**BC/NW 2022 № 1 (39):6.2**

**СПОСОБЫ РАСПОЗНАВАНИЯ И ОБЕЗЛИЧИВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ В АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ**

**Бабак Н. Г., Оцоков Ш.А.**

Предварительным этапом обезличивания персональных данных (ПДн) является их распознавание среди всей информации. Ручное выделение определённого вида информации не только сильно замедляет процесс, но и всё ещё подвержено риску человеческой ошибки.

Существует несколько основных автоматизированных способов распознать какую-либо информацию – это словари, регулярные выражения и алгоритмы машинного обучения. Поскольку существует большое разнообразие данных, прямо или косвенно идентифицирующих человека, например фамилия, имя и отчество (ФИО), серия и номер паспорта, номер телефона, то для их распознавания придется создавать и постоянно поддерживать в актуальном состоянии большие словари, что требует дополнительных временных затрат и повышает риск пропустить критичную информацию. Решить данную проблему можно с помощью машинного обучения. В частности, задача распознавания ПДн сводится к задаче распознавания именованных сущностей [1]. Существует несколько основных способов решения этой задачи, среди которых наиболее передовым является архитектура нейросетей – трансформеров [2].

Распознанные ПДн необходимо не просто удалить из текста, а заменить с учётом синонимических и морфологических особенностей, то есть произвести «умное» обезличивание. Для генерации данных, на которые производится замена, разработаны следующие алгоритмы:

1. Генерация синонимичной сущности с помощью регулярных выражений и контрольных разрядов;
2. Замена значением из словаря и сохранение морфологических особенностей текста с помощью нейросетей;
3. Замена одинаковых сущностей на одинаковые для сохранения смысла и контекста текста.

Более того, необходимо анализировать тип входных данных для сохранения структуры в обезличенных документах. Для обработки сканов документов необходимо включить в процесс обезличивания систему распознавания текста на изображениях.

**Литература**

1. **Arya Roy.** Recent Trends in Named Entity Recognition (NER) // Carnegie Mellon University, 2021. arXiv:2101.11420v1.
2. **Vaswani A. et al.** Attention Is All You Need // 31st Conference on Neural Information Processing Systems, 2017. arXiv:1706.03762v5.