Fine-tuning XLM-RoBERTa for Named Entity Recognition in Kurmanji Kurdish

Akam Nawzad and Hossein Hassani
University of Kurdistan Hewlêr
Kurdistan Region - Iraq
{akam.nawzad, hosseinh}@ukh.edu.krd

Abstract

Named Entity Recognition (NER) is the information extraction task of identifying predefined named entities such as person names, location names, organization names and more. High-resource languages have made significant progress in NER tasks. However, low-resource languages such as Kurmanji Kurdish have not seen the same advancements, due to these languages having less available data online. This research aims to close this gap by developing an NER system via fine-tuning XLM-RoBERTa on a manually annotated dataset for Kurmanji. The dataset used for fine-tuning consists of 7,919 annotated sentences, which were manually annotated by three native Kurmanji speakers. We selected the annotation based on the majority agreement, that is, when at least two of the three annotators agreed upon a certain NE class. The classes labeled in the dataset are Person (PER), Organization (ORG), and Location (LOC). A web-based application has also been developed using Streamlit to make the model more accessible. The model achieved an F1 score of 0.8735, precision of 0.8668, and recall of 0.8803, demonstrating the effectiveness of fine-tuning transformer-based models for NER tasks in low-resource languages. This work establishes a methodology that can be applied to other low-resource languages and Kurdish varieties.

1 Introduction

Despite recent advances in NLP technologies, low-resource languages such as Kurdish continue to lag behind high-resource languages. Among Kurdish varieties, Kurmanji (Northern Kurdish) is the most widely spoken, used by approximately 65% of the Kurdish population (Akin, 2011). It also constitutes the largest group in terms of geographical distribution and speaker numbers (Öpengin, 2021), yet it remains overshadowed by Sorani, which holds official language status in Iraq.

A key challenge for Kurmanji NLP is script variation. The language employs different writing systems across regions: a modified Perso-Arabic script in Iraqi Kurdistan and Iran, and the Latin-based Hawar alphabet in Turkey and Syria (Sheyholislami, 2009; Tavadze, 2019). This fragmentation complicates resource development and limits cross-regional data sharing. This research focuses on the Hawar alphabet due to its broader geographic usage and greater presence in digital text sources.

Named Entity Recognition (NER) serves as a fundamental building block for downstream NLP applications including information retrieval, machine translation, and question answering. Recent research has shown that fine-tuning pre-trained transformer models achieves superior performance for NER tasks on low-resource languages (Hanslo, 2022). Following this approach, we develop a reliable NER system for Kurmanji by fine-tuning XLM-RoBERTa on a manually annotated dataset. Our contributions include:

- The first publicly available NER system for Kurmanji Kurdish
- A manually annotated dataset of 7,919 sentences with 21,297 labeled entities
- Empirical validation of transformer fine-tuning effectiveness for Kurdish NLP

2 Related Work

2.1 Kurdish NLP and NER Research

Kurdish NLP faces unique challenges including script variation and resource scarcity (Esmaili, 2012). Previous work on Kurdish NER has been limited, with most research focusing on Sorani rather than Kurmanji.

Recent transformer-based approaches have shown promise for Kurdish varieties. Abdullah et al. (2024) fine-tuned RoBERTa on Sorani Kurdish, achieving 92.9% F-score for NER tasks. For

Kurmanji specifically, Morad et al. (2024) developed a transformer-based model for part-of-speech tagging, demonstrating that fine-tuned transformers outperform traditional approaches. However, dedicated NER systems for Kurmanji remain largely unexplored.

2.2 Transformer Models for Low-Resource NER

The introduction of transformer models, particularly BERT (Devlin et al., 2019) and its multilingual variants, has significantly advanced NER capabilities across languages. XLM-RoBERTa has emerged as particularly effective for cross-lingual tasks. Conneau et al. (2020) demonstrated that XLM-RoBERTa outperforms multilingual BERT across various benchmarks, achieving +2.4% F1 improvement on NER tasks.

Hanslo (2022) conducted comprehensive evaluations on ten low-resource South African languages, consistently finding that fine-tuned XLM-RoBERTa outperformed traditional CRF and BiLSTM approaches. Their results demonstrate that transformer fine-tuning can achieve strong performance even with limited training data, directly supporting the viability of our approach for Kurmanji.

3 Methodology

3.1 Data Collection and Annotation

We collected Kurmanji text written in the Hawar alphabet from multiple sources: Kurdish news websites (primary source), Kurdish Wikipedia, and the OSCAR dataset. Text containing higher densities of named entities was prioritized for annotation.

The collected data was manually annotated using the standard BIO (Beginning, Inside, Outside) tagging scheme for three entity types: Person (PER), Organization (ORG), and Location (LOC). Three native Kurmanji speakers performed the annotation using Label Studio, with each annotator handling a separate subset. To ensure quality, each subset was reviewed by another team member, and annotation guidelines were established to handle ambiguous cases consistently.

3.2 Data Preprocessing

Text preprocessing involved several standardization steps: normalizing punctuation, standardizing diacritics (e.g., replacing ș with ş), and normalizing whitespace. We employed XLM-RoBERTa's SentencePiece tokenizer, which operates directly on raw unsegmented text. During tokenization, we carefully maintained alignment between BIO tags and the resulting subword tokens.

3.3 Model Architecture and Training

We fine-tuned the base XLM-RoBERTa model, which was pre-trained on 100 languages including Kurdish. A token classification head was added on top of the transformer output to predict BIO tags (O, B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC).

We selected the hyperparameters through grid search, optimizing for validation performance within our computational constraints:

- Batch size: 8
- Learning rate: 2×10^{-5}
- Optimizer: AdamW with weight decay 0.01
- Training epochs: 5
- Maximum sequence length: 128
- Gradient clipping: 1.0

4 Results

4.1 Dataset Statistics

The final dataset contains 7,919 Kurmanji sentences with 231,981 tokens total. Table 1 shows the distribution across splits.

Table 1: Dataset Split and Statistics

Split	Sentences	Tokens	Entities
Training	6,414	187,893	17,238
Validation	713	20,886	1,915
Test	792	23,202	2,144
Total	7,919	231,981	21,297

Entity distribution shows LOC entities are most frequent (8,796), followed by ORG (6,414) and PER (6,087), reflecting the news-heavy nature of our corpus.

4.2 Model Performance

Table 2 shows the model's performance progression during training. The model achieved optimal validation performance in epoch 5.

Table 2: Validation performance across epochs

Epoch	Precision	Recall	F1
1	0.7905	0.8480	0.8182
2	0.8205	0.8737	0.8462
3	0.8362	0.8731	0.8543
4	0.8501	0.8809	0.8652
5	0.8480	0.8860	0.8666

4.3 Comparison with Multilingual Model

We compared our fine-tuned model against `Davlan/xlm-roberta-base-ner-hrl`, a multilingual XLM-RoBERTa model trained on 10 high-resource languages. Table 3 shows the test set performance.

Our fine-tuned model significantly outperforms the multilingual baseline across all entity types, with an overall F1 improvement of 11.0 percentage points. The largest improvements are for PER (+14.0 points) and ORG (+10.9 points) entities.

4.4 Error Analysis

Analysis of test set errors revealed that ORG entities present the greatest challenge. Common error patterns include:

- False negatives: failing to detect ORG entities
- False positives: incorrectly labeling non-entities as ORG
- Entity confusion: misclassifying ORG as LOC or vice versa

These patterns suggest that organizational naming conventions in Kurdish text lack standardization and often overlap with locational references.

5 Discussion

Our results demonstrate that relatively small, high-quality annotated datasets can achieve strong NER performance for low-resource languages. With 7,919 annotated sentences, we achieved performance competitive with systems trained on much larger datasets for high-resource languages.

The 11.0 percentage point improvement over the multilingual model is particularly significant given that both models share the same underlying architecture. This validates the importance of language-specific fine-tuning and suggests that cross-lingual transfer alone is insufficient for optimal performance on low-resource languages.

The performance variation across entity types (LOC: 0.904, PER: 0.901, ORG: 0.805) reflects inherent linguistic challenges. Kurdish organizational names often lack standardization and may incorporate location names, making them harder to distinguish.

Looking ahead, future work should explore developing script-agnostic models that can handle both Latin and Perso-Arabic Kurmanji, enabling broader accessibility. Additional directions include expanding the dataset to cover more domains such as social media, legal, and medical texts; applying data augmentation techniques to mitigate data scarcity; and extending the system to other Kurdish dialects such as Sorani and Zazaki.

6 Conclusion

We presented the first publicly available NER system for Kurmanji Kurdish, achieving an F1 score of 0.8735 through fine-tuning XLM-RoBERTa on a manually annotated dataset. Our work demonstrates that transformer-based approaches can successfully address NLP challenges in low-resource languages, even with modest amounts of training data.

Beyond technical achievements, this research contributes to digital inclusion and preservation of Kurdish linguistic heritage. The methodology established here provides a replicable framework for developing NER systems for other low-resource languages and Kurdish varieties.

Acknowledgments

We thank the native Kurmanji speakers who participated in the annotation process. Their expertise was essential for creating the high-quality dataset that made this research possible.

Limitations

While our models demonstrates strong performance on Kurmanji NER, several limitations should be acknowledged. First, our model handles only the Hawar alphabet, excluding Perso-Arabic script users in Iraq and Iran. Second, the dataset's news-domain bias may limit generalization to other text types. Third, the relatively lower ORG entity performance indicates room for improvement.

Ethics Statement

This work involved human annotators for creating the NER dataset. All annotators were native Kur-

Table 3: Test set performance comparison between the fine-tuned Kurdish NER model and a multilingual baseline.

Entity	Fine-tuned Kurdish NER			Multilingual Model		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
PER	0.8880	0.9143	0.9010	0.708	0.823	0.761
LOC	0.9179	0.8905	0.9040	0.816	0.840	0.828
ORG	0.7814	0.8308	0.8053	0.726	0.668	0.696
Overall	**0.8668**	**0.8803**	**0.8735**	**0.750**	**0.777**	**0.763**

manji speakers who volunteered for this research project and were fully informed about the purpose and intended use of their annotations.

Our dataset was constructed from publicly available sources including news websites and Wikipedia, containing no private or personally identifiable information beyond public figure names that naturally appear in news contexts.

We acknowledge that the system currently supports only the Latin-based Hawar alphabet. This decision was based on the greater availability of online Kurmanji text in this script, but we recognize that it may limit accessibility for speakers who use the Arabic script in regions such as Iraq and Iran.

We recognize that language technology development for minority languages carries both opportunities and risks. While NER systems can help preserve and promote Kurdish digital presence, they could potentially be misused for surveillance or discrimination. We encourage responsible use of our resources and will clearly document these considerations in our public release.

The developed resources will be released under an open license to benefit the Kurdish NLP research community while including clear guidelines for ethical use.

Data Availability Statement

We plan to publicly release the Kurmanji NER dataset and fine-tuned XLM-RoBERTa model upon publication. The dataset will include BIO-formatted annotations and will be distributed under an open license for research purposes. Access details, documentation, and usage guidelines will be provided via a dedicated GitHub repository.

References

Abdulhady Abas Abdullah, Srwa Hasan Abdulla, Dalia Mohammad Toufiq, Halgurd S. Maghdid, Tarik A. Rashid, Pakshan F. Farho, Shadan Sh. Sabr, Akar H. Taher, Darya S. Hamad, Hadi Veisi, and Aras T. Asaad. 2024. Ner- roberta: Fine-tuning roberta for named entity recognition (ner) within low-resource languages. *Preprint*, arXiv:2412.15252.

S. Akin. 2011. Language planning in diaspora: the case of the kurdish kurmanji dialect. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 2(1):9–27.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kyumars Sheykh Esmaili. 2012. Challenges in kurdish text processing. *Preprint*, arXiv:1212.0074.

Ricky Hanslo. 2022. Deep learning transformer architecture for named-entity recognition on low-resourced languages: State-of-the-art results. In *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 53–60. IEEE.

Peshmerge Morad, Sina Ahmadi, and Lorenzo Gatti. 2024. Part-of-speech tagging for Northern Kurdish. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 70–80, Torino, Italia. ELRA and ICCL.

Jaffer Sheyholislami. 2009. Language and nation-building in kurdistan-iraq. In *Middle Eastern Studies Association 43rd Annual Meeting*, Boston, MA, USA.

Givi Tavadze. 2019. Spreading of the kurdish language dialects and writing systems used in the middle east. *Bulletin of the Georgian National Academy of Sciences*, 13:170–174.

Ergin Öpengin. 2021. The history of kurdish and the development of literary kurmanji. In Hamit Bozarslan, Cengiz Güneş, and Veli Yadirgi, editors, *The Cambridge History of the Kurds*, pages 615–634. Cambridge University Press.