

ОТЗЫВ

научного руководителя на диссертационную работу Ерболовой Асель Серикановны «Исследование структур естественного языка в задаче идентификации ботов», представленную на соискание степени доктора философии (PhD) по специальности 8D06101 – Информатика, вычислительная техника и управление

Диссертационное исследование Ерболовой Асель Серикановны выполнено на актуальную тему, связанную с разработкой методов анализа крупномасштабной структуры естественного языка и их применением для решения важной прикладной задачи – различения текстов, созданных человеком, и текстов, сгенерированных автоматизированными системами (ботами). Актуальность темы обусловлена стремительным распространением технологий генеративного искусственного интеллекта и необходимостью обеспечения информационной безопасности, а также верификации цифрового контента в условиях роста объёмов синтетических текстовых данных.

В ходе выполнения диссертационного исследования соискатель проявил себя как сформировавшийся исследователь, способный самостоятельно формулировать и решать сложные научные задачи, применять современные математические методы и разрабатывать эффективные программные реализации предложенных алгоритмов. Работа выполнена на высоком научно-методическом уровне, с привлечением обширного эмпирического материала (52 языка из 18 языковых семей) и использованием комплекса методов статистического анализа, нелинейной динамики, фрактальной геометрии, топологического анализа данных и машинного обучения.

Структура и содержание работы полностью соответствуют поставленной цели и задачам. Диссертация состоит из введения, четырёх глав, заключения, списка использованных источников и пяти приложений.

В первой главе проведён всесторонний обзор современных подходов к анализу структуры естественного языка и методов идентификации ботов. Особое внимание уделено концепции языка как самоорганизованно-критичной системы, анализу хаотичности семантических траекторий и применению топологического анализа данных для выявления структурных особенностей языковых единиц. Обзор отличается глубиной и критическим осмыслением существующих работ, что позволило автору чётко определить место собственного исследования и сформулировать нерешённые проблемы.

Вторая глава посвящена описанию разработанных и адаптированных методов анализа. Представлены статистические критерии для проверки

гипотез о степенном характере распределений, методика построения семантических траекторий и оценки их хаотичности на плоскости энтропия–сложность, алгоритмы оценки внутренней размерности языковых фрактальных структур (включая методы Швайнхарта и Брито), процедура выявления персистентных гомологий и оконтуривания «дыр» в семантическом пространстве языка. Подробно изложена методология формирования выборок, предобработки данных и построения классификационных моделей для идентификации ботов.

Третья глава содержит основные экспериментальные результаты исследования крупномасштабной структуры естественного языка. Показано, что для подавляющего большинства естественных языков распределение длин текстов подчиняется степенному закону (исключение – искусственный язык эсперанто). Проведён анализ хаотичности семантических траекторий, выявлены кластеры языков по параметрам энтропии и сложности, установлены корреляции с лингвистическими характеристиками (порядок слов, тип маркировки и др.). Оценены внутренние размерности языковых фрактальных структур для униграмм, биграмм и триграмм, подтверждена мультифрактальная природа языка. Методами топологического анализа данных выделены значимые гомологии первого порядка, определённые как «слепые зоны» языка, и показано, что тексты, созданные ботами, статистически значимо удалены от границ этих зон по сравнению с текстами людей.

Четвёртая глава демонстрирует практическое применение разработанных методов для задачи идентификации ботов. Предложены классификаторы, основанные на различных группах признаков: положении точки на плоскости энтропия–сложность, характеристиках семантических траекторий, кластеризации n-грамм и расстояниях до границ персистентных гомологий. Эксперименты, проведённые по строгой схеме с разделением моделей-генераторов на обучающие и тестовые, показали высокую эффективность предложенных подходов: для разных языков достигнуты значения F1-меры до 0,98.

Научная новизна диссертации не вызывает сомнений. К наиболее значимым новым результатам, полученным лично автором, относятся:

1. Подтверждение гипотезы о естественном языке как самоорганизованно-критичной системе на широком спектре языков (52 языка) с выявлением уникальности искусственного языка эсперанто, демонстрирующего нормальное распределение длин текстов.
2. Разработка и апробация методологии оценки хаотичности семантических траекторий для сравнительного анализа языков,

позволившей установить типологические кластеры, коррелирующие с энтропийно-сложностными характеристиками.

3. Создание и верификация методов оценки внутренней размерности языковых фрактальных структур, показавших мультифрактальную природу языка и инвариантность размерностей относительно способов получения векторных представлений.
4. Применение аппарата персистентных гомологий для выявления и оконтуривания топологических «дыр» в семантическом пространстве языка, что дало новый инструмент для анализа крупномасштабной структуры языка и различения человеческих и бот-текстов.
5. Построение высокоэффективных классификационных моделей для идентификации ботов, сохраняющих обобщающую способность при тестировании на ранее не встречавшихся генеративных моделях.

Практическая значимость работы заключается в том, что разработанные методы и алгоритмы могут быть использованы для создания систем мониторинга информационной безопасности, автоматической модерации контента, оценки качества текстов и анализа лингвистических данных в различных предметных областях. Результаты внедрены в Национальном университете обороны Республики Казахстан, Карагандинском университете Казпотребсоюза и Лаборатории анализа семантики НИУ ВШЭ (Российская Федерация), что подтверждено соответствующими актами.

Достоверность и обоснованность полученных результатов обеспечивается корректным применением математического аппарата, тщательной верификацией алгоритмов на синтетических данных с известными свойствами, большим объёмом вычислительных экспериментов и использованием общепринятых статистических критериев. Основные положения диссертации прошли апробацию на международных научных конференциях и опубликованы в 7 печатных работах, из которых 5 статей в журналах, индексируемых в базе данных Scopus (процентили от 39 до 80), 1 статья в издании, рекомендованном КОКСНВО, и 1 публикация в материалах международной конференции. Получено авторское свидетельство на объект интеллектуальной собственности.

Личный вклад соискателя является определяющим на всех этапах работы: постановка задач, выбор и адаптация методов, разработка алгоритмов и программного обеспечения, проведение вычислительных экспериментов, анализ и интерпретация результатов, подготовка публикаций. В совместных публикациях автору принадлежат основные идеи, реализация и анализ результатов.

Заключение

Диссертационная работа Ерболовой Асель Серикановны «Исследование структур естественного языка в задаче идентификации ботов» является завершённой научно-квалификационной работой, выполненной на актуальную тему. В ней решена важная научная задача разработки методов анализа крупномасштабной структуры естественного языка и их применения для идентификации текстов, сгенерированных ботами. Работа полностью соответствует требованиям, предъявляемым к диссертациям на соискание степени доктора философии (PhD) по специальности 8D06101 – Информатика, вычислительная техника и управление. Рекомендую диссертацию к защите.

Научный руководитель:

кандидат технических наук,
ассоциированный профессор (доцент),
профессор кафедры «Информационно-коммуникационные технологии»

НАО «Северо-Казахстанский университет
имени Манаша Козыбаева»

« 16 » _____ 04 _____ 2026 г.



Курмашев И.Г.