Awakened at HIPE-2026: Knowledge-Augmented Prompting and Logic-Constrained Fine-Tuning for Person–Place Relation Extraction

Notebook for the HIPE Lab at CLEF 2026

Dragoș-Mitruț Vasile[1], Elena-Simona Apostol[1] and Ciprian-Octavian Truică[1,2,*]

[1]National University of Science and Technology POLITEHNICA Bucharest, Splaiul Independenței 313, București 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

In this paper, we describe the contribution and participation of team Awakened in the HIPE-2026 task on person–place relation extraction from historical documents. For each (person, place) pair, the task asks whether the person is connected to the place (`at`: TRUE, PROBABLE, or FALSE) and whether that connection holds around the document's date (`isAt`: TRUE or FALSE). We submitted three runs, one for each of the Accuracy, Efficiency and Generalization profiles of the shared task. Our first run uses a prompted Claude Sonnet 4 with Wikidata facts attached to each pair, few-shot examples that calibrate the PROBABLE class, and also a knowledge-graph consistency rule. The second run is a system based on multilingual XLM-RoBERTa-large, fine-tuned on the Silver Train A dataset. This run uses the same Wikidata knowledge graph, text-pattern features, and an auxiliary logic-constrained loss that helps avoid inconsistent predictions. For this run, ablation shows that the logic constraints help as a light rule but hurt when weighted heavily. The third run reuses the prompted setup from the first run, but it uses a domain-agnostic prompt to match the literary domain. Our best French run ranks fourth on this language, and the domain-agnostic prompt unexpectedly matches or exceeds the newspaper-specific one on German and English.

Keywords

relation extraction, historical documents, large language models, knowledge graphs, logic-constrained learning

1. Introduction

Historical documents are valuable for information extraction, but they are also very hard to work with. The text often comes from OCR, so it contains errors, spelling and language have changed over time, and many of the people and places mentioned are no longer well-known today. Because of these challenges, named entity processing on historical material has become an important research direction, reflected in several shared tasks, from the first CLEF-HIPE evaluation on historical newspapers [1], to its multilingual follow-up [2], and later surveys of the field [3]. HIPE-2026[1] takes this domain one step further: instead of only identifying or linking entities, it asks a relational question, i.e., when a person and a place are mentioned in the same document, what kind of relation connects them? [4, 5]

Each (person, place) pair is labeled along two dimensions. The `at` label indicates whether there is any connection between the person and the place, having three different outcomes: a confirmed connection (TRUE), a plausible one (PROBABLE), and no connection at all (FALSE). The `isAt` label captures whether that connection is valid or not at the date of the document. This makes the task interesting because a person may be mentioned together with a place for different reasons, such as origin, office, candidacy, or just a passing reference, without necessarily being present there. Moreover, a past connection does not always imply a current one.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
*Corresponding author.
dragos.vasile2603@upb.ro (D. Vasile); elena.apostol@upb.ro (E. Apostol); ciprian.truica@upb.ro (C. Truică)
https://sites.google.com/view/elena-simona-apostol (E. Apostol); https://sites.google.com/view/ciprian-octavian-truica (C. Truică)
0009-0003-1213-3839 (D. Vasile); 0000-0001-6397-4951 (E. Apostol); 0000-0001-7292-4462 (C. Truică)


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

To address this, we rely on two main ideas. First, we use structured knowledge, since many decisions depend on facts that are not stated in the document, such as whether a person was alive at the document date or where they held office. For this purpose, we build a Wikidata-derived knowledge graph [6] for the entities in the task and use it in all our submitted systems. Second, we use logical consistency between labels. For example, if `at` is FALSE, then `isAt` must also be FALSE. These constraints can be encoded explicitly instead of leaving the model to learn them on its own.

We submit three runs that combine these ideas in different ways. Run 1 uses Claude Sonnet 4 with knowledge-graph facts and few-shot examples, following the now common use of large language models for guideline-driven tasks [7]. Run 2 uses a fine-tuned XLM-RoBERTa-large [8], combining knowledge-graph features, text patterns, and a logic-constrained loss inspired by semantic loss [9]. Run 3 also uses the prompting approach but makes use of a more domain-agnostic prompt for the literary surprise test set. Overall, our best results are obtained on French, where Run 1 ranks fourth among all submitted runs. Interestingly, the more general prompt in Run 3 slightly outperforms the newspaper-specific prompt of Run 1 on German and English. The smaller fine-tuned model offers a useful accuracy-size trade-off, while the logical constraints eliminate all label inconsistencies in the test predictions.

The rest of the manuscript is structured as follows. Section 2 presents the official HIPE-2026 task and the data we use for evaluation. We describe our implementation in Section 3: the knowledge graph derived from Wikidata and each of the three submitted runs. We present our results in Section 4. In Section 5, we interpret the results in detail. Finally, Section 6 concludes and discusses future work.

2. Task and Data

The HIPE-2026 shared task [4, 5] focuses on person–place relation extraction in historical documents. For each document, the organizers provide sampled (person, place) pairs, with both entities linked to Wikidata.

Evaluation is based on macro recall [10], computed separately for the three `at` classes and the two `isAt` classes. The final score for a file is the average of these two values. This makes the rare classes, especially PROBABLE and `isAt`=TRUE important.

We evaluate our system on the data released by the organizers[2], summarized in Table 1. Development and model training for the fine-tuned run are carried out on the Silver Train A and Silver Dev A splits. The official evaluation uses two test sets. Test A comes from the *impresso* collection [11] of historical newspapers in German, English, and French, mostly from the nineteenth and twentieth centuries. Test B is a French "surprise" set of literary and historical documents from the sixteenth to eighteenth centuries, designed to test generalization beyond newspapers; it is evaluated only on `at` label.

Table 1
Test data used in the evaluation. Pair counts are from the released test files; year ranges are the span of document dates. Two French Test A documents have no date.

Set	Lang	Docs	Pairs	Years
Impresso (Test A)	DE	19	238	1808–1948
Impresso (Test A)	EN	19	162	1800–1960
Impresso (Test A)	FR	19	238	1804–1998
Surprise (Test B)	FR	30	480	1542–1797

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

3. System Overview

3.1. Knowledge Graph Construction

The dataset comes with a Wikidata QID for the person and the place, and we use this information to build a knowledge graph by querying the public Wikidata SPARQL endpoint. For each person, we collect the date of birth (P569), the date of death (P570), the occupations (P106), the place of birth (P19), the place of death (P20), the residence (P551), the work location (P937), the position held (P39), and the country of citizenship (P27). As far as the places are concerned, we retrieve the type (P31) and the country (P17). All retrieved facts are saved in a file and reused at each run. The reasoner implements two temporal rules: 1) if a person is dead at the document date, then it cannot be present (prompt instruction in Run 1, loss-constraint in Run 2, and disabled in Run 3), and 2) a person who was born after the document date cannot be related to the place at that time (applied as post-processing).

Our cache contains 1121 persons, 1224 locations, and 2942 unique pairs. We measure how often these facts count, on the German development split: 27% of pairs (117 of 432). For the rest of 73%, no person–place relation is found, even though both QIDs are present, so the decision is made entirely by the model.

3.2. Run 1: Knowledge-Augmented Prompting

Our first run uses Claude Sonnet 4 (Anthropic) through the Anthropic API. We use only prompt engineering, and the entire system consists of 1) a prompt, 2) the Wikidata facts we attach to each pair, and 3) a single deterministic post-processing rule.

The model is instructed that its role is a historian analysing person–place relations in historical newspaper articles with possible OCR noise. We also give the definitions for each label and value (to distinguish `at`=TRUE from PROBABLE and FALSE, and `isAt`=TRUE from FALSE), and the prompt also states strict rules of the task: 1) if `at` is FALSE then `isAt` is FALSE and 2) a death date before the article means `isAt`=FALSE.

For each sample, we add a small description that is extracted from the Wikidata KG (described in Section 3.1) to give the model background facts that are not present in the document. The user prompt includes seven few-shot examples (Table 2). We choose to define our few-shot examples in this way because each entry targets errors we observed during the development of other prompt versions, where the PROBABLE class is the weakest. Another important mention is that chain-of-thought prompting does not improve the scores, so we ask the model only for a short justification, rather than step-by-step reasoning.

As far as the response is concerned, the model returns a JSON array, which we parse to extract the information we need. The parser falls back to FALSE for anything unexpected (wrong labels for `isAt` or `at`), enforces `isAt`=FALSE if `at`=FALSE, and if we encounter any error during parsing, we simply return FALSE/FALSE to ensure that every pair has a prediction.

Finally, we apply the post-processing rule: if the graph gives the birth date later than the article date, the person could not have been present, and both labels are set to FALSE.

3.3. Run 2: Logic-Constrained Fine-Tuning

For the second run, we choose to fine-tune a multilingual encoder, XLM-RoBERTa large [8], on the Silver Train A training data with all three languages combined, producing a single multilingual model rather than three monolingual ones to keep the system efficient.

The architecture steps are as follows: we take the `[CLS]` representation produced by the encoder and concatenate it with two groups of features, 16 features extracted from the Wikidata knowledge graph described in Section 3.1 and 6 text-pattern features extracted from the local context, which surrounds each pair. The resulting vector is passed through a dropout layer and then into two classification heads, one for `at` (three classes: TRUE, FALSE, and PROBABLE) and one for `isAt` (two classes: TRUE and FALSE).

Table 2
The seven few-shot examples in the Run 1 prompt, each chosen to address a specific error mode observed in development. "Target" is the label the example is meant to reinforce.

#	Target	Cue the example teaches
1	TRUE	An action happens at the place (presence, not just a role).
2	PROBABLE	Only an origin marker ("X from Munich"), no scene of presence.
3	PROBABLE	Candidacy in a district is not presence.
4	TRUE/FALSE	A departure: `at` is true, but `isAt` is false.
5	FALSE	Co-mention of person and place with no real link.
6	TRUE	A role *with* an action at the place (ambassador delivering a note in Rome).
7	PROBABLE	A link asserted only by Wikidata (nationality / birthplace).

On the one hand, the 16 knowledge-graph features include: 1) presence flags (whether the person and the place have an entry in the knowledge graph at all), 2) temporal flags (whether the person is alive at the document's date), and 3) relation flags (birthplace, place of death, residence, work location, position held, and a citizenship-country match). Furthermore, we introduce 6 binary indicators determined by applying regular expressions to the context window of the mentioned person or place entity ($\pm$200 characters), that match: 1) a locative action (e.g. *arrived*, *spoke*, *resided*), 2) a role or title marker (e.g. *ambassador*, *deputy*, *minister*), 3) an origin marker (e.g. *of*, *from*, *born in*), 4) a time indicator (e.g. *today*, *yesterday*, *currently*), 5) a departure marker (e.g. *left*, *departed*), and 6) a candidacy marker (e.g. *candidate in*, *election in*). These patterns are matched with Python's standard `re` module directly on the raw character context, without any prior tokenization.

The most important component is the logic-constraint losses. We add two penalty terms that encode constraints so that the model is discouraged from producing label combinations that cannot be correct. For the `at` head we use a focal cross-entropy loss [12] with $\gamma = 1.0$ to down-weight the dominant FALSE class. Equation (1) presents the formula for calculating the total loss, where $\lambda_{\text{hard}} = 0.05$ and $\lambda_{\text{soft}} = 0.02$. The hard term ($\mathcal{L}_{\text{hard}}$) penalizes three violations. The first one is the only strict rule of the task: if `at` is FALSE then `isAt` must be FALSE, penalized as $P(\texttt{at}{=}\textsc{false}) \cdot P(\texttt{isAt}{=}\textsc{true})$. The other two use knowledge-graph facts: if the person was already dead at the document's date, the model is penalized for predicting `isAt`=TRUE, and if the person was not yet born, it is penalized for predicting either `at`$\neq$FALSE or `isAt`=TRUE. The soft term ($\mathcal{L}_{\text{soft}}$) uses the six text-pattern features as weak hints rather than hard rules. For example, a locative action verb suggests actual presence, so it makes `at`=FALSE less likely. Regarding the title, role, or candidacy marker without an action verb often points to an affiliation rather than real presence, pushing the model towards PROBABLE over TRUE. Departure markers are treated as evidence against current presence, while time expressions combined with an action verb support `isAt`=TRUE.

$$\mathcal{L} = \mathcal{L}_{\text{focal}}^{\texttt{at}} + \mathcal{L}_{\text{CE}}^{\texttt{isAt}} + \lambda_{\text{hard}}\,\mathcal{L}_{\text{hard}} + \lambda_{\text{soft}}\,\mathcal{L}_{\text{soft}}, \tag{1}$$

We train the model for 20 epochs with an effective batch size of 32 (batch size 8 with gradient accumulation over 4 steps). The learning rate is set to 1.5×10^{-5} with a linear warmup over the first 8% of steps, and we use a weight decay of 0.01 and a dropout of 0.2.

The confusion matrices in Tables 3 and 4 show the main remaining errors. For `at`, the hardest distinction is between FALSE and PROBABLE, while TRUE is predicted better. For `isAt`, the model identifies most of the FALSE cases, but still misses many of the TRUE ones.

We ran ablation tests (Table 5) and found that both weights produce good improvements over the results, even if they are kept small. Training the model with no constraints reaches a global score of 0.672, and larger weights do not help and, in fact, hurt the model's performance.

Table 3
Development-set confusion matrix for `at` (rows: gold, columns: predicted).

gold \ pred	FALSE	PROBABLE	TRUE
FALSE	787	421	56
PROBABLE	93	366	109
TRUE	19	65	165

Table 4
Development-set confusion matrix for `isAt` (rows: gold, columns: predicted).

gold \ pred	FALSE	TRUE
FALSE	1782	125
TRUE	74	100

Table 5
Effect of the constraint weights on the development global score. All runs are identical except for λ_{hard} and λ_{soft} (same seed, 20 epochs, best dev checkpoint). The submitted configuration is highlighted.

λ_{hard}	λ_{soft}	`at`	`isAt`	**global**
0	0	0.6150	0.7295	0.6722
0.05	**0.02**	**0.6432**	**0.7546**	**0.6989**
0.3	0.1	0.6020	0.7117	0.6569
1.0	0.5	0.5902	0.7409	0.6656

We also run a feature ablation (Table 6) to separate the contribution of the two feature groups (KG and FOL). The knowledge-graph features have a greater impact: removing them lowers the score by 0.055, with the largest drop on `isAt` (0.7546 to 0.6903), which indicates that temporal alignment depends on the information present in the KG. The six text-pattern features contribute a smaller amount, a 0.017 drop when removed. It is notable that without the KG features, the text-pattern features slightly decrease the overall score, suggesting that this information becomes relevant when anchored by structured knowledge-graph facts.

Table 6
Feature ablation on Silver Dev A. All runs use the submitted configuration ($\lambda_{\text{hard}} = 0.05$, $\lambda_{\text{soft}} = 0.02$, 20 epochs, seed 42); each row removes one feature group. When the text-pattern features are removed, the soft loss term is also disabled, as it depends on them.

Configuration	`at`	`isAt`	**global**
Full (KG + text-pattern + logic)	**0.6432**	**0.7546**	**0.6989**
− text-pattern features	0.6344	0.7298	0.6821
− KG features	0.5968	0.6903	0.6436
− KG and text-pattern	0.5936	0.7193	0.6564

3.4. Run 3: Domain-Agnostic Prompting

The third run uses the same Claude Sonnet 4 setup as Run 1, including the Wikidata facts and the post-processing rule, and differs only in the prompt. We implement this run because Test B differs from Test A, being the literary "surprise" set, whose documents are extracted from historical and literary works rather than newspapers. The Run 1 prompt is written around newspapers, and we expect it to transfer poorly to that genre, so we rewrote it to be domain-agnostic. Table 7 summarizes the differences between the two prompts.

Table 7
Prompt differences between Run 1 (newspaper-specific) and Run 3 (domain-agnostic). Both share the same Claude Sonnet 4 backbone, the same Wikidata facts, and the same post-processing rule.

Aspect	Run 1	Run 3
Document framing	"historical newspaper article"	"historical document": newspaper, novel, memoir, official record, etc.
`isAt` definition	present = a short window around the publication date	present = the document's narrated moment, including the narrative present
PROBABLE definition	journalistic cues (role, origin, candidacy, Wikidata-only link)	same, plus a character named near a place in a novel without being present
Few-shot examples	7, all journalistic	11: seven journalistic and four narrated-scene examples
Death-date hard rule in prompt	stated (death before article date → `isAt` false)	removed

One of the main changes is to the label `isAt`. In newspapers, the relevant "present" is the publication date, so current presence can be tied to a short period of time around that date. In literary texts, this notion is less clear: the relevant present is often the moment being narrated, not the document date. Then, we treat `isAt` as true when the person is present at the place in the narrative present, even if the scene is written in the past tense. For the same reason, we remove from the prompt the rule that checks if the Wikidata death date is before the document date, so the model is no longer instructed to set `isAt`=FALSE for characters whose Wikidata death date precedes the document date.

4. Results

This section presents our official results for the HIPE-2026 lab, comparing the two prompted runs (Run 1 and Run 3) with the fine-tuned model (Run 2) across all languages in the test dataset. For reproducibility purposes, the prompts and the code are freely available on GitHub at https://github.com/DS4AI-UPB/Awakened-CLEF2026-HIPE and the fine-tuned Run 2 model is released on the Hugging Face Hub at https://huggingface.co/DS4AI-UPB/person-place-relation-xlmr.

Table 8 shows the official test scores for our three runs. Across the board, the two prompted runs (Run 1 and Run 3) are clearly stronger than the fine-tuned model (Run 2), which is expected given the difference in scale. French is our strongest language, German is close behind, and English is the weakest. On the two newspaper languages where the comparison is most interesting, the domain-agnostic prompt of Run 3 is slightly better than the newspaper-specific Run 1.

Our strongest official results are obtained on the French data. Out of 17 participating teams and 45 submitted runs, on Test A, our Run 1 is ranked fourth among all submitted runs for French, with a global score of 0.716, while Run 3 is ranked fifth with 0.685. On the German data, the two prompted runs are almost tied (0.675 for Run 3 and 0.674 for Run 1). On the English data, Run 3 performs better than Run 1, scoring 0.642 compared with 0.585. Overall, on Test A, Run 3 is our best submission, ranking eighth, followed by Run 1 in ninth place. On the literary Test B, Run 3 outperforms Run 1 again, with scores of 0.661 and 0.634.

Run 2 is submitted as the efficiency-oriented system. With about 561M parameters, it is not a lightweight model in absolute terms, but it still achieves competitive scores on French and German and ranks eighth in the Efficiency Profile ranking. Its weakest result is on the English data, where the limited development data likely makes fine-tuning not so stable.

Overall, Run 3 is slightly more robust than Run 1, even on some newspaper test sets, despite being designed as a more domain-agnostic prompt. We discuss this unexpected behavior in Section 5.

Table 8
Official HIPE-2026 test results for team Awakened. Test A reports the global score (mean of `at` and `isAt` macro recall); Test B reports `at` macro recall only.

Test set	Run	at	isAt	global
A–DE	Run 1	0.5857	0.7621	0.6739
A–DE	Run 2	0.4952	0.5882	0.5417
A–DE	Run 3	0.5994	0.7498	0.6746
A–EN	Run 1	0.5289	0.6408	0.5848
A–EN	Run 2	0.3392	0.6586	0.4989
A–EN	Run 3	0.6009	0.6820	0.6415
A–FR	Run 1	0.6843	0.7483	0.7163
A–FR	Run 2	0.5397	0.6754	0.6076
A–FR	Run 3	0.6529	0.7179	0.6854
B–FR	Run 1	0.6338	—	0.6338
B–FR	Run 2	0.5509	—	0.5509
B–FR	Run 3	0.6613	—	0.6613

5. Analysis and Discussion

Why the domain-agnostic prompt remains competitive on newspapers. We expected the newspaper-specific prompt (Run 1) to perform best on Test A, and the domain-agnostic prompt (Run 3) to help mainly on the literary Test B. However, Run 3 also slightly improves over Run 1 on German and English. The label distributions (Table 9) suggest that this is not because Run 3 predicts a more current presence. In fact, it is more cautious: it assigns fewer pairs to TRUE and more to PROBABLE. This helps because the evaluation uses macro recall, so the rare PROBABLE class matters as much as the frequent classes. The same behavior appears on `isAt`: Run 3 predicts fewer current-presence labels than Run 1.

Table 9
Predicted label distributions of Run 1 and Run 3 on the Test A newspaper files

		at			isAt
Lang	Run	TRUE	PROB.	FALSE	TRUE
DE	Run 1	74	55	109	56
DE	Run 3	68	61	109	46
EN	Run 1	70	44	48	35
EN	Run 3	61	50	51	27
FR	Run 1	97	16	125	66
FR	Run 3	88	26	124	55

Strong on `at`, weaker on `isAt`. Our French result is worth a closer look because it separates the runs where our system is competitive and where it is not. On the French `at` label, Run 1 reaches a macro recall of 0.684, which is one of the highest `at` scores among the top systems on that language. What pulls our global score down is `isAt`: there we reach 0.748 compared to the first place (0.832). We can observe the same pattern for our runs across languages: deciding whether a person is related to a place seems to be easier for our model than deciding when the relation happened. The main direction to improve here is the temporal grounding.

Predicting PROBABLE is the hard part. Across our analysis, the main difficulty remains the PROBABLE class. For the fine-tuned model, the most frequent confusion in Table 3 is between FALSE and PROBABLE:

many pairs with no connection are predicted as plausible. This boundary is difficult because the difference between "no relation" and "a possible relation not confirmed as presence" is often very subtle. Improving this FALSE/PROBABLE distinction is therefore the main direction for future work.

A small model remains competitive, and consistent. The fine-tuned model is smaller than the prompted setup, and although it is less accurate, it remains usable. It also offers the guarantee that the logic-constrained loss and the post-processing produce zero violations of the `at`=FALSE → `isAt`=FALSE rule on all four test files. The ablation (Table 5) shows that the logic terms help when weighted lightly and hurt when heavy.

Limitations Our work has two main limitations. First, our two strongest runs rely on a closed model with paid APIs, which means they are not fully reproducible in the strictest sense. In contrast, Run 2 is based on a fine-tuned open model that can be reproduced using the provided code and dataset. Second, the small test sets may generate differences between runs, and the results should not be over-interpreted. The two ablations we report isolate the weight of the logic constraints (Table 5) and the contribution of the two feature groups (Table 6); we do not separately ablate the post-processing step.

6. Conclusion

In this paper, we present our participation in the HIPE-2026 task on person–place relation extraction. Our three-run submission includes: 1) a prompted Claude Sonnet 4 with Wikidata facts and few-shot examples, 2) a smaller fine-tuned XLM-RoBERTa-large with knowledge-graph, text-pattern features, and a logic-constrained loss, and 3) a domain-agnostic version of the first run for the literary test set. Our best results are obtained on the French data, where the first run ranks fourth among all submitted runs. The hardest part of the task is the PROBABLE class and the temporal `isAt` label, on which the models consistently return weaker results. For future work, we plan to focus on a better way to separate real connections from plausible ones and a stronger handling of when a connection is actually current. Combining the prompted and the fine-tuned systems shows promising results for future experiments.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to check the grammar, spelling, and formatting of the document. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended overview of CLEF HIPE 2020: Named entity processing on historical newspapers, in: Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS, Thessaloniki, Greece, 2020. doi:10.5281/zenodo.4117566.

[2] M. Ehrmann, M. Romanello, S. Najem-Meyer, A. Doucet, S. Clematide, Extended overview of HIPE-2022: Named entity recognition and linking in multilingual historical documents, in: Working

Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS, 2022. doi:10.5281/zenodo.6979577.

[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) 27:1–27:47. doi:10.1145/3604931.

[4] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, S. Clematide, Overview of HIPE-2026: Person–Place Relation Extraction from Multilingual Historical Texts, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. 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.

[5] J. Opitz, M. Ehrmann, C. Raclé, A. Michail, M. Romanello, E. Boros, S. Gabay, S. Clematide, Extended Overview of HIPE-2026: Evaluating Accurate and Efficient Person–Place Relation Extraction from Multilingual Historical Texts, in: E. 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.

[6] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85. doi:10.1145/2629489.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 1877–1901.

[8] 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 (ACL), 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[9] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Van den Broeck, A semantic loss function for deep learning with symbolic knowledge, in: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018, pp. 5502–5511.

[10] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, University Politehnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.

[11] M. Ehrmann, M. Romanello, S. Clematide, P. B. Ströbel, R. Barman, Language resources for historical newspapers: The impresso collection, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association, Marseille, France, 2020, pp. 958–968.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.