

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 002.087.01
на базе Федерального государственного бюджетного учреждения науки
Институт системного программирования им. В.П. Иванникова
Российской академии наук
Министерства науки и высшего образования РФ
по диссертации на соискание ученой степени кандидата наук

аттестационное дело № _____

решение диссертационного совета от 25 мая 2021 года № 2021/16

О присуждении Гукасяну Цолаку Гукасовичу, гражданину Республики Армения (РА) ученой степени кандидата технических наук.

Диссертация «Методы и программные средства для выявления заимствований в текстах на армянском языке» по специальности 05.13.11 – «математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» принята к защите 25 марта 2021 г., протокол № 2021/06 диссертационным советом Д 002.087.01 на базе Федерального государственного бюджетного учреждения науки Институт системного программирования им. В.П. Иванникова Российской академии наук (ведомственная принадлежность: Министерство науки и высшего образования РФ; адрес: 109004, г. Москва, ул. А. Солженицына, дом 25), создан Приказом Минобрнауки России о советах по защите докторских и кандидатских диссертаций от 2 ноября 2012 г. № 714/нк.

Соискатель Гукасян Цолак Гукасович, 1995 года рождения, работает научным сотрудником в Лаборатории системного программирования Межгосударственного образовательного учреждения высшего образования «Российско-Армянский Университет» (ведомственная принадлежность: Министерство науки и высшего образования РФ и Министерство образования и науки РА).

В 2019 году соискатель окончил Межгосударственное образовательное учреждение высшего образования «Российско-Армянский Университет».

Диссертация выполнена на кафедре системного программирования Института математики и информатики Межгосударственного образовательного учреждения высшего образования «Российско-Армянский Университет» (ведомственная принадлежность: Министерство науки и высшего образования РФ и Министерство образования и науки РА).

Научный руководитель – кандидат физико-математических наук Турдаков Денис Юрьевич, Институт системного программирования им. В.П. Иванникова Российской академии наук, Отдел информационных систем, заведующий отделом.

Официальные оппоненты:

1. Лукашевич Наталья Валентиновна, доктор технических наук (05.13.17), Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова, ведущий научный сотрудник лаборатории анализа информационных ресурсов,
2. Чехович Юрий Викторович, кандидат физико-математических наук (01.01.09), Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», заведующий отделом Вычислительного центра им.А.А.Дородницына Российской академии наук

дали положительные отзывы на диссертацию.

Ведущая организация Федеральное государственное автономное образовательное учреждение высшего образования «Казанский (Приволжский) федеральный университет» (г. Казань) в своем положительном заключении, подписанном Соловьевым Валерием Дмитриевичем (д.ф.-м.н., профессор, кафедра «Интеллектуальные технологии поиска» Института информационных технологий и интеллектуальных систем, кафедра прикладной и экспериментальной лингвистики Института филологии и межкультурной коммуникации) и Зуевым Денисом Сергеевичем (к.т.н., заведующий кафедрой

«Интеллектуальные технологии поиска», Институт информационных технологий и интеллектуальных систем, заместитель директора по научной деятельности), указала, что диссертационная работа содержит новые научные результаты, имеющие существенное значение для науки и практики.

Выбор официальных оппонентов и ведущей организации обосновывается их компетентностью и достижениями в данной отрасли науки, наличием публикаций в сфере исследований, соответствующей теме диссертации, и способностью определить научную и практическую ценность диссертации.

Соискатель имеет 7 работ по теме диссертации, опубликованных в рецензируемых научных изданиях.

В статьях «Subwords-Only Alternatives to fastText for Morphologically Rich Languages», опубликованной в научном журнале «Programming and Computer Software» в 2021 году (входит в список Scopus), и «Векторные модели на основе символьных n-грамм для морфологического анализа текстов», опубликованной в научном журнале «Труды ИСП РАН» в 2020 году (входит в список ВАК), представлено решение проблемы разреженности данных в методах лемматизации и морфологического анализа языков с богатой морфологией (армянский в том числе). Личный вклад соискателя заключается в разработке и реализации новых методов векторного представления слов, планировании экспериментов, участии в их проведении и обработке результатов, написании текста публикации.

В статье «ARPA: Armenian Paraphrase Detection Corpus and Models», опубликованной в сборнике международной конференции «Иванниковские чтения 2020» (входит в список Scopus), представлено решение задачи обнаружения парафраза. Для армянского языка впервые предложен программный инструмент, решающий эту задачу, а также полуавтоматический универсальный метод генерации парафразов предложений, существенно сокращающий необходимость привлечения экспертов в процессе разметки. Соискатель принимал непосредственное участие в разработке программного инструмента и метода генерации, в сборе, подготовке и ручной разметке примеров, планировании и проведении экспериментов.

В статье «Поиск заимствований в армянских текстах путем внутреннего стилометрического анализа», опубликованной в научном журнале «Труды ИСП РАН» в 2021 году, представлены результаты исследования методов по обнаружению подозрительных участков текста путем стилометрического анализа при условии отсутствия проверочного набора потенциальных источников. В рамках этой публикации, при личном участии автора были созданы лингвистические и программные ресурсы описания стиля армянских текстов, проверочные наборы данных для проведения экспериментов по обнаружению качества стилометрических методов поиска заимствований.

В статье «pIoNER: Datasets and Baselines for Armenian Named Entity Recognition», представленной на Открытой конференции ИСП РАН в 2018 году (сборник добавлен в Scopus), описано решение задачи распознавания именованных сущностей для армянского языка, впервые предложены обучающие и тестовые наборы данных, а также автоматический метод их генерации и разметки. Личный вклад соискателя состоит в разработке и реализации модификации метода генерации размеченных примеров, в планировании, проведении и обработке результатов экспериментов, написании текста публикации.

В статье «Post-OCR Correction of Armenian Texts Using Neural Networks», опубликованной в 2020 году, предложено решение задачи исправления ошибок в результате автоматического распознавания армянских текстов, разработан и реализован метод автоматической генерации обучающих и тестовых примеров. Соискатель принимал непосредственное участие в разработке метода создания размеченных примеров, планировании и проведении экспериментов.

Диссертационный совет отмечает, что соискателем получены новые научные результаты:

- представлен новый метод генерации парафразов предложения на основе обратного машинного перевода в несколько итераций и ручной проверки корректности результатов перевода;

- разработан метод автоматической генерации обучающих и тестовых примеров для задачи исправления ошибок оптического распознавания текстов;
- разработаны модификации модели векторного представления слов fastText, которые решают проблему разреженности данных для языков с богатой морфологией и сокращают размер этих моделей, не ухудшая точность и полноту решений задач морфологического анализа и лемматизации;
- разработан новый подход к использованию статей Википедии и Викиданных для полной автоматизации процесса генерации обучающих примеров для задачи распознавания и классификации именованных сущностей.

Значение полученных соискателем результатов исследования для практики состоит в том, что:

- разработана и внедрена система оценки уникальности текстов для армянского языка, которая может быть применена в работе учебных и научных заведений для анализа качества работ учащихся и сотрудников. Программа выполняет поиск заимствований из источников в Интернете, позволяет обнаруживать парафраз;
- разработанные подходы по автоматизации процесса генерации размеченных данных могут быть применены для создания аналогичных наборов для других языков и позволят сократить использование ресурсов;
- впервые для армянского языка разработаны инструменты, позволяющие выполнять внутренний стилометрический анализ текстов на наличие заимствований, исправлять ошибки в результатах оптического распознавания текстов, обнаруживать парафраз;
- впервые для армянского языка созданы тестовые наборы данных с ручной разметкой для задач распознавания именованных сущностей и определения парафраза, которые в дальнейшем могут быть применены для оценки и сравнения качества методов решения этих задач.

Достоверность результатов исследования состоит в том, что:

- принятые в диссертации решения базируются на анализе практики;
- полученные результаты подтверждаются экспериментами;
- использованы представительные выборочные совокупности с обоснованием подбора объектов (единиц) наблюдения и измерения.

Личный вклад соискателя: все представленные на защиту решения получены соискателем лично.

На заседании 25 мая 2021 года диссертационный совет принял решение присудить Гукасяну Цолаку Гукасовичу ученую степень кандидата технических наук.

При проведении голосования диссертационный совет в количестве 18 человек, из них 10 докторов наук по специальности рассматриваемой диссертации, участвовавших в заседании, из 22 человек, входящих в состав совета, проголосовали: за – 18, против – 0, воздержались – 0.

Председатель диссертационного совета,
академик РАН

Аветисян А. И.

Ученый секретарь диссертационного совета,
кандидат физико-математических наук

Зеленов С. В.

25 мая 2021 года