UMUTeam at HIPE-CLEF 2026: Person-Place Relation Extraction from Multilingual Historical Documents

Jorge Gómez-Navalón[1], Tomás Bernal-Beltrán[1], Ronghao Pan[1], José Antonio García-Díaz[1,*] and Rafael Valencia-García[1]

[1]Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain

Abstract

This paper presents the participation of UMUTeam in the HIPE-2026 shared task on person-place relation extraction from multilingual historical documents. The objective is to determine whether a semantic relation holds between a person and a location mentioned in a document and to classify this relation according to its temporal scope. Two relation types are evaluated independently: *at*, which captures whether a person was present at a location at any time prior to the publication date, and *isAt*, which determines whether the person is located at that place within the immediate temporal context of the document. We explored two complementary approaches. The first approach uses zero-shot prompting with two instruction-following large language models, `Qwen/Qwen3.5-9B` and `google/gemma-4-E4B-it`. The second approach is a supervised cross-encoder based on `FacebookAI/xlm-roberta-base` fine-tuned on the official training data. Both approaches receive as input the extracted local context, publication metadata, and a candidate person-location pair, and predict the labels for both relation types. Experimental results show that the zero-shot `Qwen/Qwen3.5-9B` model achieved the best performance among our submitted runs, obtaining a global macro recall of 0.5856 on the official Test A set and slightly outperforming the official baseline. In contrast, the supervised `FacebookAI/xlm-roberta-base` cross-encoder obtained a lower score, suggesting that, in this setting, zero-shot instruction-following models were more effective than our supervised classifier. Error analysis indicates that the main challenges of the task are temporal reasoning, implicit inference, and the distinction between explicit and implicit, uncertain evidence in noisy historical texts. These findings suggest that zero-shot instruction-following LLMs can provide competitive results for historical relation extraction, while supervised multilingual transformers remain a relevant baseline whose performance may depend strongly on the amount and nature of the available training data.

Keywords

person-place relation extraction, multilingual historical documents, large language models, zero-shot prompting, cross-encoder, temporal relation classification

1. Introduction

Historical documents are an essential source for studying the evolution of societies, institutions, public figures, and geographical spaces over time. Newspapers, gazettes, and other archival collections contain rich information about where people lived, travelled, worked, met, or were reported to be at specific moments. When extracted at scale, these references can support the reconstruction of spatial and temporal trajectories, offering valuable evidence for digital humanities research, historical analysis, and large-scale cultural heritage studies [1, 2]. However, transforming historical texts into structured knowledge remains a challenging task, since unlike contemporary and well-edited documents, historical collections often contain OCR errors, spelling variation, archaic language, inconsistent punctuation, and fragmented textual contexts. These difficulties are further increased in multilingual settings, where models must deal with different linguistic structures, historical writing conventions, and named entity variations [3]. As a result, identifying entities in a document is only a first step; understanding the relations between them requires deeper interpretation of the surrounding context.

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

*Corresponding author.

jorge.gomezn@um.es (J. Gómez-Navalón); tomas.bernalb@um.es (T. Bernal-Beltrán); ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

Among the different types of information that can be extracted from historical documents, person-place relations are particularly relevant. Linking individuals to locations makes it possible to study mobility, presence, residence, political activity, travel, and other forms of geographical association. Nevertheless, such relations are rarely expressed in a uniform or explicit way. A person and a place may appear close to each other without being semantically related, while in other cases the relation may be implicit, distributed across several sentences, or dependent on temporal expressions. Therefore, person-place relation extraction cannot be reduced to simple entity co-occurrence, but requires contextual and temporal reasoning.

Recent advances in Natural Language Processing (NLP) have opened new possibilities for addressing this type of problem. On the one hand, multilingual transformer encoders have shown strong performance in cross-lingual and multilingual transfer settings, making them suitable candidates for historical collections involving several languages [4]. On the other hand, instruction-following large language models offer a flexible alternative, as they can be prompted with task definitions and applied to new tasks without task-specific training [5]. These two paradigms represent different ways of approaching historical relation extraction: supervised learning from examples and zero-shot inference based on natural-language instructions.

In this paper, we describe our participation in the HIPE-2026 shared task[1] on person-place relation extraction from multilingual historical documents [6, 7]. Our work explores both paradigms by comparing supervised multilingual classification with zero-shot prompting using instruction-following language models. Through this comparison, we aim to analyze the extent to which current models can capture person-place relations in noisy and temporally complex historical texts, as well as the trade-offs between task-specific training and prompt-based reasoning.

The main contributions of this work are as follows. First, we implement a supervised cross-encoder approach based on a multilingual transformer model for person-place relation classification. Second, we evaluate zero-shot prompting strategies using instruction-following large language models. Third, we compare these approaches under the official shared-task setting and analyze their behavior in terms of classification performance and error patterns. Finally, we discuss the main challenges observed when applying modern language technologies to historical, multilingual, and temporally ambiguous documents.

The remainder of this paper is organized as follows. Section 2 reviews previous work on historical document processing, relation extraction, multilingual transformers, and large language models. Section 3 presents the shared task and dataset. Section 4 describes the proposed approaches. Section 5 reports and discusses the experimental results. Finally, Section 6 summarizes the main findings and outlines future work.

2. Related Work

Historical NLP and HIPE. The automatic extraction of structured information from historical documents has become an increasingly relevant research area at the intersection of NLP and Digital Humanities. The large-scale digitisation of newspapers, books, archives, and cultural heritage collections has enabled new forms of computational access to historical sources. However, transforming digitised text into structured knowledge remains a complex task. Historical documents often contain OCR errors, obsolete spellings, diachronic linguistic variation, inconsistent typography, heterogeneous layouts, and domain-specific vocabulary, all of which make standard NLP pipelines less reliable than in contemporary text collections [8, 3]. These challenges are particularly relevant when the objective is not only to retrieve documents or identify isolated entities, but also to reconstruct semantic connections between people, places, events, and time periods.

Early work in historical NLP focused primarily on improving access to digitised collections through tokenisation, normalisation, named entity recognition, and information retrieval. Rule-based systems, gazetteers, and handcrafted linguistic resources were commonly used to identify names of people,

[1] https://hipe-eval.github.io/HIPE-2026/

organisations, and places in historical corpora [9]. Although these approaches provided valuable results in controlled settings, their portability was limited by the variability of historical texts across languages, periods, and document types. As historical corpora expanded, the need for more robust and transferable methods became increasingly evident. Recent surveys have shown that Named Entity Recognition (NER) in historical documents remains a challenging task due to the combined effect of textual noise, linguistic variation, annotation scarcity, and domain mismatch between modern training data and historical collections [3].

Shared tasks have played an important role in consolidating research on historical entity processing. The HIPE evaluation campaigns have provided multilingual benchmarks for named entity recognition and linking in historical newspapers, encouraging the development of systems capable of dealing with noisy OCR, multilinguality, and historical language variation. HIPE-2020 introduced named entity processing on historical newspapers in French, German, and English [10], while HIPE-2022 extended the evaluation setting to a broader multilingual and multi-domain scenario, including named entity recognition and linking under heterogeneous historical conditions [1]. These initiatives reflect a broader movement in Digital Humanities and NLP: moving from document retrieval and surface-level entity detection towards richer semantic modelling of historical sources.

Relation extraction and document-level evidence. A natural next step after entity recognition and linking is Relation Extraction (RE), whose goal is to identify semantic relations between entities mentioned in text. Classical RE approaches relied on supervised learning with handcrafted lexical, syntactic, and semantic features. Kernel-based methods, such as those proposed by Zelenko et al., exploited structured representations of sentences to classify semantic relations between entity pairs [11]. Later, distant supervision reduced the dependence on manually labelled relation instances by aligning text with external knowledge bases, enabling relation extraction at larger scale [12]. These approaches were highly influential, but they often assumed that relation evidence appeared in relatively local and well-formed textual contexts, an assumption that does not always hold in noisy or historically complex documents.

The introduction of neural models changed the methodological landscape of relation extraction. Convolutional and recurrent neural networks reduced the need for manual feature engineering by learning relation representations directly from text [13]. These models improved the ability to capture local semantic patterns around entity mentions, but they were still largely designed for sentence-level relation classification. In many real-world settings, and especially in historical texts, relations may not be explicitly expressed in a single sentence. Evidence can be indirect, distributed across the document, or dependent on external metadata. This limitation motivated a shift towards document-level relation extraction, where systems must aggregate information across multiple sentences and mentions.

Document-level relation extraction has become an important research direction because many relations require reasoning beyond local entity contexts. The DocRED benchmark showed that a substantial number of relations require synthesising evidence across several sentences, resolving coreference, and modelling long-distance dependencies [14]. Subsequent work has explored graph-based reasoning, latent structure induction, and document-level attention mechanisms to improve relation extraction in contexts where the evidence is distributed or implicit [15]. These advances are particularly relevant for historical documents, where entity mentions may be sparse, OCR noise may obscure relevant cues, and the relation between two entities may depend on broader discourse context rather than direct syntactic proximity.

Spatial and temporal interpretation. The extraction of relations involving places also connects with spatial NLP, geoparsing, and toponym resolution. Geoparsing research has traditionally focused on detecting place names in text and linking them to geographical coordinates or knowledge-base entries [16]. In historical corpora, however, this process is complicated by historical spelling variants, administrative changes, ambiguous place names, and the mismatch between contemporary gazetteers and past geographical references. Datasets and studies on toponym resolution in nineteenth-century

newspapers have shown that historical geographical interpretation requires combining textual context, metadata, and external resources to disambiguate places accurately [17]. Other work has argued for the use of multiple historical and contemporary resources to resolve places across past and present references [18]. Nevertheless, resolving a place name is only one layer of the problem: extracting meaningful person-place relations requires determining whether the text actually supports a semantic connection between the person and the location.

Temporal interpretation is another important dimension in historical information extraction. Foundational work on temporal interval reasoning provided formal models for representing temporal relations between events and intervals [19]. In NLP, TimeML introduced a framework for annotating events, temporal expressions, and temporal links in text [20], while TempEval operationalised temporal relation extraction as a shared evaluation problem [21]. These lines of work highlight the importance of anchoring textual information to temporal references and distinguishing between past, present, and future situations. For historical documents, temporal interpretation is particularly important because texts often refer simultaneously to current reports, past events, biographical facts, and indirect historical references. As a result, relation extraction over historical corpora often requires not only semantic classification, but also temporal anchoring and evidence qualification.

Multilingual transformers and instruction-following LLMs. The emergence of transformer-based language models has substantially improved performance across many NLP tasks, including relation extraction. BERT demonstrated that deep bidirectional contextual representations can be adapted to downstream tasks through fine-tuning [22]. In relation extraction, transformer encoders enabled more effective modelling of entity pairs and their surrounding context. Entity-aware formulations and cross-encoder architectures are especially suitable for pairwise classification, since the input context and the target entities are encoded jointly, allowing the model to capture fine-grained interactions between the entities and the textual evidence [23]. This type of architecture is well aligned with relation extraction settings in which candidate entity pairs are already provided and the system must decide whether a relation is supported by the document.

Multilingual transformer models are particularly relevant for historical corpora, which often include documents in several languages and time periods. Multilingual BERT showed that a single pretrained encoder could support transfer across languages, even without explicit cross-lingual supervision [24]. `FacebookAI/xlm-roberta-base` further improved multilingual representation learning by pretraining on a large-scale multilingual corpus and achieving strong results across several cross-lingual benchmarks [4]. These models offer an effective alternative to language-specific pipelines, especially in settings where annotated data is limited or unevenly distributed across languages. For historical information extraction, multilingual encoders provide a practical way to exploit shared patterns across languages while still allowing supervised adaptation to the target task.

In parallel with supervised transformer-based approaches, recent research has explored generative and instruction-based formulations of information extraction. Instead of modelling tasks such as NER or relation extraction as fixed classification problems, generative systems can be prompted to produce structured outputs from natural language instructions. Unified information extraction frameworks have shown that multiple IE tasks can be reformulated under a common text-to-structure paradigm [25]. Instruction-tuned approaches extend this idea by using natural-language task descriptions and output schemas to guide the model across different extraction problems [26]. Multilingual generative relation extraction resources have also shown that relation extraction can be formulated as multilingual triplet generation, supporting transfer across languages and relation types [27]. These developments are relevant because they reduce the gap between task definitions and model inputs, allowing systems to exploit natural-language descriptions of relations, labels, and decision criteria.

Large Language Models (LLMs) have further accelerated interest in zero-shot and few-shot information extraction. In zero-shot relation extraction, the model is not fine-tuned on the target dataset; instead, the task is described through a prompt that includes the input text, candidate entities, label definitions, and expected output format. Recent studies have revisited LLMs as zero-shot relation extrac-

tors, showing that prompt-based formulations can identify relations without task-specific parameter updates [28]. Other work has shown that aligning relation extraction with instruction-following and question-answering formats can improve zero-shot extraction capabilities [29]. This paradigm is especially attractive in historical NLP, where manually annotated data is expensive and domain adaptation is difficult. However, LLM-based extraction also introduces important challenges, including sensitivity to prompt design, inconsistent output formatting, hallucinated evidence, and difficulty enforcing strict label spaces.

Overall, these lines of work motivate the two paradigms explored in this paper: supervised multilingual cross-encoding and zero-shot instruction-following extraction for person-place relation classification in historical documents.

3. Task Description

This section describes the dataset and the two target relation types addressed in the HIPE-2026 shared task on person-place relation extraction from multilingual historical documents. The goal is to determine whether a semantic relation holds between a person and a location mentioned in a document, and to classify this relation according to its temporal scope.

3.1. Dataset Description

The HIPE-2026 v1.0 data release[2] is composed of multilingual historical documents annotated with person-place relations. The main benchmark consists of historical newspaper articles written in English, French, and German, covering approximately two centuries (19th and 20th centuries). In addition, the organizers provide a literary surprise test set in French to assess the robustness and generalization capabilities of participating systems on out-of-domain texts.

The HIPE-2026 dataset consists of two evaluation domains. Domain A contains historical newspaper articles in German, English, and French, spanning the nineteenth and twentieth centuries. This material is derived from the entity-annotated HIPE-2022 newspaper data and was further processed for the present shared task through document selection, filtering, candidate person-location pair extraction, and relation annotation. Domain B is a surprise test set composed of French literary and historiographical texts from the sixteenth to eighteenth centuries, designed to evaluate out-of-domain generalization. The official data is divided into training, development, and test splits, as summarized in Table 1.

Table 1
Statistics of the HIPE-2026 released data by split and language.

Split	Language	#Documents	#Pairs
Gold Train A	de	34	466
Gold Train A	en	35	307
Gold Train A	fr	35	478
Silver Train A	de	64	893
Silver Train A	en	23	206
Silver Train A	fr	293	4,110
Silver Dev A	de	32	432
Silver Dev A	en	17	151
Silver Dev A	fr	107	1,498
Gold Test A	de	19	238
Gold Test A	en	19	162
Gold Test A	fr	19	238
Gold Test B	fr	30	480

[2] https://github.com/hipe-eval/HIPE-2026-data/releases/tag/v1.0

Gold Test A corresponds to the historical newspaper evaluation set and includes German, English, and French documents. Gold Test B corresponds to the surprise French out-of-domain set and is evaluated separately under the Generalization Profile.

Each document is represented in JSON Lines format, where each line corresponds to one document following the official HIPE-2026 JSON schema[3]. Every entry contains four main components: metadata, which includes information such as the document identifier, publication title, language, source type, and publication date; the complete text of the historical document; a list of sampled candidate (person, location) pairs selected for annotation; and the gold labels for the `at` and `isAt` relations in the training and development sets.

For each sampled pair, the dataset provides unique entity identifiers, Wikidata QIDs for linked non-NIL mentions, and the list of textual mentions associated with both the person and the location. In the test sets, the relation labels are replaced with `null`, and systems must predict the corresponding values.

Listing 1: Simplified example of a HIPE-2026 document.

```
{
  "document_id": "NZZ-1798-12-12-a-p0003",
  "language": "de",
  "date": "1798-12-12",
  "text": "... Nelson ... Abukir ...",
  "sampled_pairs": [
    {
      "pers_mentions_list": ["Nelson"],
      "loc_mentions_list": ["Abukir"],
      "at": "TRUE",
      "isAt": "FALSE"
    }
  ]
}
```

Listing 1 shows a simplified example of a document in the HIPE-2026 dataset, illustrating the overall structure and the annotation of a person-location pair.

3.2. Target Relation Types

The HIPE-2026 shared task is formulated as a person-place relation classification problem. Given a document and a candidate person-location pair, systems must predict two target relation types: `at` and `isAt`. These relation types are conceptually connected but are evaluated independently in the official setting.

The task requires systems to predict two target relation types for each candidate person-location pair. The `at` relation captures whether the document provides evidence that the person was at the specified location at any point in time prior to the publication date. This relation is formulated as a three-class classification problem with the labels TRUE, PROBABLE, and FALSE. The label TRUE indicates that the document explicitly states that the person was at the location, PROBABLE denotes that the relation can be reasonably inferred from implicit contextual evidence, and FALSE indicates that there is no supporting evidence or that the document contradicts the relation. This relation type encompasses a broad temporal range and includes situations such as places of residence, travel destinations, birthplaces, workplaces, or previously visited locations.

The `isAt` relation is a temporally constrained refinement of `at`. It determines whether the person is located at the specified place within the document's temporal horizon, defined by the organizers as approximately one month before the publication date. This relation is modeled as a binary classification

[3] https://github.com/hipe-eval/HIPE-2026-data/blob/main/schemas/hipe-2026-data.schema.json

problem with the labels TRUE and FALSE. The label TRUE indicates that the text suggests the person is currently at the location or was there shortly before publication, whereas FALSE indicates that no such recent association can be established.

For example, if a newspaper article reports that a political leader is visiting Berlin, both `at` and `isAt` are labeled as TRUE. In contrast, if the article mentions that the leader visited Berlin several years earlier, only the `at` relation is considered valid. Although both relations are conceptually connected, the official evaluation treats them independently, requiring systems to generate two predictions for each person-location pair.

3.3. Evaluation Setting

The official evaluation considers three profiles: the Accuracy Profile, the Efficiency Profile, and the Generalization Profile. The Accuracy Profile evaluates predictive performance on the historical newspaper Test A set. Both `at` and `isAt` are evaluated independently using macro Recall, also referred to as balanced accuracy. The `at` relation is evaluated as a three-class classification problem with labels TRUE, PROBABLE, and FALSE, while `isAt` is evaluated as a binary classification problem with labels TRUE and FALSE. The official global score for Test A is computed as the arithmetic mean of the macro Recall obtained for both relation types. The official evaluation was performed using the HIPE-2026 evaluation script[4].

The Efficiency Profile complements accuracy with information about model size and parameter count. The Generalization Profile evaluates robustness on the surprise Test B set, which belongs to a non-news domain and is evaluated only for the `at` relation using macro Recall.

4. Methodology

To address the person-place relation extraction task proposed in HIPE-2026, we explored two complementary modeling strategies. The first strategy relies on instruction-following LLMs operating in a zero-shot setting, where the task is formulated as a structured prompting problem. The second strategy uses a supervised cross-encoder architecture based on a multilingual transformer fine-tuned on the official training data. These two approaches represent different paradigms for relation extraction: prompting-based inference without task-specific parameter updates, and data-driven learning through supervised optimization.

Both strategies receive as input an extracted local context window, the publication metadata, and a candidate pair consisting of one person entity and one location entity. The systems then generate predictions for the two target relation types following the official HIPE-2026 JSON schema.

4.1. Context Extraction and Input Representation

Historical documents in the HIPE-2026 collection can be relatively long and may contain several mentions of the same person or location entity. Instead of processing the full document for every candidate pair, we use a local context extraction strategy in order to reduce computational cost and limit the amount of irrelevant information provided to the models.

Given a document, a candidate person-location pair, and the mention lists provided for both entities, the system searches for the corresponding surface forms in the document text. The earliest located mention of either entity is used as an anchor, and a local window of 2,000 characters before and after this position is extracted, resulting in a maximum context length of approximately 4,000 characters. If no mention surface form is matched, the system falls back to the first 4,000 characters of the document. This fallback was included as a robustness mechanism for possible tokenization, normalization, or formatting mismatches between mention strings and document text.

This design keeps the input manageable for both the LLM-based and cross-encoder pipelines. However, it also introduces a limitation: in long documents, the earliest mention is not necessarily the most

[4] https://github.com/hipe-eval/HIPE-2026-eval

informative one, and relevant evidence may appear outside the extracted window. This may reduce recall for cases where the person-place relation is supported by distant or later mentions.

In addition to the extracted context, the input representation includes the document publication date, the document language, and the lists of textual mentions associated with the person and location entities. This information provides essential cues for temporal reasoning and multilingual interpretation, both of which are central to the HIPE-2026 task.

4.2. Overall Pipeline

Figure 1 summarizes the overall workflow of the proposed system. Starting from the input JSONL files, each sampled person-location pair is transformed into a structured representation that combines document metadata and contextual information. This representation is then processed either by a zero-shot LLM or by the supervised cross-encoder. Finally, the predicted labels are validated and inserted into the output files following the official HIPE-2026 JSON schema.

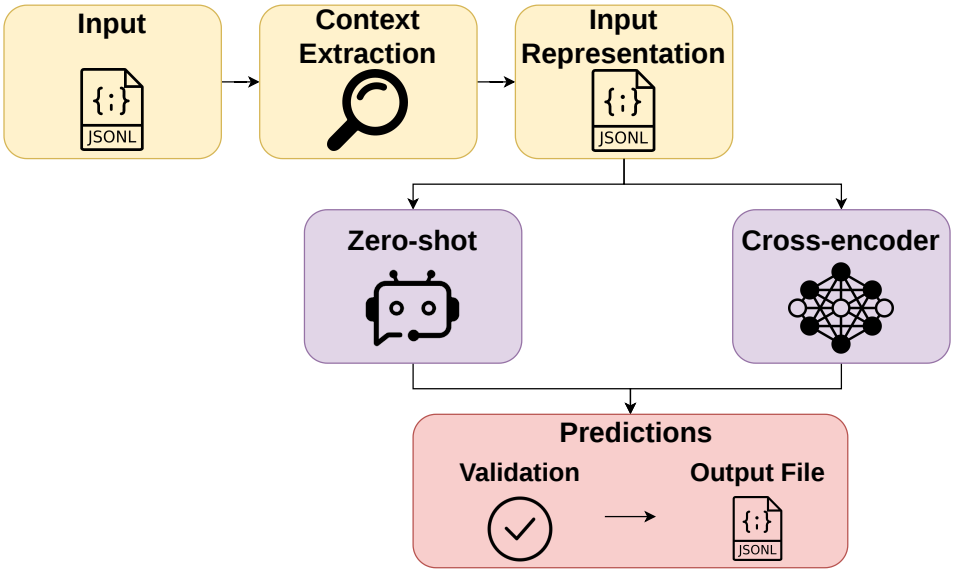

Figure 1: Overview of the proposed methodology for person-place relation extraction.

4.3. Zero-Shot Large Language Models

The first strategy treats relation extraction as an instruction-following task. We evaluated two open-source generative models, `Qwen/Qwen3.5-9B`[5] and `google/gemma-4-E4B-it`[6]. The same prompt template was used for all languages and was written in English. The prompt was designed to provide the model with the task definition, the label space, the temporal distinction between `at` and `isAt`, and the required JSON output format.

For each candidate pair, the system constructs a conversational prompt composed of two messages. The system message contains the general task instructions, including the definitions of TRUE, PROBABLE, and FALSE for `at`, the binary definition of `isAt`, and the consistency constraint that `isAt` cannot be TRUE when `at` is FALSE. The user message contains the document metadata, including the publication date and language, the candidate person and location entities, their mention lists, and the extracted local context window. The model is instructed to return a valid JSON object with the predicted labels and optional explanations.

[5] https://huggingface.co/Qwen/Qwen3.5-9B
[6] https://huggingface.co/google/gemma-4-E4B-it

Generation is performed greedily (`do_sample=False`) with a maximum of 256 output tokens. The generated text is parsed to extract the JSON object containing the predicted labels and optional explanations. If the output cannot be parsed correctly, the system defaults to assigning FALSE to both relations. After parsing, a validation step ensures that the predicted labels belong to the official label sets. Invalid labels are replaced by FALSE, and logical consistency is enforced so that whenever *at* is predicted as FALSE, the *isAt* label is also forced to FALSE.

This prompting-based approach does not require any task-specific training and can be directly applied to all languages provided in the benchmark. Furthermore, the optional explanations generated by the models offer an interpretable rationale for each decision, although these explanations are not used during evaluation.

4.4. Supervised Cross-Encoder

For the supervised approach, we fine-tuned a cross-encoder based on `FacebookAI/xlm-roberta-base`. The model was trained on the official HIPE-2026 v1.0 newspaper training files for the three languages: English, French, and German. Each labeled person-location pair was converted into a textual representation containing the publication date, document language, person and location mention lists, and the extracted local context window.

The input was tokenized with the corresponding `FacebookAI/xlm-roberta-base` tokenizer and truncated to a maximum length of 512 tokens. The contextualized representation of the first special token was used as the global representation of the input. Two independent classification heads were added on top of the encoder: one for the three-class `at` relation and another for the binary `isAt` relation.

The model was trained using the sum of two weighted cross-entropy losses. The class weights were manually set to $[1.5, 4.0, 1.0]$ for FALSE, PROBABLE, and TRUE in the `at` head, and to $[1.0, 2.0]$ for FALSE and TRUE in the `isAt` head. Training was performed with the Hugging Face `Trainer` API using AdamW, a learning rate of 2×10^{-5}, batch size 8, weight decay 0.01, and 10 epochs. Periodic evaluation was performed during training using the language-specific file passed through the `--dev-jsonl` argument, but no independent development split was used for model selection.

4.5. Implementation Details

All systems were implemented in Python using PyTorch and the Hugging Face `transformers` library. The zero-shot LLM runs were executed with `bfloat16` precision and automatic device mapping. For both LLMs, generation was performed greedily with `do_sample=False` and a maximum of 256 generated tokens. Model-specific chat templates were applied before generation.

The supervised cross-encoder was implemented as a PyTorch module on top of `FacebookAI/xlm-roberta-base`, with two task-specific classification heads. Input sequences were padded and truncated to 512 tokens. The submitted systems read and wrote files in the official HIPE-2026 JSONL format, preserving document and entity metadata and replacing the `at` and `isAt` fields with the predicted labels.

5. Results

5.1. Overall Results

Table 2 reports the official Accuracy Profile results obtained by our three submitted systems on the HIPE-2026 Gold Test A set. The table includes the macro Recall obtained for each target relation type, `at` and `isAt`, as well as the official global score computed as their arithmetic mean.

Among the evaluated systems, `Qwen/Qwen3.5-9B` achieved the best overall performance, obtaining a global score of 0.5856 and ranking 16th out of 46 submitted runs. This result slightly outperformed the official baseline provided by the organizers, which obtained a score of 0.5818. In contrast, the supervised `FacebookAI/xlm-roberta-base` cross-encoder achieved a substantially lower score of 0.4408, while `google/gemma-4-E4B-it` reached 0.4495.

Table 2
Performance of the submitted systems on the HIPE-2026 Gold Test A set.

Run	Approach	at	isAt	Global Score
Run 1	`FacebookAI/xlm-roberta-base` Cross-Encoder	0.3657	0.5159	0.4408
Run 2	`Qwen/Qwen3.5-9B` zero-shot	0.4702	0.7010	0.5856
Run 3	`google/gemma-4-E4B-it` zero-shot	0.3449	0.5542	0.4495

The lower performance of the cross-encoder may be partly related to the submitted training configuration, which did not use an independent development split for model selection, and to the difficulty of learning temporal and evidential distinctions from local context windows.

A detailed analysis of `Qwen/Qwen3.5-9B` across languages revealed relatively stable performance across all three languages, with global scores of 0.5384 in English, 0.5880 in French, and 0.6305 in German. On Gold Test B, which evaluates only the `at` relation under the Generalization Profile, the model achieved a macro Recall of 0.5723 on the surprise French literary test set. This result suggests some ability to transfer to an out-of-domain setting, although it should be interpreted as moderate rather than strong generalization.

These results indicate that zero-shot instruction-following LLMs can be highly competitive for multilingual person-place relation extraction, even without task-specific fine-tuning. In particular, `Qwen/Qwen3.5-9B` showed a useful ability to follow the task instructions and perform contextual reasoning over historical texts. On the other hand, the cross-encoder struggled to generalize effectively from the limited amount of annotated training data, suggesting that supervised learning alone may be insufficient when the task requires substantial temporal and inferential reasoning.

The comparison between the two LLMs also highlights that model architecture and instruction-following capabilities play a crucial role. Although both systems were evaluated using the same prompting strategy, `Qwen/Qwen3.5-9B` outperformed `google/gemma-4-E4B-it`, showing that not all instruction-tuned models exhibit the same degree of robustness for complex relation extraction tasks.

5.2. Performance by Relation Type

A more detailed examination reveals substantial differences between the two target relation types. In general, the *at* relation proved to be significantly more challenging than *isAt*. While *at* requires a three-class classification among TRUE, PROBABLE, and FALSE, the *isAt* relation is formulated as a simpler binary classification task focused on the immediate temporal context of the document.

Across the evaluated systems, performance on `isAt` was consistently higher than on `at`. For the best-performing model, `Qwen/Qwen3.5-9B`, macro Recall increased from 0.4702 on `at` to 0.7010 on `isAt`. Similar patterns were observed for the cross-encoder and `google/gemma-4-E4B-it`.

This difference suggests that one of the main challenges lies in distinguishing explicit evidence from implicit, uncertain evidence. In particular, the PROBABLE class introduces a substantial degree of ambiguity, as it requires systems to determine whether the relation is only weakly implied by contextual clues rather than directly stated in the text. By contrast, the binary nature of *isAt* makes the decision process more straightforward and less sensitive to uncertainty.

Although we do not report a separate binary-evaluation score here, this pattern suggests that many errors are related to evidential calibration between TRUE and PROBABLE, rather than to a complete failure to detect person-place associations. In other words, the model often identifies contextual evidence linking a person and a place, but struggles to determine whether this evidence is explicit enough for TRUE or only indirect enough for PROBABLE.

5.3. Error Analysis

The quantitative results reveal several consistent error patterns across the evaluated systems. The most frequent source of errors occurs in the *at* relation, particularly in the distinction between the TRUE and

PROBABLE classes. This behavior is expected, since the boundary between explicit and inferred evidence is often subtle and can depend on subjective interpretation.

Another common source of mistakes involves temporal reasoning. Systems may correctly identify that a person was associated with a location at some point in the past, but fail to determine whether this association falls within the temporal horizon required by the *isAt* relation. As a result, some instances are assigned TRUE for *at* but incorrectly labeled as FALSE for *isAt*.

The supervised cross-encoder exhibited a tendency toward conservative predictions, which reduced its ability to capture nuanced contextual clues. In contrast, the LLM-based systems were better able to leverage broader linguistic and world knowledge, although they still struggled with subtle distinctions involving uncertainty and temporal scope.

Overall, the error analysis confirms that the primary challenge of HIPE-2026 lies not in detecting simple co-occurrence between entity mentions, but in performing higher-level reasoning over historical texts that may contain noisy OCR, indirect references, and temporally distant events.

Table 3 presents selected error cases produced by the best-performing model, illustrating different types of incorrect predictions.

Table 3
Selected error cases produced by `Qwen/Qwen3.5-9B`.

Document ID	Person-Location Pair	Gold *at*	Predicted *at*	Gold *isAt*	Predicted *isAt*	Explanation
sn88068010-1890-10-02-a-i0001	Pallas-New York	FALSE	TRUE	FALSE	TRUE	The model interpreted the mythological figure "Pallas Athene" as a real person and treated symbolic references to travel as evidence of an actual visit to New York.
luxwort-1948-08-27-a-i0008	Brigadier M. R. L. Robinson-amerikanischen Zone	PROBABLE	TRUE	TRUE	TRUE	The model inferred physical presence from Robinson's involvement in discussions about the American Zone, although the text only states that he spoke about economic plans affecting that region.
EXP-1998-07-25-a-i0143	Joseph Cyrille N'Do-Lugano	FALSE	PROBABLE	FALSE	FALSE	The model inferred a weak relation from N'Do's participation in a previous match against Lugano, although the text does not place him physically in Lugano.

The examples in Table 3 illustrate several recurrent error patterns. In the first case, the model treats a figurative or mythological mention as evidence of physical presence, leading to a false positive prediction for both relation types. In the second example, the system overestimates the strength of the evidence and predicts TRUE instead of PROBABLE. In the third case, the model infers a weak association from contextual information despite the absence of evidence that the person was physically located in the place. Together, these examples show that the main challenges involve mention interpretation in noisy or figurative contexts, evidential calibration, and contextual inference.

These examples confirm one of the central challenges of the HIPE-2026 shared task: systems must not only identify semantic associations between persons and places, but also assess both the certainty and the temporal relevance of the evidence provided by multilingual historical documents.

6. Conclusion

In this paper, we presented our participation in the HIPE-2026 shared task on person-place relation extraction from multilingual historical documents. We explored two complementary approaches: a supervised cross-encoder based on `FacebookAI/xlm-roberta-base` and two zero-shot instruction-based large language models, `Qwen/Qwen3.5-9B` and `google/gemma-4-E4B-it`. These approaches represent two distinct paradigms for relation extraction: task-specific supervised learning and prompting-based reasoning without fine-tuning.

The experimental results show that `Qwen/Qwen3.5-9B` achieved the best overall performance among our submitted systems, obtaining a global macro recall of 0.5856 on the official Test A set. This result indicates that zero-shot instruction-following LLMs can be competitive for multilingual person-place relation extraction when the task is formulated with explicit label definitions and a structured output format. In contrast, the supervised `FacebookAI/xlm-roberta-base` cross-encoder obtained a lower score in the official evaluation, suggesting that our fine-tuned classifier struggled to generalize from the available training data.

The error analysis revealed that the main challenges of the task are closely aligned with the difficulties highlighted in the HIPE-2026 guidelines, including temporal reasoning, implicit inference, and the distinction between explicit and implicit, uncertain evidence in noisy historical texts. In particular, the models frequently over-predicted positive relations when a person and a place appeared in the same document but were not semantically linked, and they struggled to distinguish between current and historical locations when assigning the *isAt* relation.

Overall, our results confirm that person-place relation extraction from historical documents remains a challenging task that requires both linguistic understanding and temporal interpretation. The multilingual and diachronic nature of the dataset, together with OCR noise and indirect contextual cues, make this benchmark particularly demanding.

Future work will focus on improving the robustness of person-place relation extraction in three directions. First, we plan to explore instruction tuning and few-shot prompting strategies to better adapt LLMs to the distinction between TRUE, PROBABLE, and FALSE. Second, we will investigate hybrid architectures that combine the contextual reasoning capabilities of LLMs with the efficiency and stability of supervised classifiers. Third, we will study improved context selection strategies that consider multiple entity mentions rather than anchoring the input window on the earliest occurrence, which may help recover evidence that appears later in long documents.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammar and spelling checking.

Acknowledgments

This work is part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF/EU - FEDER/UE)-a way of making Europe. Mr. Tomás Bernal-Beltrán is supported by University of Murcia through the predoctoral programme.

References

[1] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Overview of hipe-2022: named entity recognition and linking in multilingual historical documents, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022, pp. 423–446.

[2] A. Hamdi, E. Linhares Pontes, E. Boros, T. T. H. Nguyen, G. Hackl, J. G. Moreno, A. Doucet, A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2328–2334.
[3] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification in historical documents: A survey, ACM Computing Surveys 56 (2023) 1–47.
[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.
[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[6] J. Opitz, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, M. Ehrmann, S. Clematide, Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS, 2026. doi:10.5281/zenodo.20344461.
[7] J. Opitz, C. Raclé, A. Michail, M. Romanello, M. Ehrmann, 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. Sánchez Salido, A. Barrón-Cedeño, A. García Seco 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.
[8] M. Piotrowski, Natural language processing for historical texts, Morgan & Claypool Publishers, 2012.
[9] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, J. Ball, Use of the edinburgh geoparser for georeferencing digitized historical collections, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368 (2010) 3875–3889.
[10] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended overview of clef hipe 2020: named entity processing on historical newspapers, in: CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum, volume 2696, CEUR-WS, 2020.
[11] D. Zelenko, C. Aone, A. Richardella, Kernel methods for relation extraction, Journal of machine learning research 3 (2003) 1083–1106.
[12] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009, pp. 1003–1011.
[13] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers, 2014, pp. 2335–2344.
[14] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, Docred: A large-scale document-level relation extraction dataset, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 764–777.
[15] G. Nan, Z. Guo, I. Sekulić, W. Lu, Reasoning with latent structure refinement for document-level relation extraction, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 1546–1557.
[16] J. L. Leidner, M. D. Lieberman, Detecting geographical references in the form of place names and associated spatial natural language, Sigspatial Special 3 (2011) 5–11.
[17] M. C. Ardanuy, D. Beavan, K. Beelen, K. Hosseini, J. Lawrence, K. McDonough, F. Nanni, D. van Strien, D. C. Wilson, A dataset for toponym resolution in nineteenth-century english newspapers, Journal of Open Humanities Data 8 (2022).
[18] M. C. Ardanuy, K. McDonough, A. Krause, D. C. Wilson, K. Hosseini, D. Van Strien, Resolving

places, past and present: toponym resolution in historical british newspapers using multiple resources, in: Proceedings of the 13th Workshop on Geographic Information Retrieval, 2019, pp. 1–6.
[19] J. F. Allen, Maintaining knowledge about temporal intervals, Communications of the ACM 26 (1983) 832–843.
[20] J. Pustejovsky, J. M. Castano, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, D. R. Radev, Timeml: Robust specification of event and temporal expressions in text., New directions in question answering 3 (2003) 28–34.
[21] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, J. Moszkowicz, J. Pustejovsky, The tempeval challenge: identifying temporal relations in text, Language Resources and Evaluation 43 (2009) 161–179.
[22] J. Devlin, M.-W. Chang, K. Lee, K. 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), 2019, pp. 4171–4186.
[23] L. B. Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 2895–2905.
[24] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert?, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 4996–5001.
[25] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, H. Wu, Unified structure generation for universal information extraction, in: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), 2022, pp. 5755–5772.
[26] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, et al., Instructuie: Multi-task instruction tuning for unified information extraction, arXiv preprint arXiv:2304.08085 (2023).
[27] P.-L. H. Cabot, S. Tedeschi, A.-C. N. Ngomo, R. Navigli, Redfm: a filtered and multilingual relation extraction dataset, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4326–4343.
[28] G. Li, P. Wang, W. Ke, Revisiting large language models as zero-shot relation extractors, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 6877–6892.
[29] K. Zhang, B. J. Gutiérrez, Y. Su, Aligning instruction tasks unlocks large language models as zero-shot relation extractors, in: Findings of the association for computational linguistics: ACL 2023, 2023, pp. 794–812.

A. Zero-shot Prompt Template

The following template was used for the zero-shot LLM runs. Placeholders between angle brackets were filled automatically for each candidate person-location pair.

```
System:
You are an information extraction system for historical documents.
Your task is to decide whether a relation holds between a PERSON
and a LOCATION mentioned in the document.

Predict two relation types:

1. at:

* TRUE: the document explicitly states that the person was at the location
  at some point before or at the publication date.
* PROBABLE: the relation is not explicitly stated, but can be reasonably
```

```
  inferred from contextual evidence.
* FALSE: there is no evidence for the relation, or the document contradicts it.

2. isAt:

* TRUE: the document indicates that the person was at the location within
  the immediate temporal horizon of the document, approximately one month
  before the publication date.
* FALSE: this recent presence cannot be established.

Consistency rule:
If at is FALSE, then isAt must be FALSE.

Return only a valid JSON object with this structure:
{
"at": "TRUE|PROBABLE|FALSE",
"isAt": "TRUE|FALSE",
"at_explanation": "...",
"isAt_explanation": "..."
}

User:
Document date: <DATE>
Document language: <LANGUAGE>

Person entity:
<PERSON_ENTITY_ID>
Person mentions:
<PERSON_MENTIONS_LIST>

Location entity:
<LOCATION_ENTITY_ID>
Location mentions:
<LOCATION_MENTIONS_LIST>

Extracted context:
<CONTEXT_WINDOW>

Return the JSON prediction.
```