ОТЗЫВ

официального рецензента на диссертационную работу PhD докторанта Мухсиной Куралай Женисбеконы на тему: «Разработка системы анализа многоязычной текстовой информации на основе машинного обучения», представленную на соискание степени доктора философии (PhD) по специальности «6D070400 – Вычислительная техника и программное обеспечение».

1. Актуальность темы исследования и ее связь с общенаучными и общегосударственными программами

Сегодня, с появлением огромных объемов текстов в информационных глобальных и корпоративных сетях, научные исследования, связанные с автоматизацией поиска и анализа текстовой информации выходят на передний план развития информационных технологий. Такие работы исследуют морфологию, синтаксис языка и, в гораздо меньшей степени, обращаются к исследованию семантики. При этом, одной из особенностей данного научного направления является необходимость создания текстовых корпусов и разработки моделей и методов анализа текстов для каждого естественного языка отдельно. Все приведенное выше показывает и подчеркивает актуальность рецензируемой работы, направленной на разработку методов и моделей семантического анализа текстов казахского языка.

Кроме того, созданные в рецензируемой диссертации модели, методы и алгоритмы направленны на обеспечение триединства казахского, русского и английского языков, служащего базой информационного обмена как внутри страны, так и за ее пределами. Разработанная в исследовании система обработки и анализа многоязычной текстовой информации позволяет извлекать структурированную информацию в виде фактов из неструктурированных текстов трех языков, рассматриваемых в Казахстане в качестве государственного языка, языка межнационального общения и языка интеграции в глобальную экономику. Данный подход позволяет существенно повысить качество обработки многоязычных текстов и тем самым получить экономический эффект во многих отраслях, зависящих от передачи и обработки тестовой информации.

Таким образов, решаемые в диссертации задачи Государственной программы по реализации языковой политики в Республике Казахстан на 2020 – 2025 годы от 09 декабря 2020 года по совершенствованию употребления казахского языка в области информатизации и коммуникации делают данную работу актуальной и востребованной.

2. Научные результаты и их обоснованность

Целью работы является исследование и разработка моделей, методов и алгоритмов семантического анализа многоязычной текстовой информации на основе подходов машинного обучения. В рамках поставленной цели решается научная задача моделирования процессов интеллектуальной обработки многоязычных текстов, осуществляющих анализ многоязычной текстовой информации, с целью определения основных характеристик текстов при построении моделей машинного обучения.

В ходе реализации этой цели были получены следующие результаты:

- разработана модель извлечения фактов из слабоструктурированных и неструктурированных текстовых массивов, которая адаптирована для казахского, русского и английского языков;
- обоснован выбор математического аппарата алгебры конечных предикатов для моделирования семантики предложений естественного языка;
- модифицирован метод автоматической морфологической и семантической разметки текстовых корпусов казахского языка, базирующийся на одновременном использовании шаблонов и скрытой Марковской модели;
- разработан метод определения семантической близости текстовых документов на казахском языке к узкоспециализированной предметной области, базирующийся на использовании VSM (Vector Space Model) и весовых коэффициентов PPMI (Positive Pointwise Mutual Information);
- предложена методика экспертной оценки качества работы системы анализа семантической близости текстов;
- создан программный комплекс, определяющий наличие криминального содержания и осуществляющий семантическую разметку текстов казахского, русского и английского языков;
- за счет использования разработанного ПО, в разработке которого реализованы созданные в диссертационном исследовании модели и методы, повышена точность семантической разметки казахских текстов до 71%, русскоязычных текстов до 82% и англоязычных текстов до 87%.

3.Степень обоснованности и достоверности каждого научного результата (научного положения), выводов и заключения, сформулированных в диссертации

Обоснованность и достоверность каждого научного результата (научного положения), выводов и заключения основывается на глубоком анализе, корректном использовании теории интеллекта, общей теории систем, системного анализа, алгебры конечных предикатов и методов машинного обучения. Решение каждой поставленной задачи опирается на результаты, полу-

ченные на предыдущих этапах исследований, что подтверждает их взаимозависимость и внутреннее единство результатов. Обоснованность и достоверность подтверждается также внедрением разработанных рекомендаций.

Очередность выполнения основных этапов работы, начиная с анализа состояния вопроса, последовательный переход к решению поставленных задач, логика изложения направлены на создание автоматизированной системы, определяющей наличие криминального содержания и осуществляющей семантическую разметку текстов казахского, русского и английского языков.

4.Степень новизны каждого научного результата (научного положения), выводов и заключения, сформулированных в диссертации

Наиболее существенные результаты диссертации Мухсиной К. Ж., обладающие высокой степенью научной новизны, включают:

- Логико-лингвистическую модель семантического анализа, идентифицирующую факты в многоязыковых текстах, которая позволяет извлекать из текстов казахского, русского и английского языков знания, явным образом представленные в виде RDF-триплетов, и формировать семантически размеченные обучающие корпуса.

- Метод автоматической морфологической и семантической разметки текстовых корпусов казахского языка, отличительной особенностью которого является одновременное использование модели HMM (Hidden Markov Model) и правил, представленных регулярными выражениями.

- Метод определения семантической близости текстовых документов казахского языка к узкоспециализированной предметной области.

- Методику экспертной оценки качества анализа семантической близости текстов, базирующуюся на вычислении среднего и минимального значения косинусного сходства.

- Программный комплекс, определяющий наличие криминального содержания и осуществляющий семантическую разметку текстов казахского, русского и английского языков.

5. Практическая и теоретическая значимость научных результатов, направленных на решение актуальной проблемы, теоретической и прикладной задачи.

Диссертация Мухсиной К. Ж. является квалификационной научной работой, содержащей научно обоснованные результаты, направленные на создание и внедрение современных систем автоматической обработки текстовой информации на основе машинного обучения. Использование предложенных в работе моделей, методов и алгоритмов позволяет существенно по-

высить эффективность семантической обработки текстов казахского, русского и английского языков.

6. Соблюдение в диссертации принципа самостоятельности

Соблюдение в диссертации принципа самостоятельности подтверждается 17 публикациями автора, среди которых 2 статьи и 5 конференций, индексированных в базах данных Web of science и Scopus; 4 статьи в изданиях, рекомендованных ККСОН МОН РК и материалы 6 международных конференций, в том числе 1-й зарубежной. На разработанную программу получено свидетельство авторского права. Индекс Хирша Мухсиной К. Ж. равен 3.

7. Соответствие аннотации содержанию диссертации

Содержание аннотации соответствует содержанию диссертации, отражает все четыре главы, заключение, научные положения, научную новизну и практическую значимость.

8. Замечания, предложения по диссертации

1. Возможно, следовало рассмотреть, как будет работать предложенная логико-лингвистическая модель извлечения фактов из текстовых массивов на документах, одновременно содержащих фрагментны на казахском, русском и английском языках.

2. В подразделе 4.1 диссертации рассмотрен разработанный параллельный корпус криминально-окрашенных текстов казахского и русского языков. Следовало рассмотреть существующие методы использования параллельных корпусов для извлечения информации из многоязычных текстов.

3. Следовало более подробно описать используемый способ вычисления коэффициента agreement в подразделе 4.3.

Изложенные замечания не снижают ценности представленной работы и носят характер пожеланий.

9. Заключение о возможности присуждения соискателю степени доктора философии (PhD) по специальности 6D070400 – «Вычислительная техника и программное обеспечение»

Не смотря на сделанные замечания считаю, что диссертационная работа Мухсиной К. Ж. на тему «Разработка системы анализа многоязычной текстовой информации на основе машинного обучения», представленная на соискание степени доктора PhD по специальности 6D070400 – «Вычислительная техника и программное обеспечение», выполнена на высоком научном уровне, представляет собой законченную научно-исследовательскую работу

и удовлетворяет требованиям, предъявляемым к докторским диссертациям на соискание степени доктора PhD.

Рекомендую Мусхину Куралай Женисбековну к присуждению ей степени доктора PhD по специальности 6D070400 – «Вычислительная техника и программное обеспечение».

Официальный рецензент:
д.т.н., ректор
«Astana IT University» **Ахмед- Заки Д.Ж.**