Named Entity Recognition for Confidential and Military Texts in the Uzbek Language

Obidjon Bozorov
National University of Uzbekistan, bozorov.obid@gmail.com

Abstract

This thesis presents a specialized Named Entity Recognition (NER) system for identifying sensitive entities in Uzbek confidential and military texts. By fine-tuning the bertbek-ner-uznews transformer model with domain-specific synthetic data, we achieved an F1-score of 89.6% across seven custom entity categories. The system addresses critical gaps in Uzbek NLP for security applications and provides a foundation for Data Loss Prevention (DLP) integration.

Keywords: Named Entity Recognition, Uzbek NLP, Military Text Processing, Information Security

1. Introduction

The widespread use of textual data in governmental and defense environments has increased the need for automated information security solutions. Named Entity Recognition (NER) plays a vital role in identifying sensitive entities within documents such as military codes, ranks, operations, and classified data. While most existing NER models are trained for global languages like English or Russian, there is a lack of domain-specific tools for the Uzbek language, especially in security contexts.

This research addresses this gap by fine-tuning a transformer-based model for NER in Uzbek, introducing new entity classes: military rank (B-RANK), military equipment (B-EQUIP), military codes (B-CODE), secret information (B-SECRET), military units (B-MIL), dates (B-DATE), and locations (B-LOC).

2. Related Work

NER has evolved from rule-based and statistical models (HMM, CRF) to transformer-based architectures like BERT, which have significantly improved entity recognition performance. Several Uzbek NER models exist, including bertbek-ner-uznews, but these lack specific labels for confidential or military-related entities. Specialized NER models for medical, legal, and financial domains demonstrate that domain adaptation is crucial for achieving optimal performance in specialized contexts.

3. Methodology

3.1 Dataset Development

Since real confidential military documents cannot be disclosed, we developed a comprehensive synthetic dataset in BIO format. The dataset generation process involved:

1. **Domain Analysis:** Analysis of publicly available military terminology and document formats
2. **Template Creation:** Development of 150 document templates representing various military document types
3. **Content Generation:** Creation of 5,000 synthetic documents with 75,000 sentences
4. **Annotation:** Manual annotation by three experts (94.2% inter-annotator agreement)

Dataset Statistics:

- Total sentences: 74,832
- Total tokens: 892,156
- Annotated entities: 156,743
- Training/Validation/Test: 80%/10%/10%

Entity Categories:

- **B-MIL:** Military units and formations
- **B-RANK:** Military ranks and positions
- **B-CODE:** Operational codes and classifications
- **B-EQUIP:** Military equipment and weapons
- **B-SECRET:** Confidentiality markers
- **B-DATE:** Temporal references
- **B-LOC:** Strategic locations

3.2 Model Architecture

We selected bertbek-ner-uznews as the base model and implemented strategic fine-tuning with:

Training Configuration:

- Batch size: 16
- Learning rate: 3e-5 (with layer-wise scheduling)
- Epochs: 15 with early stopping
- Optimizer: AdamW with weight decay
- Loss function: Cross-entropy with label smoothing

3.3 Evaluation

Performance was measured using standard metrics:

- **Precision:** $P = \frac{TP}{TP + FP}$
- **Recall:** $R = \frac{TP\ )}{(TP + FN)}$

• **F1-Score:** $F1 = \frac{2 \times (P \times R)}{(P + R)}$

4. Results and Discussion

4.1 Overall Performance

The fine-tuned model demonstrated significant improvements over the baseline:

Metric	Baseline	Fine-tuned	Improvement
Precision	76.3%	90.4%	+14.1%
Recall	71.8%	88.9%	+17.1%
F1-Score	74.0%	89.6%	+15.6%

4.2 Entity-Specific Analysis

Entity	Precision	Recall	F1-Score
B-RANK	96.1%	94.8%	95.4%
B-DATE	94.7%	93.1%	93.9%
B-MIL	93.2%	91.7%	92.4%
B-LOC	91.8%	90.3%	91.0%
B-CODE	89.7%	87.2%	88.4%
B-EQUIP	88.4%	86.9%	87.6%
B-SECRET	85.9%	82.4%	84.1%

High-performing categories like B-RANK and B-DATE benefited from structured patterns and standardized formats. **B-SECRET** presented challenges due to high linguistic variability in confidentiality expressions.

4.3 Error Analysis

Main error types identified:

• **Boundary detection errors (23%):** Incorrect entity boundaries for multi-word designations

• **Category confusion (18%):** Inter-category confusion, especially between B-EQUIP and B-CODE

• **Novel variants (15%):** Entities not represented in synthetic training data

4.4 Practical Implementation

System Specifications:

• Model size: 112M parameters

• Inference speed: 145 sentences/second (GPU)

• Memory requirement: 2.3GB VRAM

• API response time: <200ms for document analysis

5. Security and Privacy Considerations

The research prioritizes security throughout development:

• **No real data exposure:** Exclusive use of synthetic data

- **Secure development environment:** Isolated, secured development
- **Deployment security:** On-premises hosting recommendations with encrypted communications
- **Ethical compliance:** Strict adherence to confidentiality guidelines

6. Future Work

Planned extensions include:

1. **Dataset Expansion:** Additional document types and multi-lingual support
2. **Multimodal Integration:** PDF processing and OCR capabilities
3. **Active Learning:** Continuous improvement with minimal annotation
4. **Domain Extensions:** Medical, financial, and legal applications
5. **Real-world Validation:** Partnerships for practical testing

7. Conclusion

This research successfully demonstrates the development of specialized NER systems for Uzbek confidential and military texts. Key achievements include:

- **Novel Application:** First specialized NER system for Uzbek military contexts
- **Strong Performance:** 89.6% F1-score across custom entity categories
- **Security Compliance:** Methodology ensuring no confidential data exposure
- **Practical Applicability:** Integration potential with existing DLP systems

The developed system addresses critical cybersecurity infrastructure gaps for Uzbek-speaking organizations while establishing a methodological framework for secure NLP development in sensitive domains. This work demonstrates that sophisticated NLP capabilities can be developed responsibly for sensitive applications while contributing to digital equity and technological sovereignty.

Limitations include synthetic data constraints, evolving terminology requirements, and potential cross-domain generalization challenges. Despite these limitations, the research provides a solid foundation for advancing secure NLP applications in under-resourced languages.

The successful integration of security considerations throughout development provides a template for responsible AI development in sensitive applications, demonstrating the balance between technical advancement, security requirements, and ethical responsibility.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

2. Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint arXiv:1508.01991*.
3. Mengliev, D., Barakhnin, V., Abdurakhmonova, N., & Eshkulov, M. (2024). Developing named entity recognition algorithms for Uzbek: Dataset insights and implementation. *Data in Brief*, 54, 110413.
4. Mengliev, D., Barakhnin, V., Eshkulov, M., Ibragimov, B., & Madirimov, S. (2024). A comprehensive dataset and neural network approach for named entity recognition in the Uzbek language. *Data in Brief*, 58, 111249.
5. elmurod1202. (2023). bertbek-ner-uznews: BERT-based Named Entity Recognition for Uzbek News. *Hugging Face Model Hub*. Available at: https://huggingface.co/elmurod1202/bertbek-ner-uznews
6. Li, X., Li, D., Yang, Z., Zhao, H., Cai, W., & Lin, X. (2023). ND-NER: A Named Entity Recognition Dataset for OSINT Towards the National Defense Domain. In *Neural Information Processing. ICONIP 2022. Communications in Computer and Information Science*, vol 1792. Springer.
7. Nitzl, C., Cyran, A., Krstanovic, S., & Borghoff, U. (2025). The Application of Named Entity Recognition in Military Intelligence. In *Computer Aided Systems Theory – EUROCAST 2024. Lecture Notes in Computer Science*, vol 15172. Springer.
8. Wang, X. R., Xiong, Z. H., & Du, X. Y. (2020). NER in threat intelligence domain with TSFL. *Proceedings of the 9th International Conference on Natural Language Processing and Chinese Computing*, 157–169.