258/06
04.06.2022

ОТЗЫВ

официального оппонента кандидата филологических наук, доцента
Захарова Виктора Павловича о 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

Общая картина современного информационного пространства позволяет утверждать, что сегодня информация оценивается как наиболее дорогостоящий ресурс. Существенная и наиболее востребованная ее часть представляет собой текстовую информацию на естественном языке, циркулирующую в сети. Соответственно, с каждым годом возрастают роль и значение систем автоматизированной обработки текстовой информации, в которых все большее место занимают методы и средства интеллектуальной обработки. Среди всего многообразия форм автоматизированной обработки текстов особое место занимают процессы, связанные с извлечением смысла и смысловой компрессией. В этом русле и выполнена представленная нам диссертационная работа, а именно, автор исследует и разрабатывает основные компоненты формальной модели многоязычного интеллектуального контент-анализа на основе онтологических знаний на примере англо-, франко- и русскоязычных новостных сообщений предметной области «Терроризм».

При оценке **актуальности диссертационного исследования** можно рассматривать два аспекта: актуальность самой области исследования в целом — что в нашем случае, как уже было сказано, жизненно важно — и актуальность конкретной проблемы, решаемой в диссертации. Актуальность специальной проблемы, рассматриваемой в диссертационной работе, а именно, разработка модели интеллектуального контент-анализа также не вызывает сомнений. Можно утверждать, что, несмотря на множество исследований в этой области, диссертация А.Ю. Зиновьевой занимает свое особое место. Это место определяется тем, что извлечение информации из текстов новостных сообщений рассматривается как интеллектуальная, смысловая задача, решаемая на основе формальных онтологий. Особый аспект актуальности поставленной задачи придает ориентация на многоязычность. Выбранная предметная область (терроризм) еще раз подчеркивает актуальность, буквально, злободневность этой работы.

Особенностями рассматриваемого диссертационного исследования, на мой взгляд, являются:

- глубокий всесторонний анализ проблем моделирования интеллектуального контент-анализа, базирующийся на изучении широкого круга источников (175 наим.);
- подробный анализ понятия «интеллектуальный контент-анализ» и помещение его в поле информационных технологий;
- методика разработки модели многоязычного интеллектуального контент-анализа на основе многоязычного корпуса текстов предметной области;
- разработка единой корпусно-ориентированной методики анализа подъязыка многоязычной предметной области;
- разработка собственно модели интеллектуального контент-анализа;
- формализация лингвистических и экспертных знаний на основе формальных вычислительных онтологий;
- создание онтологической базы знаний предметной области как основного ресурса модели интеллектуального контент-анализа;
- использование независимой от конкретного языка онтологии MikroKosmos как основы базы знаний предметной области «Терроризм»;
- разработка алгоритма многоязычного интеллектуального контент-анализа;
- совмещение в разработанной модели лингвистических, логических и статистических методов.

Научная новизна диссертации А.Ю. Зиновьевой состоит в комплексном подходе к решению задачи, включающем методику разработки модели многоязычного интеллектуального контент-анализа, методику анализа подъязыка многоязычной предметной области, разработку модели и алгоритма интеллектуального контент-анализа на основе онтологической базы знаний предметной области. Выбранные подходы и разработанные методы и инструменты носят оригинальный характер. Обладают признаками научной новизны разработанные диссертанткой правила онтологического анализа, логического вывода и формирования динамических концептуально-лексических фреймов для представления результатов контент-анализа. Можно сказать, что диссертация содержит достаточно детальный анализ проблем и методов интеллектуального контент-анализа, что само по себе представляет вклад в развитие этих аспектов прикладной лингвистики. Все выше отмеченное позволяет сделать вывод о том, что диссертантом получены **новые результаты** при исследовании сложных теоретических и практических проблем.

Обоснованность научных положений, выносимых на защиту, вытекает как из результатов, изложенных во второй и третьей главе, так и из публикаций автора, посвященных моделированию интеллектуального контент-анализа. Результаты экспериментальной проверки подтверждают эффективность разработанной модели. Все научные положения, выводы и рекомендации, сделанные в диссертации, обоснованы и аргументированы.

Достоверность результатов обусловлена широким и качественным анализом исследуемой проблематики, аргументированностью научных положений и выводов, представленных в каждой главе и в заключении диссертации, экспериментальной проверкой разработанной модели интеллектуального контент-анализа, давшей положительные результаты.

Личный вклад соискателя виден по всему исследованию, начиная со сбора и анализа материала, который достаточно репрезентативен, и заканчивая формулированием базовых принципов интеллектуального контент-анализа и разработкой многоуровневой алгоритмической модели. Там, где описывается платформа концептуального аннотирования (ПАнТ) для обработки текстов предметной области «Терроризм», работа над которой велась в коллективе, там же особо оговаривается личный вклад автора.

Практическая ценность исследования заключается в том, что разработана алгоритмическая модель интеллектуального контент-анализа, которая прошла экспериментальную проверку и показала свою эффективность, и что научно-теоретические положения и алгоритмы, разработанные А.Ю. Зиновьевой, можно использовать для разных предметных областей и разных языков. Кроме того, материалы диссертационного исследования могут быть использованы в учебном процессе учреждений высшего и дополнительного профессионального образования в рамках курсов по автоматическому анализу текста, извлечению знаний, построению онтологий. Так, вся глава 3 и Приложения Б–Е — это, можно сказать, готовое учебное пособие.

Работа имеет четкую структуру, написана ясным научным языком и хорошо оформлена.

Тема и содержание диссертации и полученные результаты соответствуют области исследования и паспорту научной специальности 10.02.21 – прикладная и математическая лингвистика (филологические науки). Диссертация представляет собой законченное исследование. Публикации автора выполнены на высоком уровне.

Автореферат и опубликованные работы отражают основное содержание диссертации.

Таким образом, диссертация А.Ю. Зиновьевой представляет собой завершенное исследование, имеющее как новизну, так и теоретическую и практическую ценность.

Конечно, не всё в исследовании А.Ю. Зиновьевой одинаково хорошо и бесспорно. В частности, считаю необходимым высказать **ряд замечаний**.

1. Понятийный аппарат исследования. В главе 1 автор последовательно рассматривает смысловое наполнение основных аспектов своего исследования, представленных в заглавии и являющихся "краеугольными камнями" диссертации: *модель, контент-анализ, интеллектуальный контент-анализ, онтология*. Однако, как говорится, наши недостатки – продолжение наших достоинств. В частности, меня не удовлетворил раздел 1.1. «Модель и моделирование в лингвистических исследованиях». Обзор литературы по теме далеко не полон как по существу, так и по набору "знаковых" авторов, писавших о лингвистических моделях. Второе и главное, многофасетная по сути своей классификация моделей дается практически на одном уровне, где типы моделей из разных фасетов (по разным основаниям классификации) даются перечислением, так сказать, "через запятую" (лингвистическая модель, языковая, речевой деятельности, лингвистического исследования, многоязычная модель, компьютерная и др.). Тем не менее, в конце раздела А.Ю. Зиновьева четко определяет задачи моделирования применительно к своей работе (с. 16), но потом пишет: «Настоящее исследование направлено на разработку многоязычной прикладной компьютерной модели, воспроизводящей *исследовательскую процедуру* (курсив мой — В.З.) интеллектуального контент-анализа». В результате, мое любопытство, какой же все-таки тип моделирования по Апресяну (*модель речевой деятельности* или *модель лингвистического исследования*) автор реализует в своей диссертации, осталось неудовлетворенным.

2. Ресурсы интеллектуального контент-анализа (подраздел 1.3.3). С терминологической точки зрения мне не кажется удачной дихотомия "статические ресурсы – динамические ресурсы". Позволю себе процитировать диссертацию: "Под статическими понимаются неизменяемые во времени ресурсы, содержащие данные в той или иной форме; к ним могут быть отнесены корпусы текстов, словари (лексиконы), онтологии, списки терминов и т. п. Динамические ресурсы, напротив, представляют собой инструменты, обеспечивающие создание новых данных и последующую их обработку, а именно, инструменты разметки текста, морфологические анализаторы, парсеры, инструменты для извлечения лексики и т. п." (с. 29-30). Я вижу здесь опять же пренебрежение "фасетностью": первый тип - "неизменяемые во времени ресурсы", а второй, динамические ресурсы – получается, изменяемые во времени?

Отнюдь нет! Второй тип — как мы прочли, это инструменты. В этой непоследовательности можно, конечно, винить Уитта [Witt и др., 2009], на которого ссылается автор диссертации, но все же далее Анастасия Юрьевна, можно сказать, настаивает на такой классификации, причем сама же отмечает: "В ряде случаев провести границу между статическими и динамическими ресурсами может быть затруднительно". Наверное, можно было бы придумать другие термины (лингвистические vs программно-алгоритмические ресурсы, декларативные vs процедурные ресурсы и т. п.), но это уже не моя задача.

3. Замечание, которое относится к понятию "концептуальный класс", точнее, к **процедурам формирования (наполнения) концептуальных классов**. В современной лингвистике для этого успешно применяются методы корпусной лингвистики. Похоже, ими же в какие-то моменты пользуется и диссертантка. «Далее извлеченные из корпусов лексические единицы были разнесены по концептуальным классам с использованием компонентного и контекстного анализа (последний выполнялся с помощью программы [Anthony, 2020])», – читаем мы на стр. 65. Anthony, 2020 — это как раз корпусный менеджер AntConc. Но в диссертации описание данной технологии, увы, отсутствует. Более того, в процедурах, описанных в разделах "Результаты анализа русского, английского, французского корпусов" (разд. 2.3–2.5) формирование дистрибутивно-статистических тезаурусов (по-другому, семантических, или концептуальных полей) методом дистрибутивное-статистического анализа оказалось бы, на мой взгляд, крайне продуктивным. Однако, судя по тексту диссертации, этого не делалось. Цитирую: «Отметим, что отнесение лексической единицы к концептуальным классам во всех случаях выполнялось на основании контекста того предложения, в котором встречалась данная единица» (с. 71). То есть, осуществлялся интеллектуальный контекстный анализ.

4. Частное замечание – неудачное использование слова "**интерпретация**" в терминологическом значении в подразделе 1.3.3. Цитата: "Первичные ресурсы в свою очередь представляют собой ресурсы, которые были подвергнуты минимальной обработке: так, если при распознавании некачественной копии документа расшифровщик предпочел одно слово другому, он уже осуществил *интерпретацию*, приняв во внимание контекст. Наконец, обогащенные ресурсы — это первичные ресурсы, к которым вручную или автоматически с помощью инструментов разметки текста были добавлены некоторые *интерпретации*" (с. 30–31). Во-первых, мы видим, что здесь "интерпретация" имеет разные значения. Во-вторых, для второго употребления имеется стандартный термин "метаданные", который автор хорошо знает, но почему-то не использует.

Однако сделанные замечания не влияют на высокую оценку представленного диссертационного исследования.

Заключение. Диссертационное исследование А.Ю. Зиновьевой является самостоятельной и завершенной научной работой. Проведенное автором научное исследование и его результаты содержат решение актуальной научно-методической и научно-технической задачи построения модели многоязычного интеллектуального контент-анализа. Проведенный анализ работы позволяет сделать вывод, что диссертация Зиновьевой Анастасии Юрьевны по теме «Модель многоязычного интеллектуального контент-анализа (на материале англо-, франко- и русскоязычных новостных сообщений о террористической деятельности)», представленная на соискание ученой степени кандидата филологических наук по научной специальности 10.02.21 – прикладная и математическая лингвистика (филологические науки), в полной мере соответствует требованиям п. 9 «Положения о присуждении ученых степеней», утвержденного постановлением Правительства Российской Федерации № 842 от 24 сентября 2013 г., предъявляемым к кандидатским диссертациям, а автор исследования, Зиновьева Анастасия Юрьевна, заслуживает присуждения ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика.

Кандидат филологических наук, доцент,
доцент кафедры математической лингвистики
Санкт-Петербургского государственного
университета В.П. Захаров

3 июня 2022 г.

Контактная информация:
199034 Санкт-Петербург
Университетская наб., 7-9
Тел. раб. +7-812-328-95-19
Тел. моб. +7-911-937-17-69
E-mail: v.zakharov@spbu.ru

ПОДПИСЬ РУКИ Захарова В. П.

УДОСТОВЕРЯЮ

ЗАМЕСТИТЕЛЬ НАЧАЛЬНИКА
УПРАВЛЕНИЯ КАДРОВ ГУОРП

ХОМУТСКАЯ Л. П.

03.06.2022