CLEAR

Comprehensible Learning for Entity Anonymization and Recognition

Programm / Ausschreibung	KIRAS, Kooperative F&E-Projekte, KIRAS-Kybernet-Pass CS Kooperative F&E Projekte (CS KFE_2024)	**Status**	laufend
Projektstart	01.10.2025	**Projektende**	31.03.2027
Zeitraum	2025 - 2027	**Projektlaufzeit**	18 Monate
Projektförderung	€ 576.334		
Keywords	Named Entity Recognition; Deep-Learning; Regel-Lernen; Anonymisierung; Digitale Forensik		

Projektbeschreibung

Sollen Texte, die personenbezogene Daten (pb Daten) enthalten, für das Training von KI-Systemen, für Forschungs- und Schulungszwecke verwendet oder veröffentlicht werden (z.B. parlamentarische Materialien bzw. Gerichtsentscheidungen), müssen sie vorher anonymisiert oder pseudonymisiert werden. Voraussetzung dafür ist eine verlässliche und nachvollziehbare Identifikation der Personenbezüge.

CLEAR entwickelt und erforscht generische, transparente, vertrauenswürdige und nachhaltige (KI-)Lösungen für die Erkennung von Entitäten und ihre Anwendung auf die Identifikation von pb Daten. Dabei werden regelbasierte und auf Machine-Learning basierende Methoden zur Realisierung ihrer Vorteile bei gleichzeitiger Vermeidung der Nachteile kombiniert.
SOTA-Lösungen für die Erkennung von Entitäten beruhen auf der Feinabstimmung großer neuronaler Sprachmodelle, die hochwertig annotierte Trainingsdaten erfordern, aber sich schlecht verallgemeinern lassen. Die Anfälligkeit für Halluzinationen führt dazu, dass das Vertrauen der Nutzer:innen sinkt bzw. Desinformation verstärkt wird. Ihre inhärente „Black Box“-Natur konfrontiert Anwender:innen mit Entscheidungen, welche unvorhersehbar und nicht erklärbar sind. Die Modelle sind nicht konfigurierbar und anfällig für Bias. Darüber hinaus stellt das Training eine erhebliche Umweltbelastung dar. Herkömmliche regelbasierte Systeme hingegen müssen auf auch nur geringfügig abweichende Gegebenheiten neu angepasst werden.

Im Projekt CLEAR entwickeln und evaluieren wir einen hybriden Ansatz einer akkuraten NER (Named Entity Recognition) auf deutschsprachige (Fließ-)Texte:
(1) Lernen von Regeln für die Erkennung von Entitäten durch Prompting bzw. Finetuning von LLM-basierten Modellen.
(2) Generieren von bewerteten Entitätenkandidaten durch Deep-Learning-Modelle und Auswahl der passendsten Kandidaten durch einen (trainierten) anwendungsspezifischen Regelsatz.
CLEAR basiert auf dem Human-in-the-Loop-Lernparadigma für juristische NER, der die oben genannten Schwächen

überwinden soll. Die Erklärbarkeit und Vorhersehbarkeit ist gegeben, die erstellten Regeln sind für Fachanwender:innen verständlich, prüfbar und einfach konfigurierbar. CLEAR bietet ein NLP-Paradigma, das die Umweltkosten sowie den Trainingsaufwand für LLMs erheblich zu senken vermag.

Im rechtswissenschaftlichen Bereich sind wichtige Fragen zum Begriff der Anonymisierung offen. Es gilt eine praktikable und rechtlich sichere Anonymisierungsstrategie zu identifizieren, da die DSGVO ebenso wie neue EU-Rechtsakte (etwa der Data Act und der Data Governance Act) auf dem Konzept der Anonymisierung aufbauen, ohne dieses vollends zu definieren. Zusätzlich sind für die Nutzung von Trainingsdaten für KI ungeklärte Randbedingungen, wie urheberrechtliche Aspekte, zu beachten. Ebenso sollen Fragen der neue KI-Verordnung der EU im europäischen Rechtsrahmen behandelt werden (z.B. zur Forschungsausnahme oder zur Risikoeinstufung von KI-Systemen).

Die angestrebte flexible, trainierbare, vertrauenswürdige NER-Architektur kann, neben dem Schwerpunkt 9, z.B. für: 2, 10, 13, diverse KIRAS-Schwerpunkte zu Data Governance sowie in einer Reihe von weiteren Anwendungsfällen zum Einsatz kommen, etwa in der digitalen Forensik und zur Bekämpfung von Cyber Crime.

Das Konsortium setzt sich aus hochkarätigen wissenschaftlichen Partnern, bedeutenden öffentlichen Bedarfsträgern sowie einem Unternehmenspartner mit langjähriger Erfahrung im Anwendungsbereich zusammen.

Abstract

Text containing personal data must be anonymized or pseudonymized before it can be used for training AI models or for research and educational purposes. Anonymization may also be necessary for publishing parliamentary materials or judicial decisions. Anonymization requires reliable and comprehensible identification of personal information.

CLEAR intends to develop generic, transparent, reliable, and sustainable AI solutions for named entity recognition (NER) and its application to identifying personal data. This will involve a combination of rule-based and ML-based methods in a way that exploits the advantages of both paradigms.
State of the art NER solutions rely on task-specific fine-tuning of large neural language models. Such models require high quality annotated training data and still fail to generalize. Their tendency to hallucinate reduces user trust and causes misinformation. Their inherent “black box” nature results in decisions that are neither predictable, nor explainable. Such models are not configurable, prone to bias, and create a significant environmental burden. Common rule-based systems, on the other hand, require laborious manual configuration to adapt to changing requirements.

The CLEAR project seeks to develop and evaluate hybrid NER methods for the processing of German texts:
(1) Rule learning for NER via prompting and fine-tuning of LLMs.
(2) Generation of entity candidates with deep learning models followed by the selection of entities using learned rules.

CLEAR relies on a human-in-the-loop learning approach for legal NER that mitigates the afore mentioned issues. Rule-based models are explainable, predictable, and auditable, as well as configurable and comprehensible for their users. CLEAR also offers a learning paradigm that greatly reduces the need to train LLMs and therefore environmental costs.

The concept of anonymization raises important questions in the field of legal research. New EU legislation such as the Data

Act and the Data Governance Act relies on the GDPR’s concept of anonymization, without settling the open questions, thus creating the need for a practical and legally secure anonymization strategy. There are also unresolved questions around the issue of intellectual property law in conjunction with data use for training AI models, and within the European legal framework of the AI Act, e.g. regarding the research exemption or the risk classification of AI systems.

The flexible, trainable, trustworthy NER architecture to be developed in CLEAR will have impact beyond focus area 9, and could be used for e.g. areas 2, 10, 13, various KIRAS research fields on data governance, and in several other use cases including digital forensics and the fight against cybercrime.

The consortium consists of high-profile scientific partners, multiple governmental organizations as stakeholders, and a corporate partner with several years of experience in the application field.

Projektkoordinator

- m2n - consulting and development gmbh

Projektpartner

- Bundesministerium für Finanzen
- Universität Wien
- Bundesministerium für Justiz
- Republik Österreich Parlamentsdirektion
- Technische Universität Wien