Отзыв научного руководителя

на диссертационную работу Гукасяна Цолака Гукасовича на тему:

“Методы и программные средства для выявления заимствований в текстах на армянском языке”,

представленную на соискание ученой степени кандидата технических наук по специальности 05.13.11 - “математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”.

Гукасян Ц.Г. (1995 г.р.) окончил бакалавриат отделения программной инженерии НИУ ВШЭ в 2017 году. С 2017 года по 2019 год Гукасян Ц.Г. учился в магистратуре Института математики и информатики Российско-Армянского университета в Ереване по направлению “Машинное обучение”, с 2019 года там же проходит обучение в аспирантуре на кафедре системного программирования.

Студентом 3-го курса бакалавриата Гукасян Ц.Г. пришел в отдел Информационных систем ИСП РАН и с тех пор занимается вопросами разработки методов и программных инструментов для автоматической обработки текстов. Результатом его курсовой работы стало добавление поддержки русской морфологии в одну из наиболее популярных свободных библиотек для обработки естественного языка NLTK. Бакалаврская дипломная работа Гукасяна Ц.Г. была посвящена разработке синтаксического анализатора русского языка, магистерская диссертация - исследованию и разработке инструментов лемматизации армянских текстов. Представленная диссертация также посвящена проблеме из этой области - разработке программных инструментов для обнаружения текстовых заимствований для армянского языка.

Тема диссертационного исследования Гукасяна Ц.Г. имеет высокую актуальность в связи с отсутствием инструментов поиска заимствований, учитывающих специфику армянского языка. При этом университеты и другие организации Армении демонстрируют заинтересованность в проверке уникальности текстовых документов. В результате своей работы Гукасян Ц.Г. впервые создал программные средства поиска заимствований в армянских текстах, которые позволили находить полные и частичные дубликаты, парафраз, попытки технической маскировки.

Современные методы и инструменты для решения поставленной задачи, как правило, основаны на статистических методах, требующих наличие размеченных текстовых коллекций. Для построения точных инструментов желательна экспертная разметка большого числа документов, но такой подход требует много ресурсов. Для разработки инструментов обнаружения парафраза и других вспомогательных методов, необходимо было решить задачу упрощения создания размеченных текстовых коллекций для армянского языка. В работе Гукасяна Ц.Г. предлагаются методы автоматизации процесса разметки, балансирующие точность и объем итоговых коллекций.

При разработке инструментов обработки текстов армянского языка также необходимо было учитывать особенности языка, в частности богатую морфологию, которые усложняют поиск нечетких дубликатов и создают проблему разреженности данных в методах на основе машинного обучения. Для решения этой проблемы в диссертационной работе предложены новые методы векторного представления слов, которые составляют вектор слова исключительно на основе векторов его подслов, тем самым существенно сокращая размер словаря и разреженность данных. Инструменты морфологического анализа и лемматизации, использующие эти новые модели векторного представления, требуют гораздо меньше памяти и показывают качество аналогичное более сложным моделям.

Результаты диссертации имеют существенную практическую значимость с точки зрения борьбы с плагиатом. Разработанная программная система, которая была внедрена в Российско-Армянском университете, позволит более строго отслеживать за качеством студенческих работ с точки зрения их уникальности. Кроме этого, представленные исследования являются существенным вкладом в разработку методов и инструментов автоматической обработки текстов на армянском языке. Методы создания и разметки текстовых датасетов, векторного представления слов могут быть применены для других языков с ограниченными ресурсами, богатой морфологией.

Считаю, что диссертация Гукасяна Ц.Г. носит законченный характер и вносит значимый вклад в область обработки не только армянских текстов, но также других языков с ограниченными ресурсами. Его работа отвечает требованиям “Положения о порядке присуждения ученых степеней”, предъявляемых к кандидатским диссертациям по специальности 05.13.11 - “математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”, а ее автор, Гукасян Цолак Гукасович, заслуживает присуждения ему ученой степени кандидата технических наук.

Кандидат физико-математических наук,
заведующий отделом
Информационных систем ИСП РАН Д.Ю. Турдаков

22 марта 2021 г.