УДК 004.37

ПРОБЛЕМЫ АВТОМАТИЧЕСКОЙ АТРИБУЦИИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Мирославская Л.А. (Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»), **Труханов А.С.** (Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»),)
Научный руководитель – д.т.н., профессор Бессмертный И.А.
(Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»),)

Аннотация Задача атрибуции естественно-языковых текстов встречается в разных областях и представляет интерес для разных специалистов, в том числе: филологов, литературоведов, юристов, криминалистов и историков. Зачастую определение авторства неизвестных текстов осуществляется экспертными методами. При идентификации автора или определении принадлежности текста другому автору в качестве основных критериев используют характерные языковые особенности и стилистические приемы. В статье рассмотрены проблемы атрибуции авторов текстов и возможные методы их решения.

Введение. Объектом исследования настоящей работы являются методы атрибуции естественно-языковых текстов. **Предметом** – алгоритмы и методы автоматической атрибуции естественно-языковых текстов на основе характеристик автора.

В рамках данного исследования были рассмотрены формальные методы атрибуции текстов, существующие программные средства определения авторства и использование квантового формализма в задачах автоматической обработки ЕЯ текстов.

В разных работах авторы рассматривают программные средства определения автора текста, основанные на формальных математических моделях.

Основная часть. Задачи атрибуции текстов делятся на идентификационные и диагностические. К первому типу относится: подтверждение или опровержение авторства кандидата и проверка текста на принадлежность конкретному автору. Ко второму типу задач относится определение личностных характеристик автора: уровень образования, знание иностранных языков, место рождения, профессия, хобби, пол, возраст, использование речевых стилей.

При автоматическом анализе языковых особенностей обычно рассматривают:

- особенности пунктуации, чтобы найти характерные для автора ошибки и авторские знаки;
- соблюдение орфографических правил, чтобы определить уровень знаний автора и выявить типичные для него ошибки;
- употребление фразеологизмов, неологизмов и других выразительных средств для определения словарного запаса автора;
- особенности стилистики, чтобы выявить характерные для автора стилистические приемы.

Для улучшения результатов, ускорения процесса и повышения качества атрибуции текстов применяют подходы из теории распознавания образов, математической статистики и теории вероятностей. Также используются алгоритмы нейронных сетей, кластерного анализа и прочие. В то время, как возможность применения квантового формализма для автоматической атрибуции естественно-языковых текстов еще не исследована. Среди проблем традиционных методов решения задачи установления авторства ЕЯ текстов стоит отметить проблему выбора лингвостилистических параметров текста и составления выборки эталонных текстов.

Выводы. В рамках данного исследования проведен ряд экспериментов и анализ существующих методов и алгоритмов автоматической атрибуции естественно-языковых текстов, выявлены их недостатки и выдвинуты предложения по улучшению с использованием квантового формализма. Предпринята попытка поиска новых критериев атрибуции авторов текстов, а также протестирован подход, основанный на комплексном анализе различных характеристик и критериев. Предложенные подходы показали небольшой прирост точности атрибуции авторов естественно-языковых текстов.

Мирославская Л.А. (автор)	Подпись
Труханов А.С. (соавтор)	Подпись
Бессмертный И.А. (научный руководитель)	Подпись