УДК [811.411.21'373.21+003.034]:[811.111+811.161.1]

Софья Андреевна Шевцова, студент
Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Россия
***э-почта*: sashevtsova@edu.hse.ru**

Sofia Andreevna Shevtsova, Student
National Research University Higher School of Economics, St. Petersburg, Russia
***e-mail*: sashevtsova@edu.hse.ru**

МОДЕЛИРОВАНИЕ ТРАНСЛИТЕРАЦИИ АРАБСКИХ ТОПОНИМОВ НА АНГЛИЙСКИЙ И РУССКИЙ ЯЗЫКИ

Рассматривается разработка модели для автоматической транслитерации арабских топонимов с учетом особенностей арабской фонетики и морфологии, с использованием сочетания лингвистических и статистических методов для повышения точности перевода географических названий.

Ключевые слова: арабские топонимы; английский язык; русский язык; машинное обучение; автоматическая транслитерация.

MODELING OF ARABIC TOPONYM TRANSLITERATION INTO THE ENGLISH AND RUSSIAN LANGUAGES

The development of a model for automatic transliteration of Arabic toponyms is discussed considering the peculiarities of Arabic phonetics and morphology, using a combination of linguistic and statistical methods to improve the accuracy of geographical name translation.

Key words: Arabic toponyms; the English language; the Russian language; machine learning; automatic transliteration.

По причине глобализации в современном мире растет потребность в точной и корректной передаче географических названий на разных языках, включая транслитерацию арабских топонимов на английский язык. Это особенно важно для эффективного функционирования геоинформационных систем, баз данных, навигационных и картографических приложений, в работе которых точность передачи топонимов играет ключевую роль в обеспечении адекватности данных. В контексте глобализации и интернационализации ошибки в транслитерации могут привести к искажениям в информационных системах, усложняя процессы поиска, идентификации объектов и использования географических данных в аналитических целях.

Сложность задачи заключается в том, что фонетическая и морфологическая структуры арабского языка значительно отличаются от английского, что затрудняет создание универсальных алгоритмов транслитерации, способных правильно передавать арабские топонимы. Существующие решения зачастую не учитывают всех специфических особенностей арабского языка, таких как произношение, диалектные различия и сложность письменных форм, что приводит к возникновению ошибок в передаче и интерпретации названий. Эти ошибки могут существенно повлиять на качество обработки данных, а также на их систематизацию и представление, что в свою очередь усложняет работу с картографическими и другими сервисами.

В связи с этим исследование направлено на разработку модели, которая способна эффективно проводить автоматическую транслитерацию арабских топонимов на английский язык, учитывая языковые закономерности, фонетические особенности и общепринятые стандарты транслитерации. В частности, модель должна быть способна правильно передавать арабские буквы латиницей, а также учитывать диалекты и фонетические особенности различных регионов арабского мира, что повысит точность и надежность результатов.

Для достижения поставленных целей планируется использование сочетания лингвистических и статистических методов. Исходные данные собраны в виде базы топонимов, включающей арабские названия и их соответствующие английские эквиваленты. Будут применены методы обработки и разметки данных для формирования обучающего корпуса, а также методы машинного обучения для создания предсказательной модели. Помимо этого, будет проведен лингвистический анализ, направленный на формулировку правил, отражающих специфические особенности арабских топонимов, включая их морфологическую и фонетическую структуры.

Валидация полученных результатов будет осуществляться с использованием как автоматических методов, так и с привлечением экспертов в области арабского языка, что позволит проверить и подтвердить точность работы модели. Для оценки эффективности работы модели планируется использование метрик, таких как F1-score, расстояние Левенштейна и BLEU-score. После разработки и тестирования модели на исходных данных будет проведено тестирование на независимых наборах данных с целью выявления и анализа возможных ошибок, что поможет в дальнейшем улучшить алгоритмы и повысить их точность.

Таким образом, интеграция машинного обучения с алгоритмами транслитерации представляет собой комплексный подход, который позволяет обеспечить высокую точность при передаче арабских топонимов на английский язык, учитывая как фонетические, так и морфологические особенности арабского языка. Это особенно важно, поскольку арабский язык обладает значительными различиями в произношении и написании в зависимости от региона, и традиционные методы транслитерации часто не способны точно передать все нюансы. Использование методов машинного обучения позволяет моделям адаптироваться к этим особенностям, а также улучшать точность с учетом контекста и особенностей разных диалектов. Одним из ключевых преимуществ такого подхода является его устойчивость к дефициту данных: даже в условиях нехватки информации модель будет способна обрабатывать топонимы, которые не встречаются в обучающем корпусе, благодаря своей способности извлекать закономерности и обрабатывать новые данные, что делает её универсальной и применимой в различных условиях.

Перспективным направлением является расширение применения модели на другие языковые пары, а также интеграция в существующие геоинформационные системы, картографические и навигационные приложения, что позволит улучшить точность и качество обработки арабских топонимов в глобальных базах данных и сервисах.