275/06
16.06.2022

ОТЗЫВ

об автореферате диссертации Зиновьевой Анастасии Юрьевны «МОДЕЛЬ МНОГОЯЗЫЧНОГО ИНТЕЛЛЕКТУАЛЬНОГО КОНТЕНТ-АНАЛИЗА (на материале англо-, франко- и русскоязычных новостных сообщений о террористической деятельности)», представленной на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика

Исследование А. Ю. Зиновьевой посвящено развитию важного направления автоматизированного анализа текста и дискурса – контент-анализа, в данном случае – моделированию интеллектуального контент-анализа (ИКА). Более узкой областью методологии, разработанной автором, является ИКА неструктурированной текстовой информации на примере новостных сообщений предметной области (ПО) «Терроризм» на английском, французском и русском языках. Проблема ИКА в настоящее время является недостаточно разработанной в русле лингвистических дисциплин, если учесть небольшое количество языков (прежде всего английский и французский), охваченных соответствующими разработками. Тем актуальнее данное исследование, в центре внимания которого находится многоязычная модель интеллектуального контент-анализа, которая может быть использована повторно для обработки текстов на различных языках с минимизацией времени, затрат и усилий разработчиков. Важно, что данная модель охватывает и русский язык, интеллектуальному контент-анализу которого до настоящего времени не уделяется достаточно внимания.

В качестве объекта диссертации рассматриваются подъязык англо-, франко- и русскоязычных новостных сообщений предметной области «Терроризм»; предметом исследования являются моделирование концептуальной структуры рассматриваемого подъязыка и формализация извлечения проблемно-ориентированного контента из текстов предметной области.

Целью диссертационного исследования является разработка модели многоязычного интеллектуального контент-анализа на основе онтологических знаний на примере англо-, франко- и русскоязычных новостных сообщений предметной области «Терроризм» с возможностью дальнейшего ее применения к другим языкам и предметным областям. Среди важных задач, успешно решенных в исследовании – построение онтологической базы знаний модели многоязычного интеллектуального контент-анализа на основе данных анализа подъязыка новостных сообщений предметной области «Терроризм» и разработка алгоритма интеллектуального контент-анализа указанного подъязыка и предметной области. Это достигается, в том числе, выявлением ограничений и закономерностей указанного подъязыка и предметной области на уровне структуры релевантности и суперструктуры, морфосинтаксическом и лексико-семантическом уровнях на материале английского, французского и русского языков.

Новизна и оригинальность данного диссертационного исследования определяется инновационной методикой исследования, а также более глобальной методологией, включающей создание онтолексиконов для английского, французского и русского языков; разработку правил онтологического анализа, логического вывода и формирования динамических концептуально-лексических фреймов для представления результатов контент-анализа; а также разработку алгоритма ИКА, который позволяет извлекать не только явно выраженную, но и имплицитную информацию.

Достоверность и научная корректность полученных автором результатов обусловлена солидной теоретической базой исследования; эффективным использованием современных методов обработки языкового материала (моделирования, целевой и случайной выборки для создания исходных и тестовых корпусов, статистического и сопоставительного анализа, дистрибутивного анализа, метода оппозиций и т. д.); опорой на обширный эмпирический материал (три псевдопараллельных корпуса новостных сообщений о терроризме за 2014–2020 гг. равного объема общим объемом более 600 000 словоупотреблений, для доработки модели – корпусы текстов за 2019–2020 гг. объемом 20–40 тыс. с. у.).

Теоретическая значимость исследования состоит в развитии методологических аспектов многоязычного ИКА; результаты исследования способны внести вклад и в смежные направления, в том числе, компьютерную и корпусную лингвистику, информационный поиск, автоматическое реферирование и аннотирование, машинный перевод. Практическая ценность результатов работы заключается в возможности создания системы многоязычного ИКА на базе разработанной модели, лежащие в ее основе принципы могут быть экстраполированы на другие языки и предметные области.

Значимыми представляются ряд теоретических выводов и практических результатов диссертационного исследования. Основным ресурсом модели интеллектуального контент-анализа является онтология. При этом в рамках направления, в котором онтологии трактуются как зависимые от языка ресурсы, предлагаются достаточно трудоемкие процедуры: разработки инструментов для полуавтоматического создания онтологий на разных языках или методик объединения онтологий, изначально созданных для разных языков. Однако для решения многих практических задач, в том числе, моделирования многоязычного ИКА онтология должна быть независимой от конкретного естественного языка и многоязычной (т. е. ориентированной на обработку нескольких языков).

Интересна разработанная трехэтапная методика обработки материала: анализ структуры релевантности и суперструктуры каждого новостного сообщения; морфосинтаксический анализ, автоматическое экстрагирование с помощью экстрактора LanaKey лексических групп от 1 до 4 компонентов (программное ограничение экстрактора), извлечение релевантных для предметной области лексических групп, лемматизация, подсчет частоты встречаемости лексических групп и т.д.; лексико-семантический анализ и распределение выделенных лексических единиц по концептуальным классам, релевантным для предметной

области, на основании общих сем с помощью прескриптивно-дескриптивной методики. На основании этой процедуры выявлены концепты, причем ряд лексических единиц предметной области обнаруживают свойства концептуального синкретизма и концептуальной неоднозначности (поэтому отнесены более чем к одному концептуальному классу).

Наконец, важен разработанный алгоритм ИКА, который включает ряд последовательных процедур по отношению к каждому тексту собранного корпуса: онтологический анализ, разметка текста тегами концептов на основе онтолексиконов, снятие концептуальной неоднозначности; определение фрагмента онтологии, содержащего релевантные для поставленной задачи концепты (и их теги); логический вывода, если релевантная информация не представлена в тексте эксплицитно; процедура извлечения информации и заполнения концептуально-лексического фрейма, формируемого динамически на основе вводных данных с помощью разработанного автором экстрактора; числовая обработка данных в заполненных фреймах; представление результатов исследования в форме таблиц или графиков.

Ценно применение и развитие платформы концептуального аннотирования, состоящей из модуля сбора и хранения знаний и концептуального теггера, для обработки текстов. Платформа может быть использована для полной автоматизации формальной концептуальной разметки и частичной автоматизации снятия концептуальной неоднозначности.

В целом диссертационное исследование А. Ю. Зиновьевой представляется значимым вкладом в моделирование интеллектуального контент-анализа и имеет широкие перспективы использования на практике в разных подъязыках и предметных областях. Оно выполнено на высоком научном уровне, соответствует требованиям п. 9 Положения о присуждении научных степеней, а ее автор заслуживает присвоения ему ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика.

Доктор филологических наук Шелестюк Елена Владимировна
по специальности 10.02.19 (Теория языка)
доцент, профессор кафедры теоретического
и прикладного языкознания
ФГБОУ ВО «Челябинский государственный университет»
07.06.2022

Место работы: ФГБОУ ВО «ЧелГУ»
454001, г. Челябинск, ул. Братьев Кашириных, 129.
тел.: +79634745702
e-mail: shelestiuk@yandex.ru

Подпись Елены Владимировны Шелестюк заверяю

Специалист по кадрам
В.И.Акутина