Impact of Noise on Named Entity Recognition

Léo Bernouin
ENSAE Paris
leo.bernouin@ensae.fr

Abstract

This paper investigates the impact of data quality on Named Entity Recognition (NER) systems. Using standard datasets such as CoNLL-2003 and various noise types including spelling mistakes, keyboard typos, OCR artifacts, and sentence shortening, we measure the robustness of modern NER models. Our findings align with recent literature, highlighting that OCR and combined noise significantly deteriorate performance, while sentence shortening has limited impact. We evaluate both transformer-based (BERT) and classical (CRF) models and propose noise-aware training strategies to improve robustness.

1 Introduction

Named Entity Recognition (NER) plays a key role in extracting structured information from unstructured text. However, real-world data often contains noise such as typos, spelling mistakes, and OCR errors, especially in historical or user-generated content. While state-of-the-art models achieve high scores on clean benchmarks, their resilience to noisy input is less studied. This work evaluates NER robustness under varying levels and types of noise using CoNLL-2003 and implements models including CRF and BERT.

2 Related Work

Several studies have explored the sensitivity of NLP models to noise Belinkov and Bisk (2018); Bodapati et al. (2019). The paper by Bhadauria et al. (2024) systematically introduced four types of noise—spelling errors, typos, OCR errors, and sentence shortening—and evaluated their impact on three NER models: CRF, BERT, and BiLSTM-Flair. They concluded that the combination of errors is most harmful, while sentence shortening has negligible impact.

3 Dataset Description

We use the CoNLL-2003 dataset, comprising labeled tokens in the IOB format with four entity types: PER, LOC, ORG, MISC. The dataset is divided into train, validation, and test splits. Figure 1 shows the sentence length distribution, while Figure 2 displays the tag distribution per split.

4 Experiment Setup

Following the procedure in Bhadauria et al. (2024), we introduce controlled noise into the datasets using NLPAug. We simulate five levels of noise (5% to 25%) for each of the following types:

- Spelling Errors
- Keyboard Typos

Preprint. Under review.

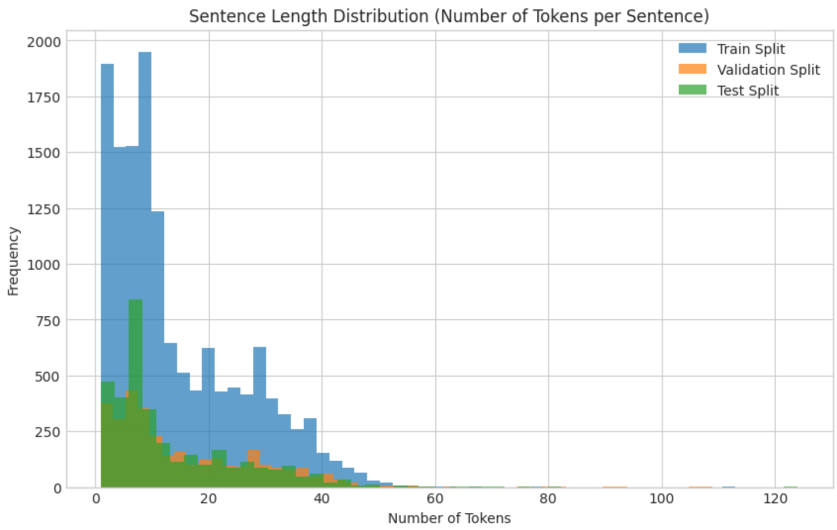

Figure 1: Sentence length distribution across dataset splits.

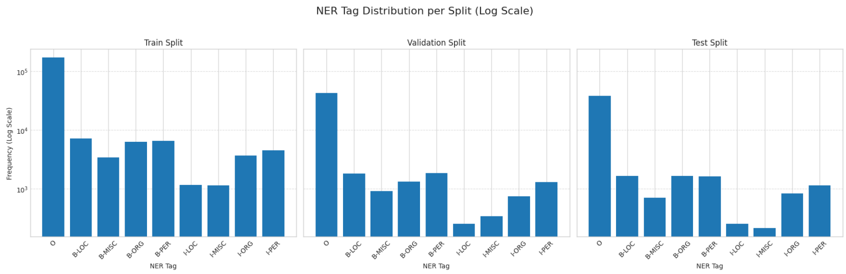

Figure 2: NER tag distribution in Train, Validation, and Test sets (log scale).

- OCR Errors
- Sentence Shortening Errors (SSE)

We evaluate:

1. Original model tested on noisy data
2. Model trained on noisy data and tested on clean and noisy sets

Models evaluated:

- BERT (bert-base-cased)
- CRF

5 Results

5.1 Noise Sensitivity of BERT

As shown in Figure 3, BERT's F1 score decreases significantly with increasing OCR and typo noise. Spelling errors have a smaller impact, and the model shows some resilience up to 10% noise. However, past this threshold, performance drops quickly, especially with OCR-induced distortions.

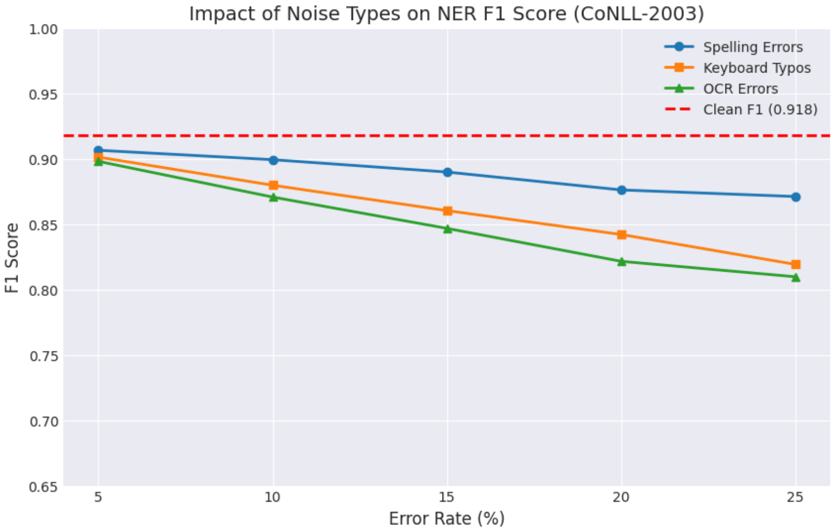

Figure 3: Impact of different noise types on BERT F1 score (CoNLL-2003).

5.2 Impact of Combined Noise

Figure 4 illustrates how performance declines progressively across noise combinations A to E. The cumulative effect of spelling, OCR, and typo noise leads to the steepest F1 drops, especially for BERT, whose performance dropped by nearly 30 points under combination E.

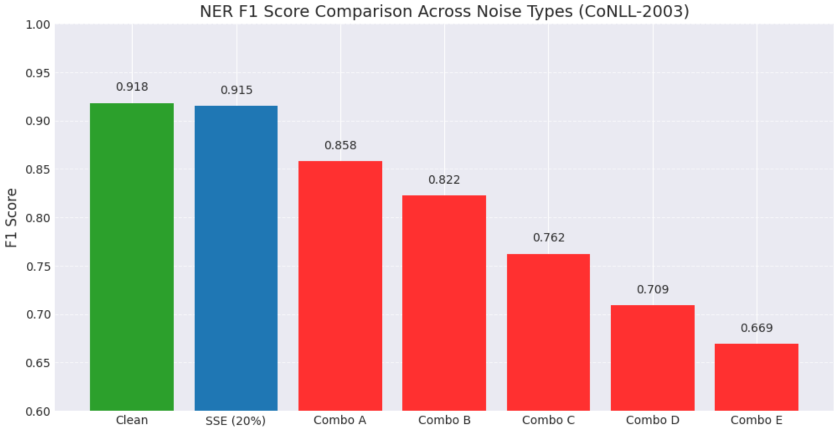

Figure 4: NER F1 Score comparison for different combinations of noise.

5.3 CRF Noise-Aware Training

In Figure 5, we observe that CRF trained on noisy data maintains better performance under OCR noise compared to CRF trained on clean data. The noise-aware CRF is more robust to distortions and shows more stable performance across all noise types, especially OCR.

5.4 Model Comparison at 25% Noise

As depicted in Figure 6, CRF outperforms BERT under heavy noise, especially in the mixed setting. This suggests that simpler models may generalize better when trained with noisy inputs. While BERT performs best on clean data, its robustness degrades sharply with combined noise.

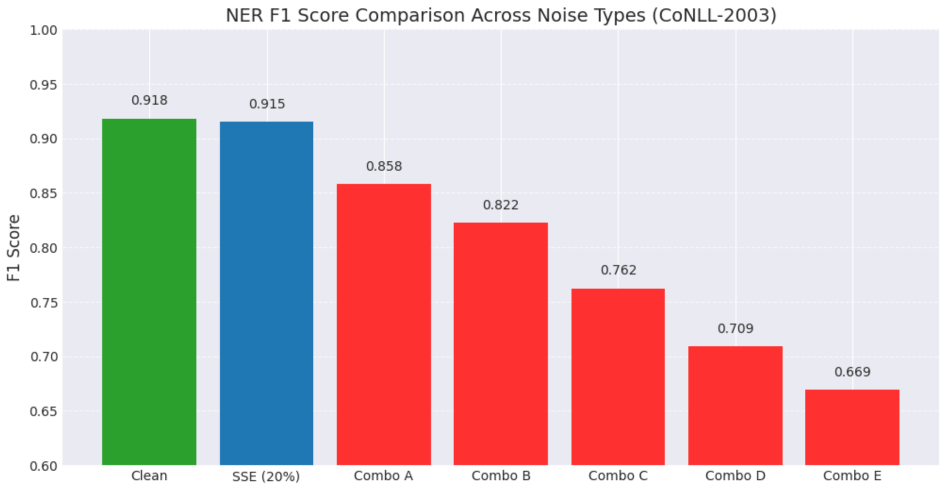

Figure 5: Clean vs. noise-aware CRF performance across noise types and test sets.

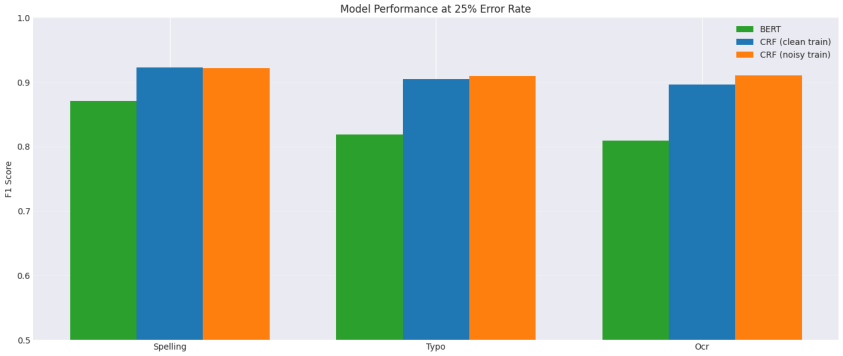

Figure 6: Model comparison under 25% noise (Spelling, Typo, OCR).

6 Discussion

Our results corroborate prior findings that noise has varying effects depending on its type. Sentence shortening is largely benign, while OCR and compound noise have drastic impact. Importantly, noise-aware training improves CRF robustness, offering a lightweight yet effective strategy for noisy environments. BERT, while powerful, is more sensitive to noise.

7 Conclusion

We analyzed the robustness of NER models under diverse noise scenarios. Our experiments, aligned with prior research, show that OCR and multi-error conditions most strongly affect model performance. Noise-aware training for simpler models like CRF offers promising mitigation. Future work may explore adversarial noise injection during training or ensemble approaches.

References

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bhadauria, D., Sierra-Múnera, A., and Krestel, R. (2024). The effects of data quality on named entity recognition. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT)*. Association for Computational Linguistics.

Bodapati, S., Yun, H., and Al-Onaizan, Y. (2019). Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT)*, pages 237–242. Association for Computational Linguistics.