GippLab at HIPE 2026: Unlocking Person–Place Relation Extraction with the Qwen Model Family

Notebook for the HIPE Lab at CLEF 2026

Shakib Yazdani*, Jan Philip Wahle and Bela Gipp

University of Göttingen, Göttingen, Germany

Abstract

In this paper, we present our participation in the HIPE 2026 shared task on person–place relation extraction from multilingual historical texts. We propose a joint multilingual training setting across all three languages using parameter-efficient fine-tuning through LoRA adapters with mid-sized Qwen language model families. Our best-performing approach, based on *Qwen3.5-4B*, achieved first place in the *at* relation sub-task for French, obtaining a macro recall of 0.694 and outperforming the Ministral3-3B baseline score of 0.589. The same model also demonstrates competitive performance on English and German for the *at* relation sub-task. For the *isAt* relation sub-task, however, performance remains slightly below the baseline on German and French. Nevertheless, in the surprise test setting designed to evaluate out-of-domain generalization, our model achieves a macro recall of 0.664, indicating robust generalization capabilities.

Keywords

Information Extraction, Relation Extraction, Parameter-efficient Fine-tuning

1. Introduction

Relation Extraction (RE) is one of the critical stages of information extraction (IE) which aims to select the relation from a set of two entities in text. Traditional supervised RE methods [1, 2] and more recent large language model (LLM)-based RE methods [3, 4] have achieved significant advancements in recent years. However, these models are optimized for cleaned, English-centric data and are less explored in historical settings, where performance is often worse not least because of OCR-induced noise.

Most existing RE benchmarks are developed in controlled settings with predefined relation schemas and relatively clean textual inputs, such as TACRED for sentence-level RE [5] and DocRED for document-level RE [6]. Although domain-specific RE datasets, including the Biographical dataset [7] and SciNLP [8], demonstrate the potential of specialized RE systems, they do not focus on historical or multilingual scenarios. In contrast, only a limited number of benchmarks address multilingual historical relation extraction, such as [9]. Furthermore, a recent survey has emphasized the lack of resources targeting noisy multilingual environments [10]. These limitations show a need for benchmarks and evaluation settings specifically designed for RE in multilingual historical texts, where OCR noise and domain shift pose additional challenges.

To address these challenges, the CLEF HIPE 2026 shared task on *Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts* [11, 12] aims to advance automatic detection of spatio-temporal links between individuals and locations within a given textual context. The shared task comprises two sub-tasks: *at* and *isAt*. The *at* sub-task requires determining whether a text indicates that a person was located at a specific place at any point in time, given the document context. In contrast, the *isAt* sub-task focuses on whether the context explicitly supports that an individual was physically present at a location during the temporal scope expressed in the document. Both sub-tasks

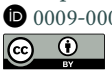
CLEF 2026 Working Notes, 21 - 24 September 2026, Jena, Germany

*Corresponding author.

shakib.yazdani@uni-goettingen.de (S. Yazdani); wahle@uni-goettingen.de (J. P. Wahle); gipp@uni-goettingen.de (B. Gipp)

https://shakibyzn.github.io/ (S. Yazdani); https://jpwahle.com/ (J. P. Wahle); https://bela.gipp.com/ (B. Gipp)

0009-0006-7220-9681 (S. Yazdani); 0000-0002-2116-9767 (J. P. Wahle); 0000-0001-6522-3019 (B. Gipp)



are defined across three languages: German, English, and French. Furthermore, the shared task features two evaluation settings: **Test A**, which consists of in-domain historical newspaper articles, and **Test B**, a surprise out-of-domain evaluation set designed to assess the generalization ability of systems on literary texts.

To address the shared task, we investigate parameter-efficient fine-tuning strategies. Specifically, we fine-tune LoRA adapters of the Qwen language model family [13, 14, 15] in a joint multilingual setting across German, English, and French. Based on the experiments on the development set and official shared task submissions, we found that *Qwen3.5-4B* achieves the strongest overall performance. Our study includes six mid-sized Qwen models, allowing for a systematic evaluation of the effectiveness of the Qwen model family for person–place relation extraction in a potentially noisy historical context. [1]

2. Related Work

2.1. Relation Extraction

RE aims to identify relationships between entity pairs within a given context. Early work focused on specialized neural architectures, such as the position-aware attention mechanisms applied to LSTMs on the TACRED dataset [5]. Subsequent work used encoder-based pre-trained models. For example, Chen and Li [16] introduced a zero-shot framework for BERT utilizing multi-task learning objectives. Most recently, with the rise of LLMs, research has shifted toward LLMs, using both fine-tuning strategies [17] and in-context learning (ICL) paradigms, ranging from zero-shot [18] to few-shot settings [19, 20]. Wadhwa et al. [21] evaluated the few-shot capability of GPT-3 [22] across four RE benchmarks and compared it to supervised fine-tuning of Flan-T5 [23], finding that GPT-3 performs comparably to fine-tuned Flan-T5 models. Wan et al. [24] extended this line of work, demonstrating that carefully crafted in-context demonstrations for few-shot RE can yield improvements over fully-supervised counterparts.

2.2. Relation Extraction for Multilingual Historical Domain

Recent advancements in RE have focused on historical and multilingual contexts to address the specific challenges of domain-specific data, such as resource scarcity and natural language variability [10]. Motivated by the fact that traditional supervised learning methods are often hindered by the high cost of manual annotation, Plum et al. [7] introduced a semi-supervised approach for biographical information, leveraging structured data from sources like Wikidata to automatically compile datasets for digital humanities research. Building on the need for diverse historical resources, Yang et al. [9] presented HistRED, a bilingual dataset covering both Korean and Classical Chinese (Hanja), which enables document-level RE across varied temporal contexts. Ali et al. [10] provides a comprehensive survey of the current landscape of multilingual RE, emphasizing the ongoing difficulties posed by cultural variations, language's morphological complexity, and low-resource languages. To bridge the performance gap in classical Chinese historical documents, recent work by Tang et al. [25] introduces SLCoLM, a collaborative framework that integrates small pre-trained language models (SLMs) with LLMs. By leveraging the fact that SLMs can be efficiently adapted to downstream tasks using supervised data, alongside the few-shot reasoning capabilities of LLMs, SLCoLM shows benefits in low-resource language domains. In this paper, we investigate whether parameter-efficient fine-tuning of modern open-weight LLMs can provide robust performance across multiple historical languages within a unified training setup. In contrast to previous studies that primarily focus on supervised encoder-based architectures, semi-supervised pipelines, or hybrid SLM–LLM frameworks, we evaluate multilingual LoRA fine-tuning of the Qwen model family for person–place relation extraction in low-resource historical documents.

[1] Our code is available at: https://github.com/shakibyzn/person-place-hipe

3. Methodology

3.1. Dataset

The HIPE 2026 dataset for person–place relation extraction spans three languages (German, English, and French) and supports two complementary sub-tasks. The first, *at* sub-task, requires systems to determine whether a given text explicitly indicates that a person was at a specific place at some point in time. The second, *isAt* sub-task, assesses whether a person has ever been associated with a given location within the document's temporal scope. The dataset is designed to evaluate a system's ability to infer semantic links between individuals and locations as expressed in historical documents. It includes two evaluation settings. *Test A* consists of historical newspaper articles in German, English, and French, while *Test B* introduces a surprise evaluation set composed of French literary texts, specifically constructed to assess out-of-domain generalization capabilities.

Table 1: Dataset statistics for HIPE 2026 person-place relation extraction.

Split	Lang.	#Docs	#Pairs
Train	German	122	1690
Train	English	91	803
Train	French	352	4928
Dev	German	32	432
Dev	English	17	151
Dev	French	107	1498
Test	German	19	238
Test	English	19	162
Test	French	19	238
Test (surprise)	French	30	480
Total	-	758	10120

The historical newspaper portion of the dataset (Test A) is derived from the entity-annotated HIPE–2022 shared task [26], and has been newly annotated and formatted to suit LLM-based approaches better. Table 1 summarizes the distribution of documents and entity pairs (person–location) across languages and dataset splits. Among the three languages, French provides the largest training portion with 4,928 entity pairs, whereas English contains the smallest training set with 803 entity pairs.

3.2. Prompt Design and Data Formatting

In our supervised fine-tuning experiments, we follow a standard prompt template composed of three sections: `Instruction`, `Input`, and `Output`. The Instruction describes the person–place relation extraction task and constrains the model to return only a JSON object containing the keys *at* and *isAt*. The `Input` serializes each training instance by including the document identifier, publication date, document language, target person mention(s), target location mention(s), and the corresponding document text. During supervised fine-tuning, the `Output` section contains the gold JSON annotation for the target entity pair, enabling the model to learn the desired mapping from the serialized input to the structured prediction. During inference, the same prompt template is used, except that the `Output` section is left empty and the model is required to generate only the JSON prediction.

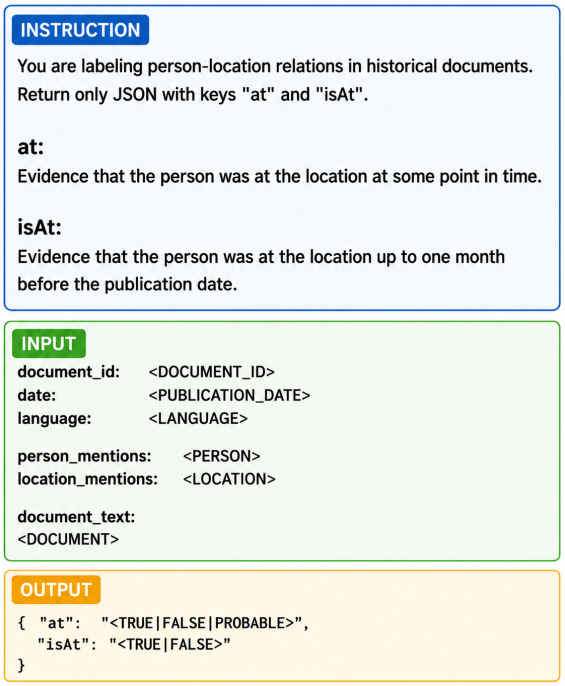


Figure 1: Prompt template used in our experiments.

3.2.1. Models

We utilize only models from the Qwen family [13, 14, 15] built by Alibaba Cloud. The Qwen

series offers a range of open-weight language
models with competitive performance across reasoning, coding, and multilingual benchmarks, making it a suitable choice for our experiments. Specifically, we selected *Qwen2.5-7B*, *Qwen3-8B*, *Qwen3.5-0.8B*, *Qwen3.5-2B*, *Qwen3.5-4B*, *Qwen3.5-9B* to compare their performance on the task. Table 2 provides details for the LLMs used in our experiments, including exact parameters and the Hugging Face model repository.

Table 2
Detailed information about models used in our experiments.

Model	Citation	Parameters	Model Repository
Qwen2.5-7B	[13]	7B	https://huggingface.co/Qwen/Qwen2.5-7B
Qwen3-8B	[14]	8B	https://huggingface.co/Qwen/Qwen3-8B
Qwen3.5-0.8B	[15]	0.8B	https://huggingface.co/Qwen/Qwen3.5-0.8B
Qwen3.5-2B	[15]	2B	https://huggingface.co/Qwen/Qwen3.5-2B
Qwen3.5-4B	[15]	4B	https://huggingface.co/Qwen/Qwen3.5-4B
Qwen3.5-9B	[15]	9B	https://huggingface.co/Qwen/Qwen3.5-9B

3.3. Evaluation Metrics

To evaluate the performance of our proposed approach, we employ *Macro Recall*, which assigns equal importance to all class labels and is therefore well suited for class-imbalanced settings [27]. *Macro Recall* for each label $\ell \in \mathscr{L}$ is defined as:

$$\text{Recall}(\ell) = \frac{TP_\ell}{TP_\ell + FN_\ell},$$

where TP_ℓ is the number of true positives and FN_ℓ denotes the number of false negatives for label ℓ. The macro Recall is then computed as the average over all labels:

$$\textit{Macro Recall} = \frac{1}{|\mathscr{L}|} \sum_{\ell \in \mathscr{L}} \text{Recall}(\ell).$$

3.4. Experimental Setup

We experiment with fine-tuning models using LoRA adapters [28] for person–place relation extraction. Figure 2 illustrates the overall architecture for our proposed method. We train models with all three languages concurrently to benefit from learning with more instance diversity, even if the source language is different.

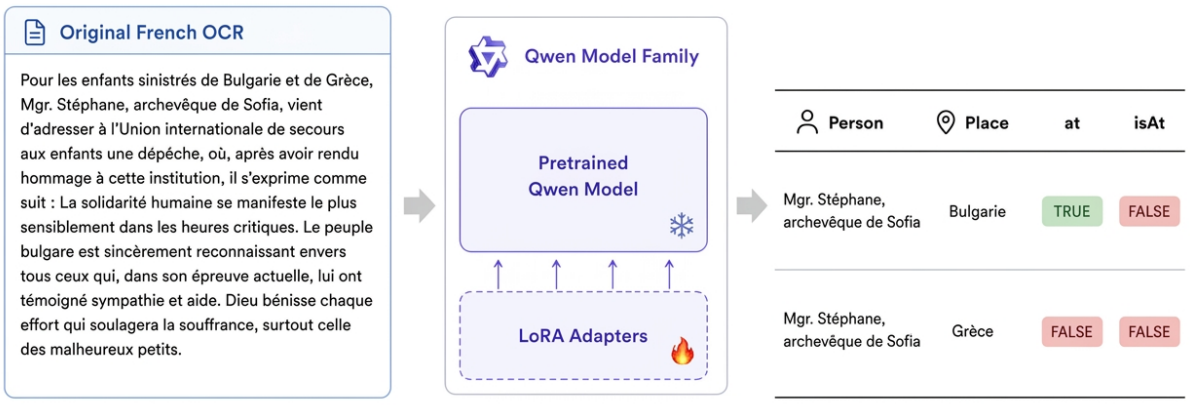


Figure 2: System overview for person–place relation extraction using LoRA fine-tuned Qwen language models.

Table 3
Macro Recall (%) for the *at* and *isAt* relations on the development set. Best results for each language are in **bold**.

	Qwen2.5-7B		Qwen3-8B		Qwen3.5-0.8B		Qwen3.5-2B		Qwen3.5-4B		Qwen3.5-9B	
Lang	at	isAt	at	isAt	at	isAt	at	isAt	at	isAt	at	isAt
DE	.687	.635	.717	.683	.414	.500	.610	.589	**.754**	.725	.734	**.758**
EN	.596	.764	.651	**.879**	.379	.500	.619	.710	.695	.731	**.725**	.846
FR	.648	.737	.714	.777	.440	.500	.519	.591	.706	.797	**.720**	**.803**

Table 4
Macro Recall (%) for the *at* and *isAt* relations on the Test sets A and B (surprise set). '–' indicates that the result is not reported because all *isAt* labels in the Surprise Test set are FALSE, making recall for this relation not applicable. Best results for each language are in **bold**.

		Qwen3.5-2B		Qwen3.5-4B		Qwen3.5-9B	
Split	Lang	at	isAt	at	isAt	at	isAt
Test A	DE	.509	.526	**.606**	.582	.587	**.604**
	EN	.433	.610	**.578**	.656	.554	**.674**
	FR	.405	.555	**.694**	.644	.588	**.675**
Test B (Surprise set)	FR	.538	–	**.664**	–	.608	–

3.4.1. Parameter-Efficient Fine-Tuning

We fine-tuned LoRA adapters for the models described in Section 3.2.1 using the Hugging Face Transformers library [2]. We explore six models on the development set, and based on the performance on the development set, we selected *Qwen3.5-2B*, *Qwen3.5-4B*, and *Qwen3.5-9B* as our shared task submission models (test sets). [3]

For all experiments, we employed a LoRA rank of 128, except for *Qwen3.5-4B*, where we reduced the rank to 64 due to comparatively lower performance on the English development set. We set the `lora_dropout` to 0.05 and used a LoRA scaling factor (`alpha`) equal to twice the LoRA rank. We fine-tuned LoRA adapters for all linear layers of the models. All models were trained for 2 epochs with a learning rate of 2×10^{-5}, which we found sufficient to avoid overfitting. Experiments were conducted on a single NVIDIA A100 GPU with 80 GB memory. Adapter fine-tuning was performed jointly across all three languages in the training set, rather than training separate adapters for each language.

4. Results

Tables 3 and 4 summarize the evaluation results of our systems on the development, test, and surprise test sets provided by the shared task organizers. On the development set, *Qwen3.5-9B* achieves the best overall performance across all three languages (German, English, and French). Interestingly, despite its smaller size, *Qwen3.5-4B* outperforms the larger *Qwen3-8B* and *Qwen2.5-7B* models, which belong to earlier Qwen generations, in most settings, except for the *isAt* relation in English, where *Qwen3-8B* achieves the highest macro recall of 0.879. [4]

On the official test A set, *Qwen3.5-4B* demonstrates more stable performance across the three languages compared to *Qwen3.5-9B*. In French, *Qwen3.5-4B* achieved a macro recall of 0.694, which ranked first among all participating teams in the *at* relation sub-task. In contrast, the model obtains lower scores on German and English, reaching 0.606 and 0.578 macro recall, respectively. This trend differs from the development set results, where performance was more balanced across languages. We hypothesize that training language-specific adapters could mitigate this discrepancy and lead to more stable language-specific performance. For the *isAt* relation sub-task, *Qwen3.5-9B* consistently outperforms *Qwen3.5-4B*,

[2] https://github.com/huggingface/transformers
[3] The final fine-tuned models are available at: https://huggingface.co/collections/Shakibyzn/clef-hipe2026
[4] Qwen3.5-4B, Qwen3.5-9B, and Qwen3.5-2B correspond to Runs 2, 1, and 3 on HIPE 2026 website, respectively.

with improvements ranging between 1–3 macro recall points across languages. In the official accuracy profile ranking provided by the organizers, our best run, Qwen3.5-4B, achieved a mean profile score of 0.6271, placing our team 12th out of 46 submitted runs. Additionally, we observe that macro recall on the *isAt* sub-task is sometimes lower than on the *at* sub-task. We hypothesize that explicitly encoding the dependency between the two relations in the prompt could improve performance on *isAt*. Specifically, as provided by the organizers, adding the constraints *"If `at` is FALSE, then `isAt` must also be FALSE"* and *"If `isAt` is TRUE, then `at` must also be TRUE"* may encourage more consistent predictions and improve recall on the *isAt* sub-task.

Finally, on the test set B (surprise set), which evaluates generalization to out-of-domain data, *Qwen3.5-4B* achieves a macro recall of 0.664, substantially outperforming the Ministral3-3B baseline [5] score of 0.5062 provided by the shared task organizers. Moreover, the relatively small gap between the out-of-domain score (0.664) and the in-domain French test score (0.694) suggests that the model generalizes robustly.

5. Conclusion

In this paper, we presented our approach to person–place relation extraction from multilingual historical texts in the context of the HIPE 2026 shared task. We used the Qwen language model family, applying parameter-efficient fine-tuning via LoRA adapters in a joint multilingual setting across German, English, and French. Our method demonstrated reliable performance across both the *at* and *isAt* sub-tasks, achieving first place in the *at* relation sub-task for French with a macro recall of 0.694 using the *Qwen3.5-4B* model. The same model also exhibited competitive and stable performance on English and German for the *at* relation sub-task. For the *isAt* relation sub-task, however, performance remained slightly below the baseline on German and French.

These findings are limited by evaluation on a single model family and a single multilingual LoRA fine-tuning strategy. Future work should extend this study by validating results across additional model families, exploring language-specific LoRA adapters, and investigating more recent LoRA variants to further improve robustness and generalization.

Acknowledgments

This work was supported by the Lower Saxony Ministry for Science and Culture (MWK) through the grant "KI in Museen" (11-76251-2050/2025, ZN4747).

Declaration on Generative AI

In the conduct of this research project, we used specific artificial intelligence tools and algorithms GPT-5.3 Chat and Claude Opus 4.7 to assist with Writing. While these tools have augmented our capabilities and contributed to our findings, it's pertinent to note that they have inherent limitations. We have made every effort to use AI in a transparent and responsible manner. Any conclusions drawn are a result of combined human and machine insights [29].

References

[1] S. Wu, Y. He, Enriching pre-trained language model with entity information for relation classification, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2361–2364. URL: https://doi.org/10.1145/3357384.3358119. doi:10.1145/3357384.3358119.

[5] https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512-GGUF

[2] H. Zheng, R. Wen, X. Chen, Y. Yang, Y. Zhang, Z. Zhang, N. Zhang, B. Qin, X. Ming, Y. Zheng, PRGC: Potential relation and global correspondence based joint relational triple extraction, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 6225–6235. URL: https://aclanthology.org/2021.acl-long.486/. doi:10.18653/v1/2021.acl-long.486.

[3] X. Ma, J. Li, M. Zhang, Chain of thought with explicit evidence reasoning for few-shot relation extraction, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 2334–2352. URL: https://aclanthology.org/2023.findings-emnlp.153/. doi:10.18653/v1/2023.findings-emnlp.153.

[4] Z. Li, F. Zhang, J. Cheng, AlignRE: An encoding and semantic alignment approach for zero-shot relation extraction, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2957–2966. URL: https://aclanthology.org/2024.findings-acl.174/. doi:10.18653/v1/2024.findings-acl.174.

[5] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware attention and supervised data improve slot filling, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 35–45. URL: https://aclanthology.org/D17-1004/. doi:10.18653/v1/D17-1004.

[6] 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: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074/. doi:10.18653/v1/P19-1074.

[7] A. Plum, T. Ranasinghe, S. Jones, C. Orasan, R. Mitkov, Biographical semi-supervised relation extraction dataset, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3121–3130. URL: https://doi.org/10.1145/3477495.3531742. doi:10.1145/3477495.3531742.

[8] D. Duan, J. Peng, Y. Zhang, C. Zhang, SciNLP: A domain-specific benchmark for full-text scientific entity and relation extraction in NLP, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 14473–14486. URL: https://aclanthology.org/2025.emnlp-main.732/. doi:10.18653/v1/2025.emnlp-main.732.

[9] S. Yang, M. Choi, Y. Cho, J. Choo, HistRED: A historical document-level relation extraction dataset, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3207–3224. URL: https://aclanthology.org/2023.acl-long.180/. doi:10.18653/v1/2023.acl-long.180.

[10] M. Ali, R. Speck, H. M. Zahera, M. Saleem, D. Moussallem, A.-C. Ngonga Ngomo, Multilingual relation extraction: A survey, IEEE Access 13 (2025) 151907–151933. doi:10.1109/ACCESS.2025.3604258.

[11] 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, volume 3180, CEUR-WS, 2026.

[12] 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.

[13] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, CoRR (2025). URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[14] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, CoRR (2025). URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[15] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

[16] C.-Y. Chen, C.-T. Li, ZS-BERT: Towards zero-shot relation extraction with attribute representation learning, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3470–3479. URL: https://aclanthology.org/2021.naacl-main.272/. doi:10.18653/v1/2021.naacl-main.272.

[17] Q. Sun, K. Huang, X. Yang, R. Tong, K. Zhang, S. Poria, Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction, in: Proceedings of the ACM Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 4407–4416. URL: https://doi.org/10.1145/3589334.3645678. doi:10.1145/3589334.3645678.

[18] K. Zhang, B. Jimenez Gutierrez, Y. Su, Aligning instruction tasks unlocks large language models as zero-shot relation extractors, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 794–812. URL: https://aclanthology.org/2023.findings-acl.50/. doi:10.18653/v1/2023.findings-acl.50.

[19] P. Zhang, W. Lu, Better few-shot relation extraction with label prompt dropout, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6996–7006. URL: https://aclanthology.org/2022.emnlp-main.471/. doi:10.18653/v1/2022.emnlp-main.471.

[20] N. Popovic, M. Färber, Few-shot document-level relation extraction, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5733–5746. URL: https://aclanthology.org/2022.naacl-main.421/. doi:10.18653/v1/2022.naacl-main.421.

[21] S. Wadhwa, S. Amir, B. Wallace, Revisiting relation extraction in the era of large language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15566–15589. URL: https://aclanthology.org/2023.acl-long.868/. doi:10.18653/v1/2023.acl-long.868.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot

learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[23] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, arXiv (2022). URL: https://arxiv.org/abs/2210.11416. doi:10.48550/ARXIV.2210.11416.

[24] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, GPT-RE: In-context learning for relation extraction using large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3534–3547. URL: https://aclanthology.org/2023.emnlp-main.214/. doi:10.18653/v1/2023.emnlp-main.214.

[25] X. Tang, L. Wang, J. Wang, Language model collaboration for relation extraction from classical chinese historical documents, Information Processing & Management 63 (2026) 104286. URL: https://www.sciencedirect.com/science/article/pii/S0306457325002274. doi:https://doi.org/10.1016/j.ipm.2025.104286.

[26] M. Ehrmann, M. Romanello, A. Doucet, S. Clematide, Introducing the hipe 2022 shared task: Named entity recognition and linking in multilingual historical documents, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2022, p. 347–354. URL: https://doi.org/10.1007/978-3-030-99739-7_44. doi:10.1007/978-3-030-99739-7_44.

[27] J. Opitz, A closer look at classification evaluation metrics and a critical reflection of common evaluation practice, Transactions of the Association for Computational Linguistics 12 (2024) 820–836. URL: https://aclanthology.org/2024.tacl-1.46/. doi:10.1162/tacl_a_00675.

[28] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[29] J. Wahle, T. Ruas, S. M. Mohammad, N. Meuschke, B. Gipp, Ai usage cards: Responsibly reporting ai-generated content, in: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 282–284. URL: https://doi.ieeecomputersociety.org/10.1109/JCDL57899.2023.00060. doi:10.1109/JCDL57899.2023.00060.