

УДК 004.912

ПОДБОР ЛЕКСИЧЕСКИХ ПРИЗНАКОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОПРЕДЕЛЕНИЯ ДИКТОРА

Рудалева Е.А (Университет ИТМО)
Научный руководитель – Попов Д.В.
(ООО ЦРТ-Инновации)

Доклад посвящен сравнительному анализу лексики внутри формального делового регистра, а также выявлению лексических признаков, характерных речи конкретного диктора. На основе подхода выделения признаков производится разработка методики выявления ошибок диаризации, связанных с лингвистической информацией.

Введение. В области речевых технологий активно развивается сфера детектирования дикторов. Большое внимание в исследованиях процесса диаризации уделяется разработке методов для повышения ее качества. В данной работе производится выделение признаков специфической лексики, а также дискурсивных паттернов для задачи определения диктора с использованием методов цифровой обработки текстов.

Основная часть. Работа посвящена компаративному анализу лексических паттернов внутри диалоговой речи формального регистра из офиса обслуживания. В рамках основной задачи производится автоматизация метода формирования лексических словарей специфической лексики дикторов. Процесс выявления паттернов включает в себя этапы предобработки данных, выделения лексических признаков (*word2vec*, *tf-idf*) и формирования более широких дискурсивных единиц, являющихся структурными частями диалога.

Выводы. В качестве результатов исследования выявлены лексические и дискурсивные признаки дикторов внутри используемой базы данных офиса обслуживания, возможные к применению для задачи определения диктора и разработки методики выявления ошибок диаризации.