АННОТАЦИЯ
диссертационной работы Касекеевой Айслу Бисеновны
«Модель семантической библиотеки, содержащей информацию на английском, русском и казахском языках», представленной на соискание степени доктора философии (PhD) по направлению подготовки кадров «8D061 – Информационно-коммуникационные технологии»: специальность «6D070300 – Информационные системы»

Цифровая грамотность - основа безопасности в информационном обществе, одно из важнейших знаний XXI века, одна из наших основных тем сегодняшнего общества. Цифровая грамотность - это готовность и способность надёжно и эффективно использовать цифровые технологии во всех сферах жизни человека. Использование этой технологии позволяет улучшить качество жизни населения.

В Послании от 31 января 2017 года «Третье возрождение Казахстана: глобальная конкурентоспособность» Президент Нурсултан Назарбаев подчеркнул необходимость разработки государственной программы «Цифровой Казахстан»: «... Мы должны развивать новые отрасли, созданные с использованием цифровых технологий. Это важная комплексная задача. Стране необходимо развивать такие направления, как 3D-печать, интернет-магазины, мобильный банкинг, цифровые услуги, здравоохранение, образование и другие перспективные направления. Эти отрасли уже провели реструктуризацию экономики развитых стран и придали новое качество традиционным отраслям. В связи с этим поручаю Правительству разработать и принять индивидуальную программу «Цифровой Казахстан». ...Развитие цифровой индустрии даст толчок всем остальным направлениям. Поэтому Правительству следует уделять особое внимание развитию ИТ-индустрии».

Каждая страна выбирает определённые факторы, влияющие на экономическое развитие и интеграцию в мировое пространство. Республика Казахстан одним из главных факторов ставит образование, которое отвечает требованиям мировой экономики, повышает жизненный уровень и общее благополучие населения. Правительства многих стран одной из основных задач считают повышение конкурентоспособности экономики за счёт развития качества образовательных услуг. Это связано с тем, что уровень образования общества и научного потенциала представляет собой важное условие экономического роста. В условиях современных глобализационных процессов формируется международная система высшего образования, которая представляет собой совокупность национальных взаимосвязанных систем.

Образование является одним из важнейших критериев Стратегии «Казахстан - 2030». Основной целью образовательных программ является адаптация образовательной системы к новой экономической среде. Президентом Республики Казахстан была поставлена задача о вхождении республики в число 50-ти конкурентоспособных стран мира. Также, согласно Закону РК «О национальной безопасности Республики Казахстан», одной из

угроз является ухудшение качества образования и интеллектуального потенциала страны, что доказывает огромное значение создания и совершенствования качественной системы образования.

До недавнего времени цифровые библиотеки воспринимались обычными пользователями как электронные версии каталогов традиционных библиотек, которые содержат описания физических объектов библиотеки (как правило, книг или других печатных изданий). Определение тематики, содержания и структуры объектов рассматриваются и воспринимались как дополнительные, но необязательные функции таких библиотек. Развитие интернета и семантических технологий вносит свои коррективы и позволяет шире взглянуть на понятие цифровых библиотек и обобщить накопившийся опыт реализации информационных систем в разных областях знаний для формирования нового типа библиотек.

Само понятие библиотеки в контексте стремительного развития интернета приобретает совершенно другой смысл и обозначает активное вовлечение пользователя в процессы, предлагаемые библиотеками. Такая библиотека предполагает участие пользователей в процессе создания, поиска и классификации того контента библиотеки, который необходим этому конкретному пользователю.

Актуальность темы исследования. В связи с быстрым ростом объёма текстовой информации исследования в области компьютерной лингвистики на естественном языке остаются актуальными. На сегодняшний день количество информации, которую люди и машины воспроизводят на естественном языке, очень сильно возросло. Создание алгоритмов и создание систем сбора данных, модель семантических библиотек, содержащих информацию на разных языках, классификация и кластеризация библиотек с текстовыми документами по-прежнему являются сложными задачами.

Многие исследователи склоняются к необходимости проведения глубокого семантического анализа текстов для создания их семантических образов, на основе которых можно было бы проводить тонкое ранжирование документов. Этот подход, несомненно, наиболее разумный, однако требует тщательной и долгой работы над созданием подходящих инструментов для автоматической обработки текстов. В частности, может потребоваться детальное описание различных областей знаний. Поэтому также имеет смысл поиск частичных решений, одно из которых представлено в данной работе.

Непрерывное увеличение интенсивности потока текстовой информации делает все более важной задачу семантической модели библиотеки, содержащую информацию на разных языках.

Семантика – раздел лингвистики, изучающий смысловое значение единиц языка: отдельных слов, словосочетаний, предложений, фрагментов текста. В настоящее время существует несколько машинно-ориентированных методов представления значений операторов.

Например, И.А. Мельчук ввёл понятие лексической функции, развил понятие синтаксической и семантической валентности и рассмотрел в рамках пояснительно-комбинаторного словаря. В.Ш. Рубашкин и Д.Г. Лахути ввели

иерархию синтаксических связей для эффективного функционирования семантического анализатора.

Непрерывное увеличение интенсивности потока текстовой информации делает все более важной задачу семантической модели библиотеки, содержащую информацию на разных языках.

Теоретической основой послужили научные работы, содержащие исследования по грамматикам связей, синтаксическим анализаторам текстов на естественном языке, методам сравнения предложений и определения тем текстов, алгоритмам на графах и математической логике.

Цель диссертационного исследования и научные результаты.

Целью диссертационного исследования является разработка метода и модели семантической библиотеки, содержащей информацию на английском, русском и казахском языках, на основе исследований и результатов, полученных при анализе пространственно-временных конструкций в естественном языке, и разработанного приложения Semantics.

Объект исследования является семантическая библиотека для поддержки научно-образовательной деятельности.

Предмет исследования – модели, алгоритмы и методы семантических библиотек, а также методы классификации и кластеризации ресурсов в ЭБ.

Научная новизна. Наибольший научный интерес представляют следующие вопросы: Предложена классификация пространственно-временных конструкций в предложениях казахского, русского и английского языков. Предложен метод и алгоритм семантической эквивалентности предложений, содержащих пространственно-временные конструкции, и простых предложений казахского, русского и английского языков. Разработаны модель и прототип семантической библиотеки с использованием семантически эквивалентных предложений казахского, русского и английского языков.

Реализован алгоритм эквивалентности пространственно-временных отношений с помощью анализаторов Link Grammar Parser и Диалинг, серверная часть PHP, клиентская часть HTML, CSS, базы данных MySQL. Создана информационная система семантической библиотеки, содержащей информацию на английском, русском и казахском языках, с учётом модели формальных представлений.

Задачи исследования. В соответствии с поставленной целью в диссертационной работе решаются следующие задачи:

1. Провести анализ и классификацию пространственно-временных конструкций;
2. Создать математическую модель множества формальных представлений предложений на языке эквивалентности;
3. Разработать информационную систему семантической библиотеки Semantics.

Положения, выносимые на защиту:

1. Формальная классификация пространственно-временных конструкций.

2. Математическая модель множества формальных представлений на языке эквивалентности.

3. Программная реализация информационной системы семантической библиотеки Semantics.

Практическое значение полученных результатов. Предложенные модели, методы в виде прототипа интеллектуальной системы могут быть использованы в интеллектуальных системах поиска информации. Итоги исследования также представляет интерес для учёных лингвистов.

Практическая значимость основных положений данной диссертации подтверждена результатами использования разработанных моделей, методов и алгоритмов для разработки семантической библиотеки в Институте систем информатики им. А.П.Ершова Сибирского отделения Российской академии наук.

Личный вклад соискателя. Модели и методы семантической библиотекой, содержащей информацию на английском, русском и казахском языках были предложены, описаны и разработаны автором лично. Анализ и выборка наиболее важных понятий, относящихся пространственно-временным отношениям, из толкового словаря С.И. Ожегова и других источников и их переводы, выявления пространственно-временные формы в документах для использования в семантической библиотеке на трех языках разрабатывалась лично автором. Создание массива перефразированных вариантов различных предложений и метод оценивания их похожести тоже личный вклад автора. Экспериментальная оценка алгоритма и модели проводилась совместно с научным консультантом и Институтом систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук.

Основные положения, выносимые на защиту, являются персональным вкладом автора в опубликованные работы.

Апробация результатов диссертации. Обоснованность полученных в диссертационной работе результатов основана на использовании апробированных методов исследования, корректном применении математического аппарата теории графов, теории множеств и векторной алгебры, статистических методов обработки данных, согласовании полученных результатов с известными теоретическими положениями в области обработки текстов на естественном языке и интеллектуальной поддержки принятия решений.

Адекватность предложенных методов и алгоритмов подтверждается результатами по реализованным интеллектуальным методам на представленном текстовом корпусе, а также результатами апробации и актом внедрения прототипа интеллектуальной системы. Достоверность полученных результатов обеспечивается применением известных методов и подходов к классификации текстов, корректной статистической обработкой данных.

Основные положения и результаты работы докладывались и обсуждались на следующих научных конференциях:

- «Математикалық және компьютерлік модельдеудің заманауи мәселелері Қазақстанның цифрлы индустриясының дамуы жағдайында» Республикалық ғылыми-практикалық конференция тезистер жинағы, (3-5 мамыр, 2018 ж, Астана);

- «Ғылым, білім және өндіріс интеграциясы - Ұлт жоспарын іске асырудың негізі» (№10 Сағынов оқулары) Халықаралық ғылыми-практикалық конференциясының Е Ң Б Е К Т Е Р І (14-15 июня 2018 г. 2 часть, Қарағанда 2018);

- «Advances in Science and Technology» XIX Международная научно-практическая конференция 15 марта 2019 Научно-издательский центр «Актуальность. РФ», (15 март, 2019 г, г. Москва);

- BIG DATA и анализ высокого уровня, Сборник материалов V международной конференции. (13-14 марта, 2019 . Минск, Белоруссия);

- Шестая международная научно-практическая конференция BIG DATA and Advanced Analytics BIG DATA и анализ высокого уровня (20-21 Мая, 2020 Минск, Беларусь);

- Сборник материалов XV Международной научной конференции студентов и молодых ученых «ǴYLYM JÁNE BILIM - 2020», РК, (10 апреля, 2020г., г. Нур-Султан).

Результаты диссертационной работы внедрены в Институте систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, что подтверждается приведенным в приложении актам о внедрения научных положений и разработок диссертации в практику деятельности образовательных организаций.

Публикаций. По результатам диссертационного исследования, опубликовано 13 печатных работ, в том числе 4 работы в рецензируемых печатных изданиях, рекомендованных ККСОН, 1 работа в изданиях, индексируемых в Scopus, и 8 публикаций в других научных журналах и сборниках трудов конференций. Получено 1 авторское свидетельство о внесении сведений в государственный реестр прав на объекты, охраняемые авторским правом программы для ЭВМ.

Объем и структура диссертации. Диссертационная работа состоит из введения, трех разделов, заключения и списка литературы. Объем работы составляет 115 страниц, включая 22 рисунка, 3 таблиц, 3 приложения. Список литературы содержит 107 наименований.

Во введении представлен научный аппарат исследования, обосновывается актуальность темы, степень ее разработанности в теории и практике, определены цель, объект, предмет и задачи исследования, раскрывается научная новизна, теоретическая и практическая значимость работы, определены методы исследования, представлены положения, выносимые на защиту, личный вклад автора, список публикации и апробаций результатов работы.

Раздел 1 включает в себя анализ современного состояния логико-философских исследований семантических библиотек. Приводится анализ методов информационного поиска, которые являются активно

развивающимися, актуальными в научно-практическом аспекте на сегодняшний день. Рассмотрены следующие модели информационного поиска: булева модель, векторная модель, вероятностная модель. Выявлены недостатки некоторых моделей. Определяются ключевые проблемы сопоставления современных используемых семантических библиотек.

В разделе 2 описываются формальный анализ перефразированных предложений естественного языка анализаторами Link Grammar Parser и Диалинг для анализа словарных статьей и примеров использования в художественной литературе, относящихся пространственно-временным отношениям.

Раздел 3 посвящена описанию модели и алгоритмов семантической библиотеки, содержащей информацию на английском, русском и казахском языках, а также программной реализации анализа диаграмм предложений. Описывается интерфейс системы для формирования модели семантической библиотеки и базы данных по классификации пространственно-временных конструкций трех языков. А также приведен математическая модель семантической библиотеки, содержащий информацию на трех языках.

В заключении обобщены результаты исследования, сформулированы основные выводы, подтверждающие и доказывающие истинность положений, выносимых на защиту.

В приложении представлены практические материалы исследования.

Автор выражает благодарность своему научному руководителю д.ф.-м.н., профессору кафедры «Информационных систем» Тусупову Джамалбек Алиаскаровичу за постановку задачи и помощь в ходе диссертационного исследования.

Автор признателен своему зарубежному консультанту к.ф.-м.н., доценту заместителю директора по научной работе Института систем информатики им. А.П. Ершова СО РАН Мурзину Федор Александровичу за консультации в ходе исследований.

Автор также выражает благодарность за консультацию д.филол.н, профессору кафедры общего языкознания и теории перевода Тажибаевой Сауле Жаксылыкбаевне.