УДК 004.82 + 81’32 + 811.1 + 323.28 DOI: 10.14529/ling210208

МОДЕЛИРОВАНИЕ МНОГОЯЗЫЧНОГО ИНТЕЛЛЕКТУАЛЬНОГО КОНТЕНТ-АНАЛИЗА

С.О. Шереметьева, А.Ю. Зиновьева
Южно-Уральский государственный университет, г.Челябинск, Россия

В настоящей статье представлен опыт разработки модели интеллектуального контент-анализа – необходимого ресурса компьютерных технологий обработки неструктурированной информации. Отличительной чертой модели является возможность ее применения для анализа текстов на различных национальных языках и механизм извлечения соответствующего задаче анализа контента, не представленного эксплицитно на поверхностном уровне текста. Модель состоит из двух основных компонентов: базы знаний, включающей ориентированную на предметную область многоязычную онтологию, онтолексиконы, динамические фреймы, а также правила обработки текстов и представления результатов контент-анализа. Методология разработки многоязычной модели и собственно процедуры контент-анализа описаны на примере их применения к корпусу новостных сообщений предметной области «Терроризм» на английском языке.

Ключевые слова: интеллектуальный контент-анализ, моделирование, многоязычность, онтология, терроризм.

Введение

Контент-анализ – это широко используемый в настоящее время метод анализа содержания текста в соответствии с определенной информационной задачей. Этот метод универсален и в настоящее время особенно интенсивно применяется для анализа е-информации [2]. Дифференциация видов контент-анализа обычно сводится к его классификации как количественного или качественного методов исследования. Количественный контент-анализ подразумевает подсчет слов и словосочетаний в искомом тексте. Качественный, или интеллектуальный, контент-анализ основан на категоризации текстовых единиц. Эта методологическая дихотомия качественного и количественного исследования часто подвергается обоснованной критике. Например, авторы работ [3, 6] определяют качественный (интеллектуальный) контент-анализ как подход смешанных методов (содержащий как качественные, так и количественные этапы анализа) и выступают за применение к ним общих исследовательских критериев. В частности, отмечается, что интеллектуальный контент-анализ как качественный шаг включает присвоение категорий текстовым элементам (кодирование), а обработку множества текстов и анализ частот категорий (кодов) – как количественный шаг. Выделение категорий кодирования при создании моделей интеллектуального контент-анализа имеет решающее значение, является основным и наиболее проблематичным этапом. В исследовательских техниках категориального кодирования текста в свою очередь выделяется три различных подхода: корпусный, прескриптивно-корпусный и суммативный. При корпусном подходе категории кодирования выводятся непосредственно из текстовых данных. При прескриптивно-корпусном подходе выделение категорий кодирования основывается на теоретических положениях или релевантных результатах предыдущих исследований. Суммативный подход к категориальной кодировке принципиально отличается от двух предыдущих. Вместо того чтобы анализировать данные в целом, текст кодируется в основном ключевыми относительно конкретного содержания словами [5]. В рамках любого из указанных подходов интеллектуальный контент-анализ достаточно часто производится на материале конкретного языка с ориентацией на конкретную информационную задачу и при изменении национального языка материала контент-анализа и (или) его задач даже в рамках одной предметной области требует нового набора категорий, что, по сути дела, исключает повторное использование ресурсов категоризации и отрицательно сказывается на трудозатратности, оперативности и совместимости моделей.

С появлением общедоступного Интернета и популяризацией Семантической паутины наиболее перспективным инструментом решения проблем категориальной кодировки текстов становится онтологический анализ [4], который может обеспечить категоризацию контента информации на одном или нескольких языках [10]. При этом в связи с отсутствием возможности построения всеобъемлющей онтологии практически каждая из основанных на онтоанализе моделей интеллектуального контент-анализа использует предметно-обусловленную онтологию [7].

В настоящей статье описывается основанная на многоязычном онтологическом ресурсе пошаговая модель интеллектуального контент-анализа текстов предметной области «Терроризм», реализованная на материале русского, английского и французского языков. Статья организована сле-

дующим образом. В разделе 1 представлена методология разработки модели. А в разделе 2 описана многоязычная онтологическая база знаний. В разделе 3 приведен алгоритмический компонент модели. В разделе 4 применение разработанной модели иллюстрируется на конкретном примере. В заключении рассмотрены итоги и дальнейшие этапы исследования.

1. Методология разработки модели

Под интеллектуальным контент-анализом (ИКА) предметной области (ПО) мы пониманием извлечение из корпусов текстов определенной ПО соответствующего информационному запросу контента, его интерпретацию и представление в удобной для пользователя форме. В качестве основного метода достижения сформулированной задачи является выявление релевантной концептуальной структуры (концептуальное кодирование) текстов, которая в отличие от общесемантической структуры, отражающей универсальные семантические признаки лексики («конкретный», «одушевленный», и т. д.), сигнализирует о принадлежности текста к конкретной предметной области.

В настоящем исследовании определение концептуальной структуры (т. е. категоризация и кодирование) текста представляет собой процедуру онтологического анализа с использованием специально построенной предметно-ориентированной многоязычной онтологии. Многоязычие в онтологиях понимается в двух основных смыслах: 1) как адаптация (или понятность) названий (labels) онтологических концептов для пользователей – носителей различных национальных языков или 2) как возможность применения одной онтологии к обработке текстов на различных языках независимо от того, какой язык используется для обозначения названий концептов. В настоящем исследовании мы придерживаемся второй трактовки, что определяет как архитектуру модели, которая состоит из двух основных компонентов: лексико-онтологической базы знаний и алгоритмических процедур интеллектуального контент-анализа (рис. 1), так и методологию ее построения. На первом этапе создания модели определяется общий подход к контент-анализу и строится проблемно-ориентированная база знаний, а на втором апробируется и финализируется алгоритмический компонент модели.

2. База знаний

База знаний – это основной, наиболее трудоемкий и проблематичный для создания компонент модели ИКА, который содержит:

- многоязычную онтологию предметной области «Терроризм»,
- онтолексиконы предметно-релевантных лексических единиц русского, английского и французского языков,
- правила онтологического анализа,
- правила логического вывода,
- правила формирования результатов работы модели ИКА,
- концептуально-текстовые фреймы.

Построение, организация и содержание предметно-ориентированной многоязычной онтологии, фрагмент которой представлен на рис. 2, основаны на корпусном анализе русских, английских и французских текстов предметной области и подробно описаны в работе [8].

Второй компонент базы знаний модели ИКА – предметно-релевантные одноязычные лексиконы одно- и многокомпонентных лексических единиц, соотнесенных с концептами онтологии. При этом между лексическими единицами каждого из языков и концептами онтологии имеют место следующие отношения: «один к одному», «один ко

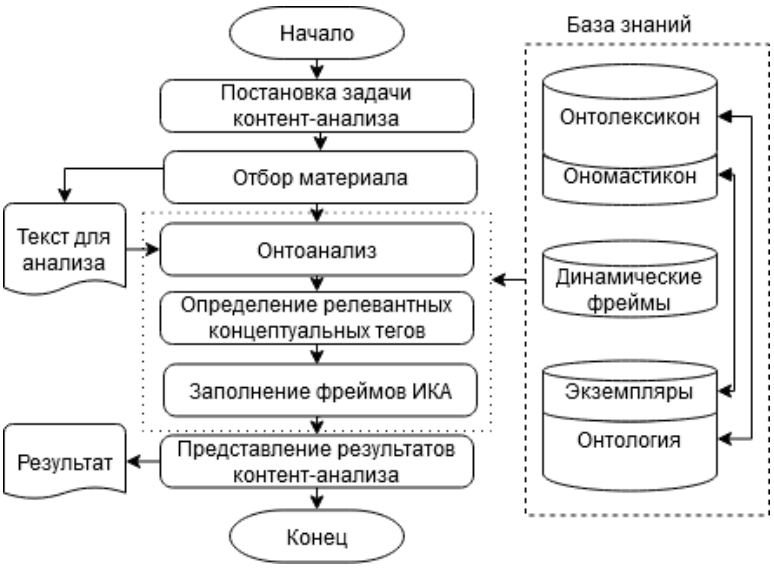

Рис. 1. Архитектура модели интеллектуального контент-анализа

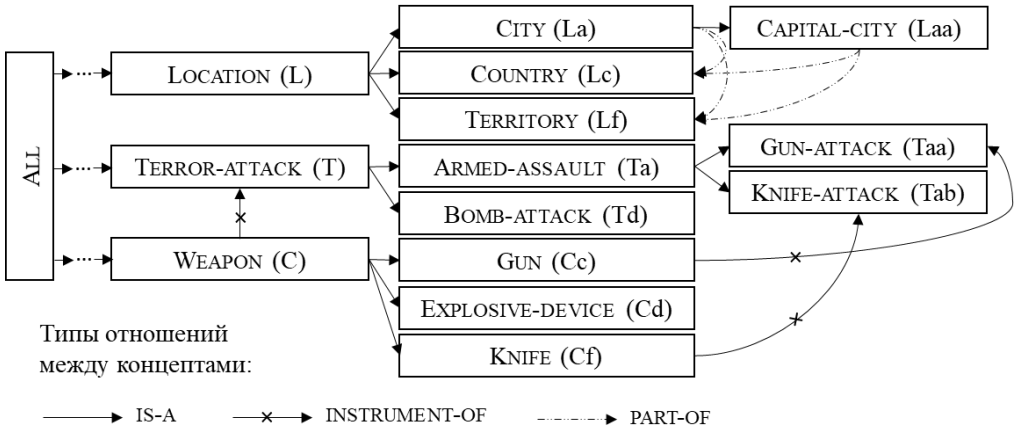

Рис. 2. Фрагмент многоязычной онтологии предметной области «Терроризм»
(для наименования концептов использован английский язык; в скобках указаны концептуальные коды (теги))

многим», «многие к одному» и «многие ко многим», что объясняется разной концептуальной природой лексических единиц, которые могут быть концептуально однозначными, т. е. связанными с одним концептом онтологии, а также концептуально неоднозначными и концептуально синкретичными, что определяет их отражение на несколько концептов онтологии. *Концептуально неоднозначными* являются лексические единицы, которые в зависимости от контекста могут иметь различные *противоречащие друг другу* концептуальные значения и в каждом конкретном случае *реализуют только одно* из этих значений. *Концептуально синкретичными* являются единицы, одновременно реализующие *несколько не противоречащих* друг другу концептуальных значений. Например, англоязычное словосочетание *Iranian victim* сочетает в себе концептуальные значения ‘национальность’ и ‘последствия теракта’ и поэтому отражена на два концепта онтологии NATION и CONSEQUENCES-FOR-PEOPLE, однако кодирование этой единицы различными концептуальными тегами, в отличие от случаев концептуальной неоднозначности, не требует устранения ни одного из этих тегов. Подробный анализ явлений концептуальной неоднозначности и синкретичности дан в работе [1]. Кроме концептуальной информации единицы онтолексиконов снабжены морфосинтаксическими признаками.

В онтолексиконе базы знаний ИКА выделен особый блок «Ономастикон», содержащий именованные сущности, связанные с экземплярами онтологических концептов. Этот ресурс необходим для систематизации именованных сущностей, поскольку многие экземпляры имеют несколько вариантов именования. Например, экземпляр концепта TERRORIST-ORGANIZATION с наименованием TERRORIST-ORGANIZATION-3 в одноязычных лексиконах связан с единицами *ТАК*, *Ястребы свободы Курдистана*, *Соколы свободы Курдистана*, *Kurdistan Freedom Hawks*, *Kurdistan Freedom Falcons*, *Faucons de la liberté du Kurdistan*, которые являются синонимами.

Правила онтологического анализа включают в себя разметку (кодирование) текстов кодами (тегами) онтологических концептов и разрешение концептуальной неоднозначности (подробнее см. [1]).

Правила логического вывода, основанные на онтологических знаниях, и правила презентации результатов работы модели ИКА описаны в разделе 4.

Концептуально-лексические фреймы представляют собой динамически формируемые шаблоны для заполнения текстовой информацией, релевантной категориям задачи ИКА.

3. Алгоритмический компонент модели

Алгоритм модели ИКА, использующий базу знаний, описанную в разделе 2, представлен в центральной части блок-схемы архитектуры модели ИКА на рис. 1. Он включает в себя: а) этап определения задачи ИКА и б) процедуры ее решения (на вход каждой из них подается информация, полученная на выходе предыдущей процедуры алгоритма), выполняемые либо автоматически (на основе авторских программ), либо вручную.

После формулировки задачи ИКА первой выполняется процедура создания релевантного корпуса текстов на одном или нескольких языках в зависимости от поставленной задачи. Корпус создается методом критериальной выборки через соответствующий интернет-агрегатор по ключевым словам. Ключевыми считаются слова, входящие в формулировку задачи ИКА. Для обеспечения репрезентативности выборки выданные агрегатором тексты фильтруются по следующим критериям:

1. Текст опубликован в открытом источнике, находящемся в свободном доступе и не требующем платной подписки.

2. Текст не является републикацией текста, уже отобранного в корпус, т. е. одному событию должно соответствовать одно новостное сообщение.

3. Текст содержит информацию о нескольких релевантных запросу событиях, но отдельных сообщений о каждом из событий не обнаружено.

Следующие процедуры алгоритма модели ИКА применяются последовательно к каждому отдельному тексту результирующего корпуса.

Сначала выполняется онтоанализ, который состоит из: а) подпроцедуры разметки (кодирования) текстов кодами (тегами) концептов онтологии на основе онтолексиконов, выполняемой автоматически с помощью платформы концептуального аннотирования [9], б) подпроцедуры разрешения концептуальной неоднозначности, осуществляемой вручную. Коды концептуально синкретичных лексем разрешения не требуют.

Следующая процедура определяет набор релевантных задаче ИКА тегов и включает в себя а) подпроцедуру определения фрагмента-дерева онтологии, содержащего релевантные для ИКА концепты (теги); б) подпроцедуру логического вывода, которая актуализируется в том случае, если в каком-либо тексте корпуса отсутствуют лексические единицы, размеченные релевантными для поставленной задачи ИКА концептуальными тегами, т. е. релевантная информация не представлена на текстовом уровне эксплицитно. Подпроцедура логического вывода основана на анализе связей онтологических концептов и на выходе выдает дополнительные релевантные теги (и размеченные ими лексические единицы) анализируемого текста. Например, концепт GUN («дочка» концепта WEAPON, INSTRUMENT-OF TERROR-ATTACK) связан отношением INSTRUMENT-OF с концептом GUN-ATTACK *(стрельба)* и может быть использован для получения информации о типе теракта, если в тексте такая информация отсутствует.

После определения в тексте релевантных меток (кодов) актуализируется процедура формирования концептуально-лексического фрейма со слотами, соответствующими выделенным в тексте релевантным тегам (концептам), которые заполняются размеченными данными тегами фрагментами текста. Эта процедура выполняется автоматически с помощью разработанного А.Ю. Зиновьевой экстрактора.

Затем информация в заполненных фреймах обсчитывается и выдается пользователю. Представление результатов ИКА на настоящем этапе исследования осуществляется в виде таблиц или графиков (см. рис. 4, 5).

4. Пример применения модели ИКА

Постановка задачи: сравнить количество терактов в странах Евразии, освещенных в англоязычных СМИ, и способы осуществления этих терактов.

Отбор материала: с учетом критериев отбора через агрегатор «Google Новости» по ключевым словам *terror attack, terrorist attack, act of terrorism* с ограничением по датам публикации 01.01.2019–31.03.2019 и 01.01.2020–31.03.2020 собрано два корпуса англоязычных сообщений о терактах.

Онтоанализ: в качестве примера на рис. 3 показан один из текстов на выходе процедуры онтоанализа, включающей кодирование и разрешение концептуальной неоднозначности. Напомним, что коды (теги), свидетельствующие о концептуальной синкретичности лексических единиц, не устраняются.

Определение релевантных концептуальных тегов: в соответствии со сформулированной выше задачей ИКА требуется выделить фрагмент онтологии с вершинами LOCATION и TERROR-ATTACK, а теги узлов этого дерева считать релевантными (см. рис. 2). В размеченном тексте примера на рис. 3 отсутствуют лексические единицы, эксплицитно репрезентующие концепты COUNTRY и (или) TERRITORY (т. е. в размеченном тексте нет тегов Lc и Lf), однако присутствуют единицы, репрезентующие концепт CITY (La), который отношением PART-OF связан с концептами COUNTRY и (или) TERRITORY. Это предложные группы *in Kiryat Arba* ‘в городе Кирьят-Арба’ и *near Hebron* ‘недалеко от Хеврона’, включенные в онтологию как экземпляры CITY-102 и CITY-86 соответственно и связанные отношением PART-OF с экземпляром TERRITORY-2

{Attempted stabbing terror attack}~Tab~K **{near Hebron}~La-86,** {terrorist}~Pb~A {killed}~Pba~Rb

{An}~DEF **{attempted stabbing terror attack}~Tab**~K {occurred}~NC **{near Hebron}~La-86** {on Monday}~B {as}~NC {the}~DEF {terrorist}~Pb~A {was shot}~Pba~Rb {by}~NC {IDF forces}~Ra, {according to}~S {IDF Spokesperson’s Unit}~S.

{According to}~S {the}~DEF {report}~D, {the}~DEF {terrorist}~A {attempted}~K {to}~NC **{stab}~Tab** {soldiers}~Zab {stationed}~UNK {at}~NC {an}~DEF **{IDF post}~Le {in Kiryat Arba}~La-102.** {The}~DEF {force}~Ra {shot}~Rb {and}~NC {killed}~Pba~Rb {the}~DEF {terrorist}~Pb~A {as}~NC {he}~NC {entered}~NC {a}~DEF **{building}~Le** {close by}~NC.

{No}~DEF {Israelis}~Pa~N {were injured}~Pae {in}~NC {the}~DEF {**attempted attack**}**~T**~K.

Рис. 3. Один из текстов корпуса на выходе процедуры онтоанализа
(жирным выделены лексические единицы, кодированные релевантными тегами (концептами)

(West Bank, ‘Западный берег реки Иордан’*)*. Следовательно, на основании аксиомы «Если некий теракт был совершен в некоем городе и этот город является частью некой территории, то этот теракт был совершен на этой территории» можно заключить, что территорией, на которой был совершен описываемый теракт, является Западный берег реки Иордан.

Формирование концептуально-лексического фрейма: фрейм, сформированный на основе автоматической экстракции релевантной информации из текста на рис. 3, имеет следующий вид (результаты логического вывода выделены жирным шрифтом):

Текст ID:	20190312JP
TERRITORY:	**West Bank (TERRITORY-2)**
CITY:	in Kiryat Arba (CITY-102), near Hebron (CITY-86)
SPECIFIC LOCATION:	IDF post, building
TERROR ATTACK:	attempted attack
KNIFE-ATTACK:	attempted stabbing terror attack, stab

Представление результатов ИКА: результаты представлены графически средствами MS Excel. На рис. 4 показаны картограммы распределения терактов по странам и территориям Евразии в первом квартале 2019 и 2020 гг. соответственно. На рис. 5 представлена сравнительная информация о типах терактов, совершенных в первом квартале 2019 и 2020 гг.

Заключение

В настоящей статье представлены методология разработки и функционал модели интеллектуального контент-анализа многоязычной неструктурированной информации. Модель основана на лексико-онтологической базе знаний, построенной на основе корпусного анализа новостных сообщений о террористических атаках на русском, французском и английском языках. Алгоритмический компонент модели включает процедуры онтологического анализа, извлечения из текста релевантной задаче контент-анализа информации, представле-

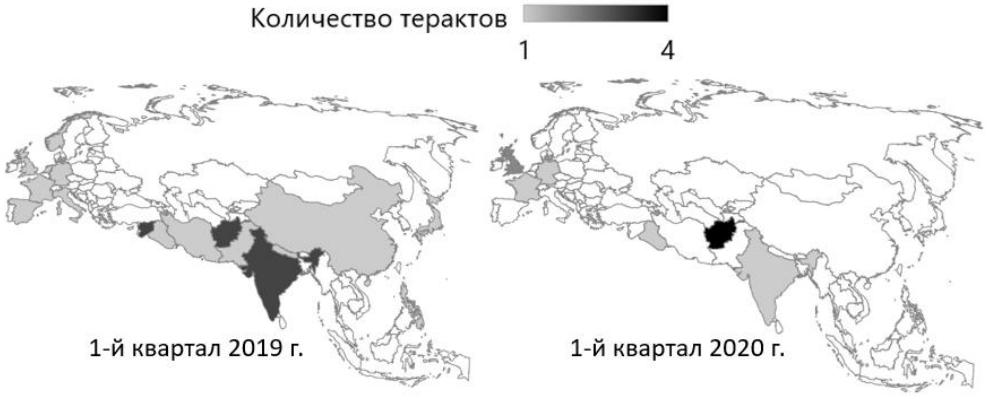

Рис. 4. Распределение терактов в странах Евразии в первых кварталах 2019 и 2020 гг.

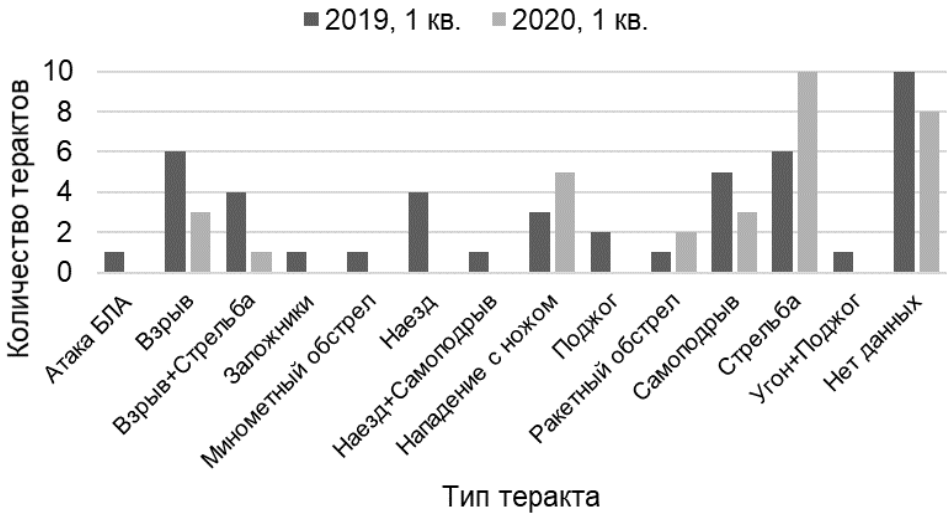

Рис. 5. Сравнение частоты различных типов терактов в первых кварталах 2019 и 2020 гг.

ния ее в виде фреймов с последующей визуализацией в виде картограмм или графиков. Дальнейшие этапы исследования будут направлены на увеличение покрываемости модели путем ее экстраполяции на другие языки и предметные области.

Литература

1. Зиновьева А.Ю. Анализ неоднозначности концептуальной разметки русскоязычного текста / А.Ю. Зиновьева, С.О. Шереметьева, Е.Д. Неручева // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. – 2020. – Т. 6, № 3 (23). – С. 38–60.

2. Погорецкий, В.Г. Контент-анализ – методические основания исследования в электронных СМИ / В.Г. Погорецкий // Труды ИСА РАН. – 2006. – Т. 26. – С. 95–111.

3. Gauch, H.G. Scientific Method in Practice / H.G. Gauch. – Cambridge University Press, 2002. – 456 p.

4. Green, P.S. The Practice of Ontological Analysis / P.S. Green, M. Rosemann, M. Undulska. – 2005. – https://pdfs.semanticscholar.org/513c/a04a8132a723cf47d9d9504983a98dd 9ec08.pdf.

5. Hsieh, H.-F. Three Approaches to Qualitative Content Analysis / H.-F. Hsieh, S.E. Shannon // Qualitative Health Research. – 2005. – Vol. 15 (9). – P. 1277–1288.

6. Mayring, Ph. Qualitative content analysis: theoretical foundation, basic procedures and software solution / Ph. Mayring. – Klagenfurt, 2014. – 144 p.

7. Nirenburg, S. Ontological Semantics / S. Nirenburg, V. Raskin. – Cambridge: MIT Press, 2004. – 440 p.

8. Sheremetyeva, S. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content / S. Sheremetyeva, A. Zinoveva // Communications in Computer and Information Science. – Springer, Cham, 2018. – Vol. 859. – P. 368–379.

9. Sheremetyeva, S. Towards creating interoperable resources for conceptual annotation of multilingual domain corpora / S. Sheremetyeva // The Proceedings of the 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16), Marseille, 2020. – P. 102–109.

10. The NEWS ontology: Design and applications / N. Fernández, D. Fuentes, L. Sánchez, J.A. Fisteus // Expert Systems with Applications. – 2010. – Vol/ 37 (12). – P. 8694–8704.

Шереметьева Светлана Олеговна, доктор филологических наук, доцент, профессор кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), sheremetevaso@susu.ru

Зиновьева Анастасия Юрьевна, аспирант кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), zinovevaaiu@bk.ru

Поступила в редакцию 29 декабря 2020 г.

DOI: 10.14529/ling210208

ON MODELING MULTILINGUAL INTELLIGENT CONTENT ANALYSIS

S.O. Sheremetyeva, *sheremetevaso@susu.ru*
A.Yu. Zinoveva, *zinovevaaiu@bk.ru*
South Ural State University, Chelyabinsk, Russian Federation

This article presents an experience of developing a model for intelligent content analysis, which is a necessary resource for computer technologies of processing unstructured information. A distinctive feature of the model is the possibility of its application for the analysis of texts in various national languages and the mechanism for extracting task-oriented content that is not explicitly presented at the surface level of the analyzed text. The model consists of two main components: first, a knowledge base that includes a domain-oriented multilingual ontology, ontolexicons, dynamic frames, and, second, rules for text processing and presentation of content analysis results. The methodology for developing a multilingual model and the actual procedure of content analysis are described in regard to their application to the corpus of news reports on terrorist attacks in English.

Keywords: intelligent content analysis, modeling, multilingualism, ontology, terrorism.

References

1. Zinoveva A.Yu., Sheremetyeva S.O., Nerucheva E.D. The Analysis of Ambiguity in Conceptual Annotation of Russian Texts. *Tyumen State University Herald. Humanities Research. Humanitates*, 2020, vol. 6, no. 3 (23), pp. 38–60. DOI: 10.21684/2411-197X-2020-6-3-38-60. (in Russ.)
2. Pogoretsky, V.G. *Kontent-analiz – metodicheskie osnovaniya issledovaniya v elektronnykh SMI* [Content analysis – methodological foundations of research in electronic mass media]. *Trudy ISA RAN [The Proceeding of ISA RAS]*. 2006, 26, pp. 95–111. (in Russ.)
3. Gauch H.G. *Scientific Method in Practice*. Cambridge University Press, 2002. 456 p.
4. Green P.S., Rosemann M., Indulska M. *The Practice of Ontological Analysis*. 2005. Available at: https://pdfs.semanticscholar.org/513c/a04a8132a7 23cf47d9d9504983a98dd9ec08.pdf.
5. Hsieh H.-F., Shannon S.E. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*. 2005, 15 (9), pp. 1277–1288.
6. Mayring Ph. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Klagenfurt, 2014. 144 p.
7. Nirenburg S., Raskin V. *Ontological Semantics*. Cambridge: MIT Press, 2004. 440 p.
8. Sheremetyeva S., Zinovyeva A. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. *Communications in Computer and Information Science*. 2018, 859, pp. 368–379.
9. Sheremetyeva S. Towards creating interoperable resources for conceptual annotation of multilingual domain corpora. *Proceedings of the 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation* (ISA-16), Marseille, 2020, pp. 102–109.
10. Fernández N., Fuentes D., Sánchez L., Fisteus J.A. The NEWS ontology: Design and applications. *Expert Systems with Applications.* 2010, 37 (12), pp. 8694–8704. DOI: 10.1016/j.eswa.2010.06.055.

Svetlana O. Sheremetyeva, PhD (Habilitation), professor of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), sheremetevaso@susu.ru

Anastasia Yu. Zinoveva, post-graduate student of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), zinovevaaiu@bk.ru

Received 29 December 2020

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Шереметьева, С.О. Моделирование многоязычного интеллектуального контент-анализа / С.О. Шереметьева, А.Ю. Зиновьева // Вестник ЮУрГУ. Серия «Лингвистика». – 2021. – Т. 18, № 2. – С. 52–58. DOI: 10.14529/ling210208

FOR CITATION

Sheremetyeva S.O., Zinoveva A.Yu. On Modeling Multilingual Intelligent Content Analysis. *Bulletin of the South Ural State University. Ser. Linguistics*. 2021, vol. 18, no. 2, pp. 52–58. (in Russ.). DOI: 10.14529/ling210208