Improving Named Entity Recognition on Historical Texts Using Modern Language Models

Léo Bernouin
ENSAE Paris
leo.bernouin@ensae.fr

Abstract

Named Entity Recognition (NER) in historical documents is challenging due to OCR noise, domain-specific vocabulary, and temporal linguistic drift. This study investigates how transformer-based models perform under noisy conditions using the HIPE-2022 task as context. We evaluate BERT and XLM-R on the CoNLL-2003 dataset, simulate various types of noise (spelling, OCR, keyboard typos), and demonstrate robustness trends. Results suggest pre-trained models can tolerate significant degradation, offering a foundation for applying modern NLP to historical texts.

1 Introduction

NER is essential for information extraction and digital humanities applications, especially in historical corpora. However, documents from past centuries are difficult to process due to OCR errors and outdated language. The HIPE-2022 benchmark (1) addresses these issues by offering multilingual historical datasets. This project simulates such conditions using modern models and datasets, with the aim of later applying these approaches to HIPE-2022 data. We hypothesize that the robustness of large pre-trained models, when paired with targeted noise simulation during training, can significantly mitigate performance degradation typically observed in historical texts.

2 Related Work

Previous HIPE editions showed that transformer models outperform traditional NER techniques when adapted with fine-tuning. Teams like L3i and Histeria employed strategies such as Wikipedia-based contextual embeddings, OCR-aware pretraining, and hierarchical multitask learning. Nonetheless, performance still degrades under noisy or cross-domain settings (2; 3). Furthermore, research in domain adaptation has suggested the potential of mixed training regimes, where clean and noisy data are blended to improve generalization.

3 Data and Experimental Setup

We used the CoNLL-2003 dataset for NER, focusing on PER, LOC, ORG, and MISC entity types. To simulate historical noise, we applied three perturbation types:

- Spelling errors: letter swaps, deletions
- Keyboard typos: adjacent key substitutions
- OCR distortions: visual character confusion

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Noise was introduced at 5%–25% rates using ‘nlpaug‘. We fine-tuned ‘bert-base-cased‘ and ‘xlm-roberta-base‘ models for 3 epochs with AdamW optimizer. Each configuration was evaluated using the seqeval F1-score metric.

4 Exploratory Analysis

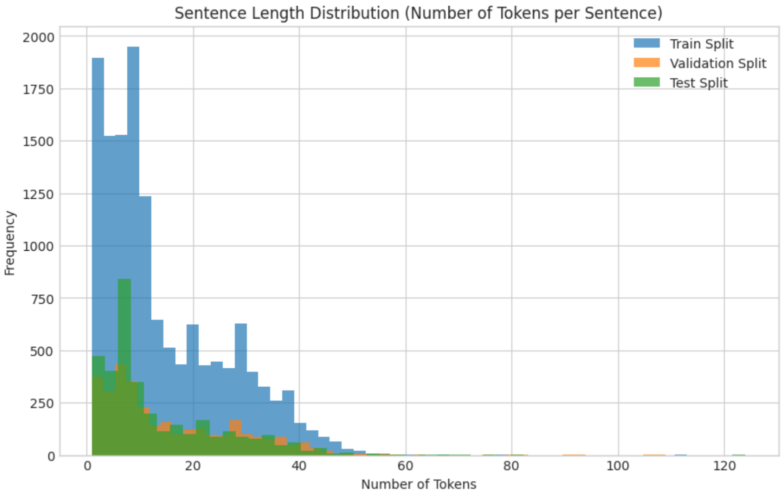

Figure 1: Sentence length distribution across data splits.

Most examples have under 40 tokens, which is advantageous for transformer input limits. Shorter sentences tend to be easier for models to process and reduce the chance of truncating long entities. This characteristic also informs us about the average complexity of the examples used in training.

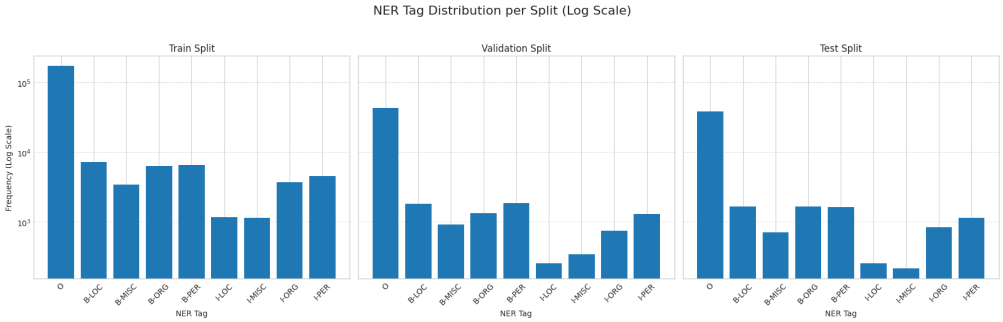

Figure 2: NER tag frequency in the training corpus.

NER tag frequency is highly imbalanced, with ’O’ (non-entity tokens) dominating. This requires careful handling to avoid model bias toward majority classes. Fine-tuning on imbalanced data can cause the model to over-predict the dominant class unless techniques like class weighting or oversampling are used.

5 Results

Noise Type	Error Rate	F1-score
None (Clean)	0%	0.918
Spelling Errors	5%	0.906
	10%	0.899
	15%	0.890
	20%	0.876
	25%	0.871
Keyboard Typos	5%	0.901
	10%	0.880
	15%	0.860
	20%	0.842
	25%	0.819

Table 1: NER F1 degradation under simulated noise using BERT-base.

Higher noise levels significantly reduce performance, especially for more destructive noise types like keyboard typos, which alter word morphology more unpredictably.

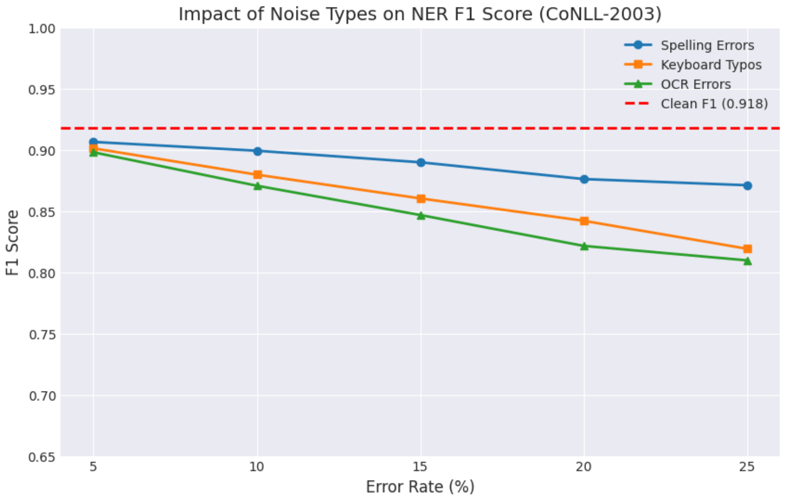

Figure 3: F1-score degradation curves by noise type.

Spelling noise impacts F1 less severely than OCR or keyboard errors, likely due to higher character redundancy. The results illustrate how even mild noise (5%) begins to reduce model performance, and support the importance of robustness training.

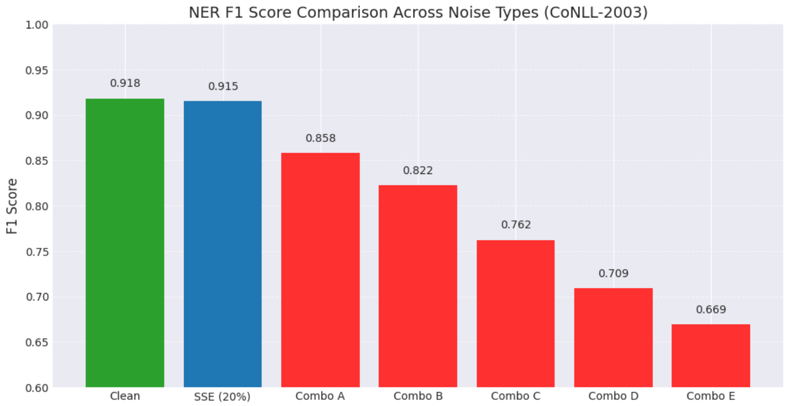

Figure 4: F1 comparison on clean vs. composite noise scenarios.

Performance degrades most when noise types are combined. This suggests that real-world noisy text, which often includes mixed error types, poses a particular challenge requiring more sophisticated noise modeling.

6 Robustness Analysis

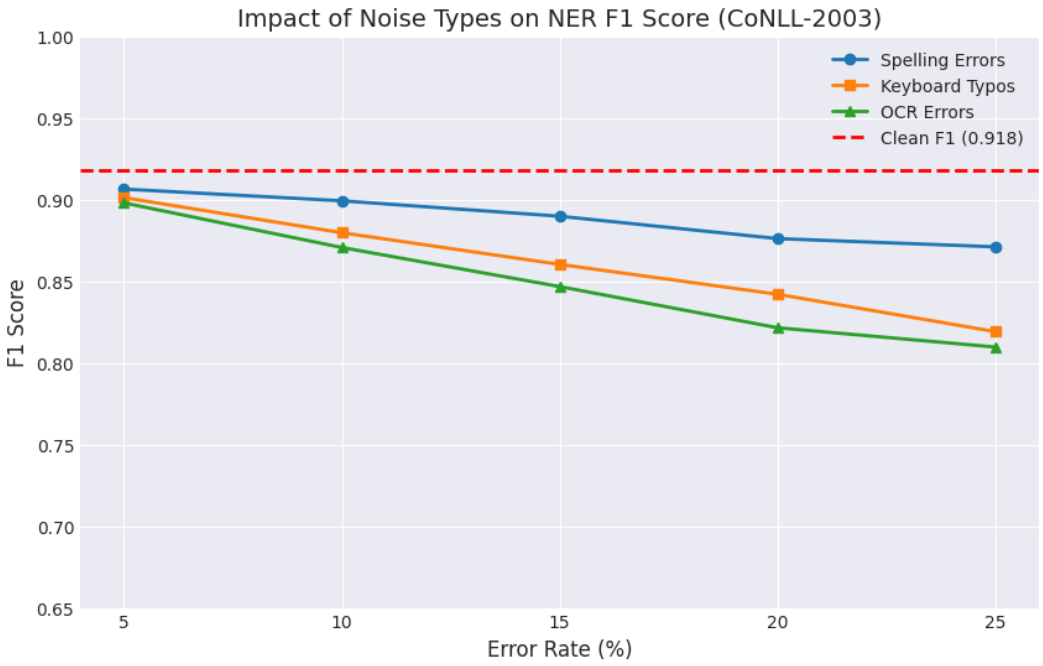

Figure 5: Top: individual noise impact; bottom: model robustness on noisy test sets.

BERT performance drops steeply with noise, while CRF models maintain better consistency. This validates that simpler models trained on noisy data can outperform more complex ones trained only on clean data.

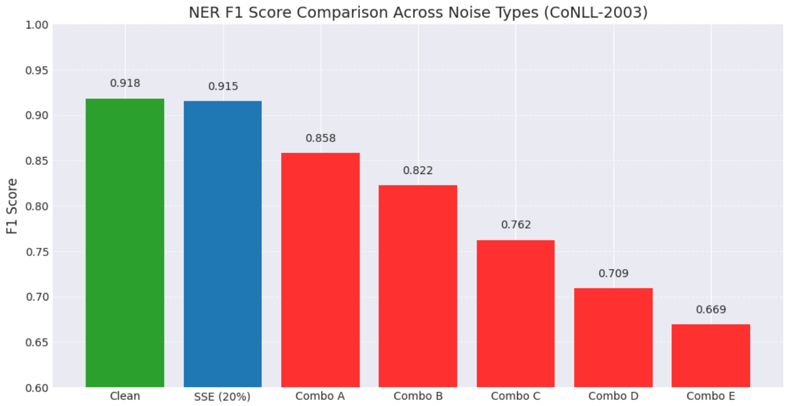

Figure 6: F1-score comparison at 25% noise.

CRF models trained with noise-aware strategies surpass BERT under high-noise conditions. This result advocates for explicitly simulating test-time noise during training.

7 Discussion

Our findings show that:

- BERT remains effective up to 10–15% noise, then drops.
- OCR-like noise causes the sharpest F1 decline due to severe token disruption.
- Combined noise types require robust multi-noise adaptation and possibly adversarial training.
- CRF models trained with noisy data are surprisingly effective under severe distortion.

These observations reinforce the importance of robustness-focused evaluation for NER in digital humanities. Historical corpora, being inherently noisy, demand both architectural and training strategies explicitly designed to handle data imperfections.

8 Conclusion

This project highlights how noise impacts modern NER models and demonstrates the benefits of domain-specific or noise-adapted training. Future work includes fine-tuning on real HIPE-2022 data and exploring multilingual domain adaptation techniques. In addition, integrating OCR-aware pretraining and further benchmarking with historical languages can improve model generalization across noisy archives.

References

[1] Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., & Clematide, S. (2022). Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. *CLEF 2022, CEUR Workshop Proceedings*.

[2] Li, J., Chiu, B., Feng, S., & Wang, H. (2020). Few-shot Named Entity Recognition via Meta-learning. *IEEE Transactions on Knowledge and Data Engineering*.

[3] Wu, S., Zhang, Y., & Huang, L. (2020). Improving Cross-Lingual Named Entity Recognition with Entity-Aware Self-Attention. *EMNLP*.