Figure 3: Entity-marker RE: mark arguments, encode, classify using marker representations.

Relevance to HIPE-2026. While entity markers help in noisy OCR, HIPE-2026 frequently requires evidence distributed across multiple sentences or mentions. This motivates Multiple Instance Learning (MIL): instead of trusting a single window, models aggregate multiple candidate evidence snippets for the same person–place pair.

4.4 Prompting and generation for relation extraction

LLM prompting and sequence-to-sequence formulations can generate relation labels or evidence snippets and may generalize better across domains. However, cost and reproducibility can be challenging, which is why HIPE-2026 includes an **efficiency profile**. Our Approach A borrows an instruction-like prompt format while remaining a compact fine-tuned encoder.

4.5 Historical-domain considerations

RE in historical OCRed text adds noise robustness, diachronic phrasing, and entity drift, motivating multilingual encoders (XLM-R), augmentation, and explicit argument marking.

5 Implemented Approaches

We implemented two transformer-based systems for multi-task relation classification. Both share the same objective:

$$\hat{y}^{(\texttt{at})} = f_\theta^{(\texttt{at})}(x), \quad \hat{y}^{(\texttt{isAt})} = f_\theta^{(\texttt{isAt})}(x),$$

where x is a constructed input containing document context and the person/place arguments.

5.1 Common pipeline

Both approaches share the same end-to-end pipeline:

1. Parse JSONL documents and build candidate pairs (p, ℓ) by cartesian product of persons and places per document.
2. Locate evidence in the OCR text using entity offsets if available; otherwise fall back to string match.
3. Extract context windows around evidence.
4. Construct model input (prompt-style or bag-of-instances).
5. Predict `at` and `isAt` with a shared encoder and two heads.

5.2 Approach A: Prompt-style dual-head XLM-R classifier

5.2.1 Input format

`Person:` `{p}` `Place:` $\{\ell\}Context : \{window\}$

5.2.2 Oversampling

We use **WeightedRandomSampler** to upweight minority labels and reduce dominance of `FALSE` pairs.

5.2.3 Augmentation

With probability $p = 0.5$, we jitter the window center and truncate context to simulate evidence shifts and OCR inconsistencies.

5.2.4 Training

Implemented with Hugging Face `Trainer`: AdamW, linear warmup (0.06), dropout 0.2, early stopping on average macro recall.

5.2.5 Architecture

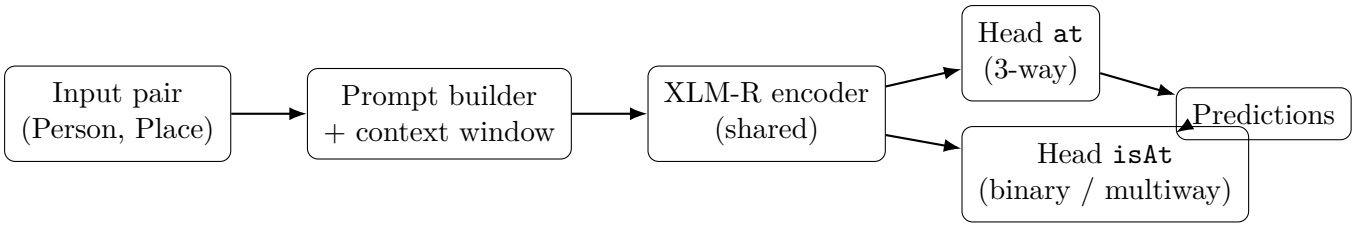

Figure 4: Approach A: prompt-style input, shared encoder, dual heads.

5.2.6 Implementation sketch

Listing 1: Approach A: dataset item and forward pass (simplified).

```
context = extract_window(doc_text, person, place, W=900)
text = f"Person: {person}  Place: {place}  Context: {context}"
enc = tokenizer(text, max_length=182, truncation=True, padding="
    max_length")

z = encoder(**enc).pooler_output  # or CLS
logits_at   = head_at(z)
logits_isAt = head_isAt(z)
loss = CE(logits_at, y_at) + CE(logits_isAt, y_isAt)
```

5.3 Approach B: MIL dual-head classifier (bag-of-windows evidence aggregation)

5.3.1 Motivation

A single fixed window can miss evidence when the person and location appear multiple times, or when relevant cues are spread across sentences (common in historical newspapers and OCR). We therefore model each person–place pair as a **bag of instances** (multiple candidate evidence windows) and use Multiple Instance Learning (MIL) to predict bag-level `at`/`isAt` labels.

5.3.2 Bag construction: extracting K evidence windows per pair

For each document and pair (p, ℓ), we identify multiple anchor positions based on mentions:

- Find up to m_p occurrences of the person mention(s) and up to m_ℓ occurrences of the location mention(s) (via offsets when available, otherwise string match).
- Form candidate anchors from mention pairs (e.g., midpoint between a person mention and a location mention), and optionally include anchors at each single mention location.
- Extract a fixed-size character window around each anchor (e.g., 600–900 chars).
- Keep the top K windows (instances) using a simple heuristic (e.g., shortest person–location distance, or earliest occurrences) and deduplicate near-identical windows.

Each instance is then formatted as:

`Person: {p} Place: {`ℓ`}`$Context : \{window\_k\}$

5.3.3 Encoder and MIL aggregation

Let a bag contain K instances $\{x_1, \ldots, x_K\}$. We encode each instance with a shared transformer:

$$h_k = \mathrm{Enc}_\theta(x_k) \in \mathbb{R}^d.$$

We then aggregate instance representations into a bag representation z using **attention pooling**:

$$\alpha_k = \frac{\exp(u^\top \tanh(W h_k))}{\sum_{j=1}^{K} \exp(u^\top \tanh(W h_j))}, \quad z = \sum_{k=1}^{K} \alpha_k h_k.$$

Finally, we apply two classification heads:

$$\mathrm{logits}_{\texttt{at}} = W_{\texttt{at}} z + b_{\texttt{at}}, \quad \mathrm{logits}_{\texttt{isAt}} = W_{\texttt{isAt}} z + b_{\texttt{isAt}}.$$

5.3.4 Training loss: bag-level weighted cross entropy

Labels are provided at the bag (pair) level. We compute class-weighted cross entropy per head:

$$\mathcal{L} = \mathrm{CE}(\mathrm{logits}_{\texttt{at}}, y_{\texttt{at}}; \mathbf{w}_{\texttt{at}}) + \lambda\, \mathrm{CE}(\mathrm{logits}_{\texttt{isAt}}, y_{\texttt{isAt}}; \mathbf{w}_{\texttt{isAt}}),$$

where $\mathbf{w}$ are inverse-frequency weights (smoothed), and $\lambda = 1$.

5.3.5 Inference

At inference, we build the same bag of K instances for each pair, compute z with MIL pooling, and output `at` and `isAt` predictions. Attention weights α_k can be used to identify which window(s) contributed most.

5.3.6 Architecture diagram

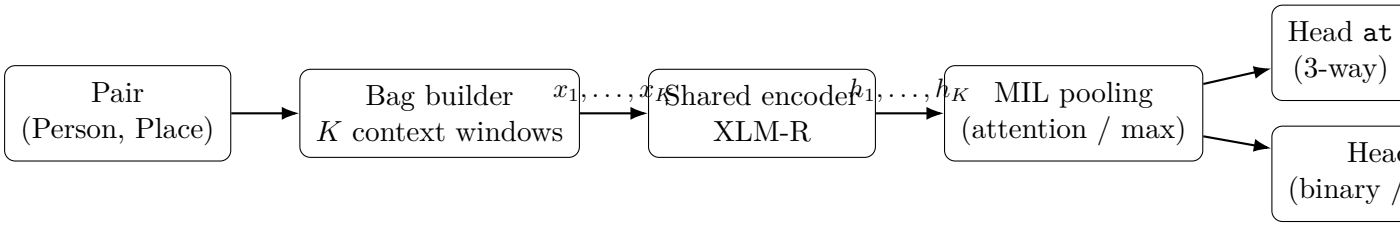

Figure 5: Approach B (MIL): multiple windows per pair, shared encoding, MIL pooling to a bag representation, then dual-head classification.

6 Experimental Results

6.1 Datasets used in this project

Experiments were run on the publicly released HIPE-2026 **sandbox** data. For model training, we combine available training subsets across languages and evaluate on a held-out English development subset.

6.2 Exploratory Data Analysis (EDA)

This section integrates the exploratory checks and summary statistics computed on the sandbox dataset.

6.2.1 Document-level integrity and coverage

- **Uniqueness:** all `document_id` values are unique (617 documents; 617 unique IDs).
- **Language distribution (documents):** French dominates the corpus (`fr`=424), followed by German (`de`=120) and English (`en`=73).
- **Temporal range:** earliest date is 1790-01-02 and latest date is 2018-01-16; 16 documents have missing dates.
- **Media schema:** each `media` entry contains `publication_title`, `time_period`, and `source_type`.

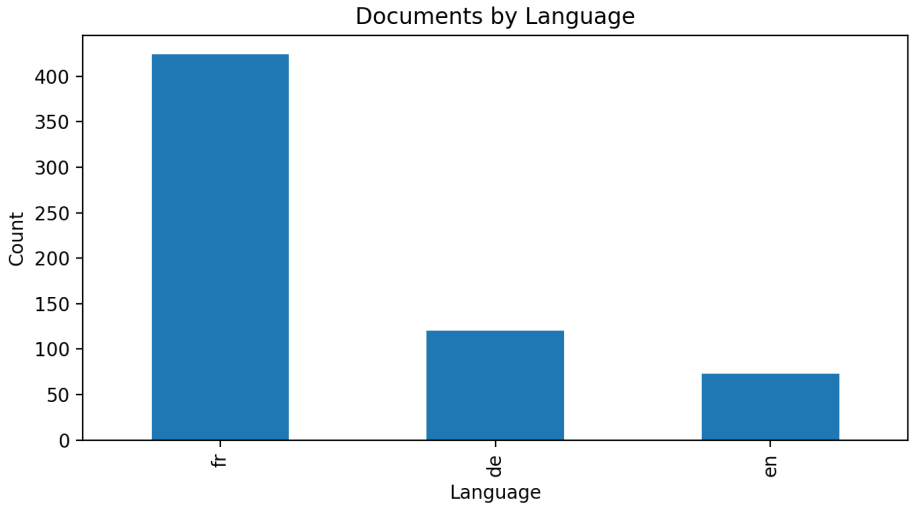

Figure 6: Distribution of documents by language in the sandbox dataset (617 documents).

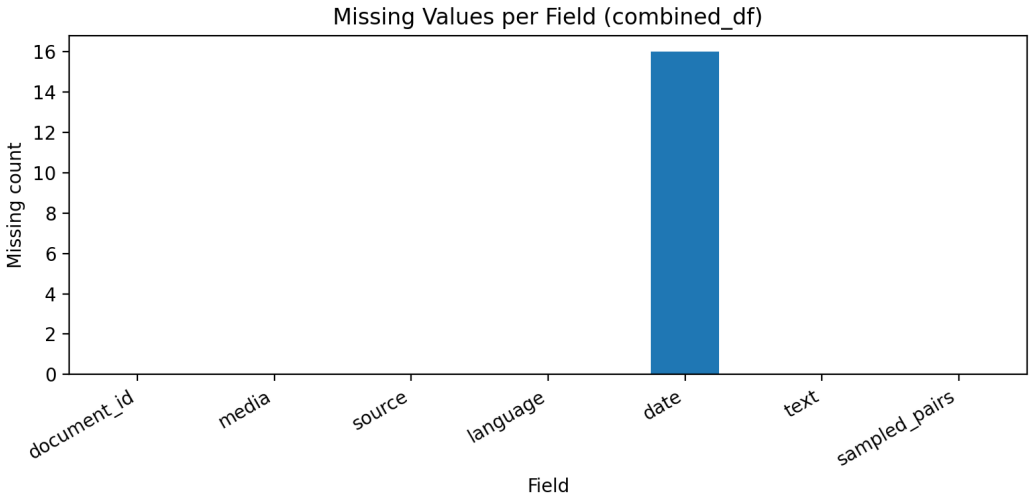

Figure 7: Missing-value overview for key document-level fields. The only missingness reported here is in `date` (16 entries).

6.2.2 Flattened person–location pairs

We flatten `sampled_pairs` across documents to study mention frequency and knowledge-base coverage:

- **Pairs extracted:** 8,251 person–location pairs.
- **Language distribution (pairs):** `fr`=5,948, `de`=1,656, `en`=647.

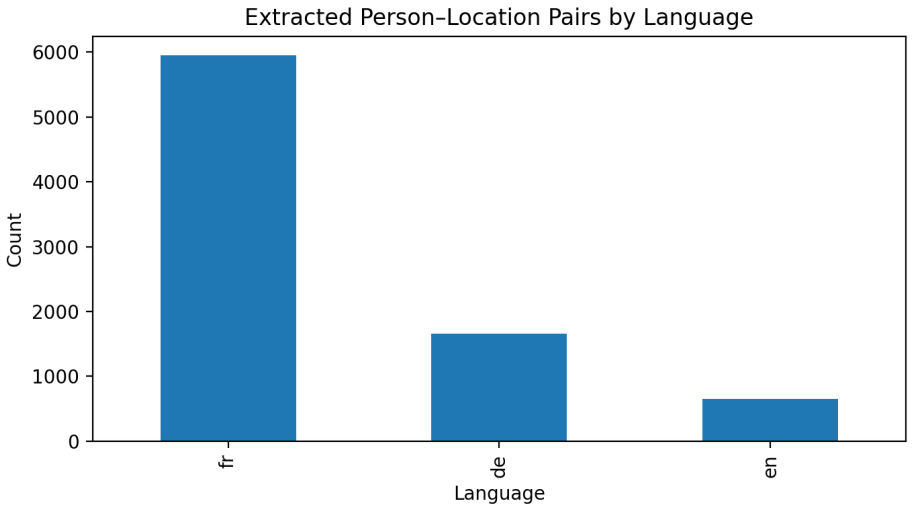

Figure 8: Distribution of flattened person–location pairs by language (8,251 pairs).

6.2.3 Most frequent mentions and co-occurrences

- **Unique mentions:** 4,161 unique person mentions; 2,835 unique location mentions.
- **Top persons (examples):** Cochrane (25), Wellington (22), Bismarck (18), Nicolas (18), M. Lloyd George (17).

- **Top locations (examples):** Paris (289), France (218), Lausanne (133), Suisse (125), Genève (124), Londres (112), Berlin (109).

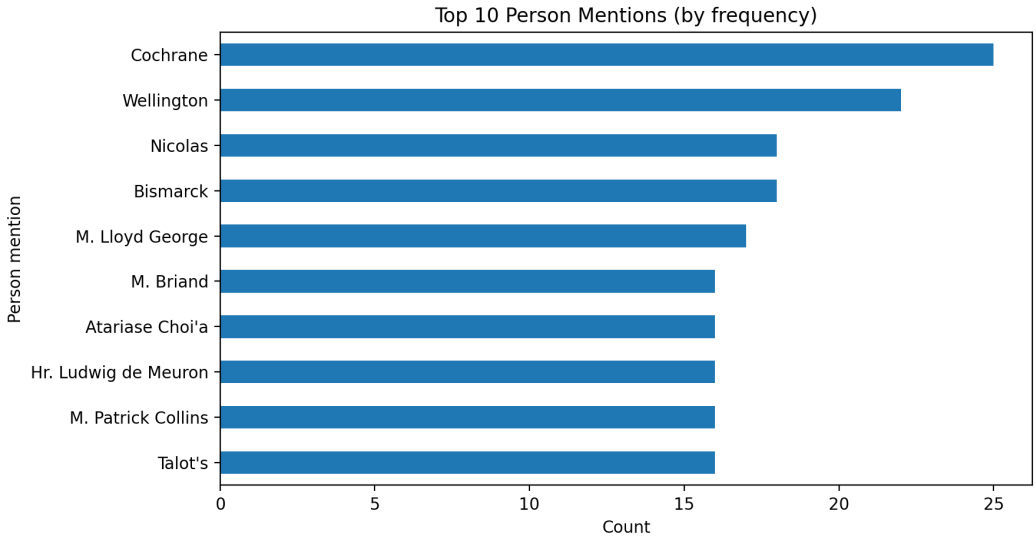

Figure 9: Top-10 most frequent person mentions in flattened pairs.

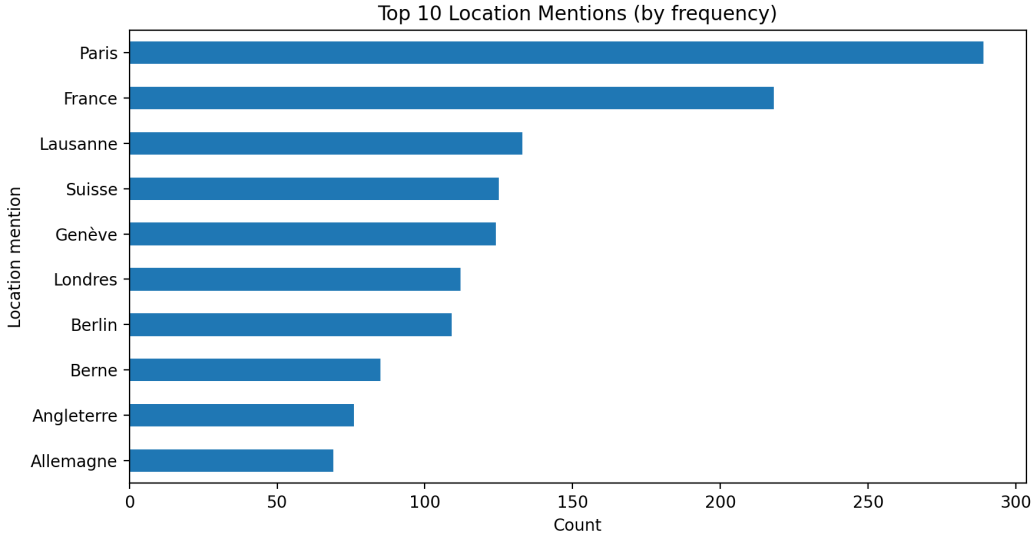

Figure 10: Top-10 most frequent location mentions in flattened pairs.

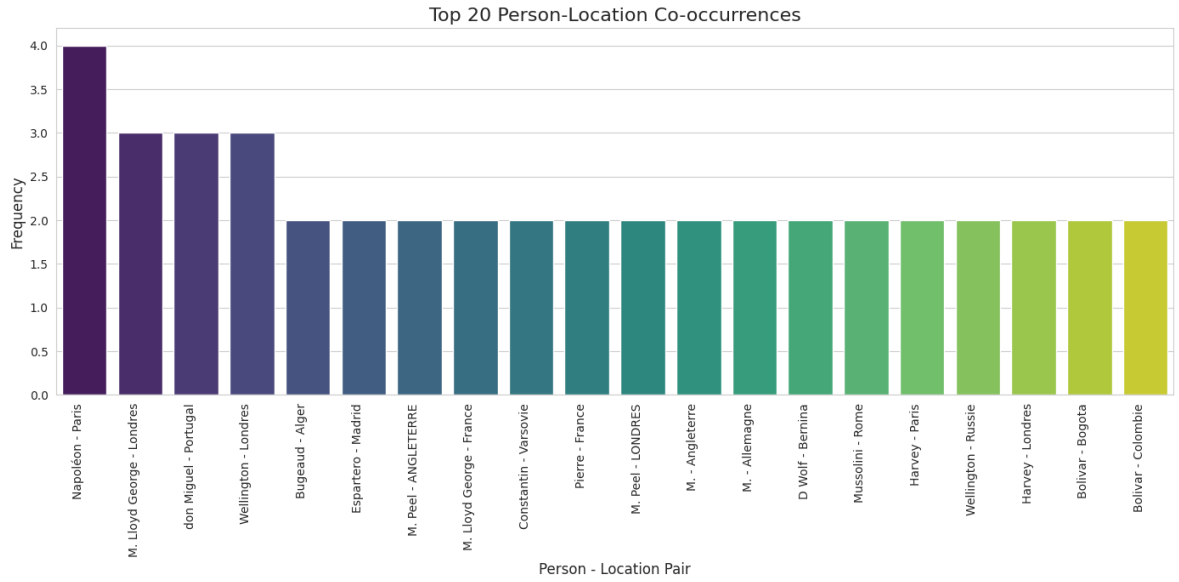

Figure 11: Top-20 person–location co-occurrences (frequency) from flattened pairs. Co-occurrence alone is not sufficient to infer `at`/`isAt` labels.

6.3 Experimental settings

6.3.1 Approach A (prompt-style XLM-R)

Encoder: `xlm-roberta-base`; max length 182; batch size 32; LR 2×10^{-5}; warmup 0.06; dropout 0.2; window 900 chars; oversampling + augmentation; early stopping (patience 2).

6.3.2 Approach B (MIL)

Encoder: `xlm-roberta-base`; bag size K (e.g., 4) windows per pair; per-instance max length 182; MIL pooling (attention; max pooling as ablation); weighted CE per head; early stopping on average macro recall.

6.4 Evaluation metrics

We report Macro Recall(`at`), Macro Recall(`isAt`), and their mean, consistent with HIPE-style balanced accuracy emphasis.

6.5 Results

Table 2: Development-set results on sandbox (English dev; as logged in our runs).

Model	Macro Recall(`at`)	Macro Recall(`isAt`)	Avg. Macro Recall
Approach A: Prompt dual-head XLM-R	0.4581	0.6552	**0.5566**
Approach B: Second model baseline	0.4848	0.4688	0.4768

6.6 Discussion

Approach A performs better overall due to stronger `isAt` macro recall on this sandbox dev set. The MIL formulation (Section 5.3) targets a known failure mode of single-window models—missing dispersed evidence—and is expected to be most beneficial when relation cues are localized in different parts of the document. Key ablations to validate MIL include: (i) max vs.

attention pooling, (ii) bag size K, and (iii) heuristics for selecting anchors (distance-based vs. mention-based).

7 Conclusion

We implemented two transformer baselines for HIPE-2026 person–place relation extraction: a prompt-style dual-head classifier and an MIL dual-head classifier that aggregates multiple candidate evidence windows. On our sandbox dev runs, the second baseline achieved an average macro recall of 0.4768, while the prompt-based model achieved 0.5566. Next steps include ablations over MIL bag size/pooling and efficiency optimizations aligned with the HIPE-2026 efficiency profile.

Acknowledgements

This project builds on HIPE resources and the HIPE-2026 shared task infrastructure, and uses multilingual historical datasets derived from the HIPE-2022 ecosystem.

References

[1] HIPE-2026 Organizers. *HIPE 2026 – Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts*. Project website, 2025–2026. `https://hipe-eval.github.io/HIPE-2026/`

[2] HIPE-2026 Organizers. *HIPE-2026-data* (GitHub repository), 2025–2026. `https://github.com/hipe-eval/HIPE-2026-data`

[3] M. Ehrmann, M. Romanello, A. Flückiger, and S. Clematide. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *Working Notes of CLEF 2020*, CEUR-WS Vol. 2696, 2020.

[4] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, and S. Clematide. Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Working Notes of CLEF 2022*, CEUR-WS Vol. 3180, 2022.

[5] M. Ehrmann, A. Hamdi, E. Linhares Pontes, M. Romanello, and A. Doucet. Named Entity Recognition and Classification on Historical Documents: A Survey. *arXiv:2109.11406*, 2021.

[6] M. Ehrmann, G. Colavizza, Y. Rochat, and F. Kaplan. Diachronic Evaluation of NER Systems on Old Newspapers. In *Proceedings of KONVENS 2016*, pp. 97–107, 2016.

[7] A. Hamdi et al. A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers. In *Proceedings of the 44th International ACM SIGIR Conference (SIGIR'21)*, 2021.

[8] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, and M. Sun. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *ACL*, 2019.

[9] L. Baldini Soares, N. FitzGerald, J. Ling, and T. Kwiatkowski. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL*, 2019.

[10] A. Plum et al. Biographical: A Semi-Supervised Relation Extraction Dataset. *arXiv:2205.00806*, 2022.