MIDI 2025 Conference

Beyond Randomization: Multi-Factor, Ontology-guided Synthetic Personal Data Generation for Document-centric AI Systems

Abstract

Modern ICT systems are expected to process stored information responsibly, particularly when handling personal information. Yet, the developed models/systems must be trained before they can protect privacy effectively. This creates a paradox: how can models learn to recognize and handle personal data if they should not be trained using personal data records for security reasons? One of the available options is synthetic data generation. However, conventional randomization-based techniques often produce unrealistic distributions or internally inconsistent records, limiting their utility for building synthetic training document datasets, especially for classification, extraction, and de-identification systems.

This work proposes multi-factor synthetic personal data generation methodology, designed to preserve statistical realism, semantic coherence, and inter-attribute dependencies. It uses empirical-frequency sampling, algorithmic construction of structured identifiers, enforcing ontology-guided rules to maintain consistency across related attributes. Additionally, some of the attributes are generated by language models, if creativity is required. The resulting synthetic records are semantically rich, and distributionally aligned with population-level statistics. By merging statistical grounding with ontology-based modeling of inter-attribute dependencies, the proposed approach offers a practical, privacy-preserving synthetic data generation framework for modern document-centric Natural Language Processing systems, contributing to future research on fairness assessment, personal data datasets synthesis, and de-identification systems.