УДК 811.111 + 004.822 + 004.738.5 + 323.28 DOI: 10.14529/ling200106

АНАЛИЗ АНГЛОЯЗЫЧНЫХ ИНТЕРНЕТ-СООБЩЕНИЙ О ТЕРРОРИСТИЧЕСКИХ АКТАХ НА ОСНОВЕ МНОГОЯЗЫЧНОЙ ОНТОЛОГИИ

С.О. Шереметьева, О.И. Бабина, А.Ю. Зиновьева, Е.Д. Неручева
Южно-Уральский государственный университет, г. Челябинск, Россия

В статье исследуются возможности онтологического анализа электронных текстовых потоков на различных языках на основе независимого от конкретного языка предметно-ориентированного онтологического ресурса. Методология исследования сочетает дескриптивный и прескриптивный подходы и включает несколько итеративных этапов: сбор, анализ и формализацию текстовых данных с последующим выявлением их концептуальной структуры, что, в свою очередь, используется как для определения и совершенствования контента онтологии, так и для уточнения процедуры онтологического анализа конкретных одноязычных текстов. Процедура анализа описана на примере его применения к англоязычным интернет-сообщениям о террористических актах. Особое внимание уделено важной проблеме онтологического анализа – разрешению концептуальной многозначности при установлении соответствий между лексическим и онтологическим уровнями представления информации.

Ключевые слова: онтологический анализ, многоязычная предметная онтология, интернет-сообщения, терроризм, английский язык.

Введение

С появлением общедоступного интернета анализ онлайн-новостей, содержащих информацию из тысяч различных источников, составляет неотъемлемую часть решения многих практических задач, среди которых борьба с терроризмом является одной из основных и предъявляет особенно высокие требования к оперативности и качеству обработки текстовых интернет-потоков. Большая часть существующих компьютеризированных методов классификации, поиска, извлечения и интерпретации данных/знаний, применяемых для обработки неструктурированной информации, куда относятся новостные сообщения о террористических актах, в основном осуществляется на основе текст-майнинга или парсинга без глубокого лингвистического анализа (компьютерного «понимания» смысла текста) из-за сложности последнего, что, как правило, не обеспечивает высококачественных результатов. При этом признается, что наиболее перспективным инструментом повышения корректности обработки информации является онтологический анализ [5]. С популяризацией Семантической паутины (Semantic Web) [1], предназначенной обеспечить операционную совместимость данных на семантическом уровне, количество работ в этой области, в частности, работ по онтологическому анализу электронных новостей, резко возросло. Фокус онтологических исследований и разработок варьирует от лингвистических и методологических вопросов до инструментов и реальных баз знаний, которые, как правило, строятся для конкретных приложений и учитывают ограничения предметной области. Например, система ePaper [12] использует собственную онтологию в качестве общего языка для персонализированной фильтрации электронных новостей на основе контента; в системе NEWS [4] онтологические знания обеспечивают классификацию по контенту на трех языках: английском, испанском и итальянском. Онтология PiT (Profiles in Terror) [7] предназначена для представления знаний о террористической сети; она включает в себя перечень отдельных лиц, организаций и связей между ними. Онтология, описанная в работе [8], создается для прогнозирования террористических атак на основе данных о террористических организациях, их намерениях и оружии. В [6] описан онтологический анализ, основанный на предметной онтологии по терроризму для извлечения информации о террористических событиях из электронных новостей. В связи со сложностью соотнесения поверхностного и концептуального уровней структуры текста практически каждое из проводимых в настоящее время онтологических исследований предназначено для решения частных практических задач и до сих пор не существует ни универсальных ресурсов, ни методик универсального онтологического анализа.

В настоящей статье представлен опыт разработки модели онтологического анализа на примере предметной области «Терроризм» и применения модели для анализа англоязычных интернет-новостей о террористических актах. Статья организована следующим образом: в разделе 1 описывается методология исследования; в разделе 2 представлен многоязычный онтологический ресурс предметной области «Терроризм»; в разделе 3 рассмотрена проблема концептуальной многозначности и представлены результаты онтологического

анализа англоязычных интернет-новостей о террористических актах. В заключении рассмотрены итоги и перспективы проведенного исследования.

1. Методология исследования

Под онтологическим анализом понимается изучение контента, или, точнее, процесс извлечения знаний о сущностях, включенных в определенную предметную область [3]. На практике онтологический анализ заключается в отображении лексических единиц текста на концепты онтологии с последующей формализацией и интерпретацией результатов такого отображения в зависимости от конкретной задачи. При этом исследователям, разрабатывающим техники онтологического анализа, приходится сталкиваться с серьезными ограничениями. Первое ограничение связано с доступностью подходящей и хорошо разработанной онтологии. Несмотря на то, что довольно большое количество онтологических библиотек можно найти онлайн [2], их применимость в каждом конкретном исследовательском проекте, как правило, проблематична. Поэтому в большинстве работ по онтологическому анализу первая задача (которая часто бывает и конечной целью исследования) состоит в создании предметной онтологии и/или онтологии, настроенной на решение конкретной прикладной задачи. Второе основное ограничение заключается в возможностях практической реализации онтологического анализа как такового. Проблемными являются как четкое определение границ анализа, так и ограниченность представленных в онтологии наборов концептов и взаимосвязей между ними. Между единицами текста и онтологическими концептами могут существовать отношения типа «один-к-нулю» (для текстовых единиц не находится отображения на концепты онтологии), «один-ко-многим», «многие-к-одному» или «многие-ко-многим», что ведет к концептуальной многозначности. Мы решаем эти проблемы с применением корпусного анализа.

Методология нашего исследования основана на эмпирических данных, использует сочетание дескриптивного и прескриптивного подходов и включает несколько итеративных этапов: сбор, анализ и формализацию текстовых данных с последующим выявлением их концептуальной структуры, что, в свою очередь, используется как для определения и совершенствования контента онтологии, так и для процедуры онтологического анализа текстов соответствующей предметной области. На рис. 1 показана дорожная карта проведенного исследования, которое основано на следующих постулатах:

- Онтология – это независимый от конкретного языка ресурс, допускающий использование для анализа информации на различных языках.
- Для онтологического анализа неструктурированной информации (текстов на естественном языке) определенной предметной области должна быть построена соответствующая предметная онтология.
- Специфичные для предметной области знания являются неотъемлемой частью общих

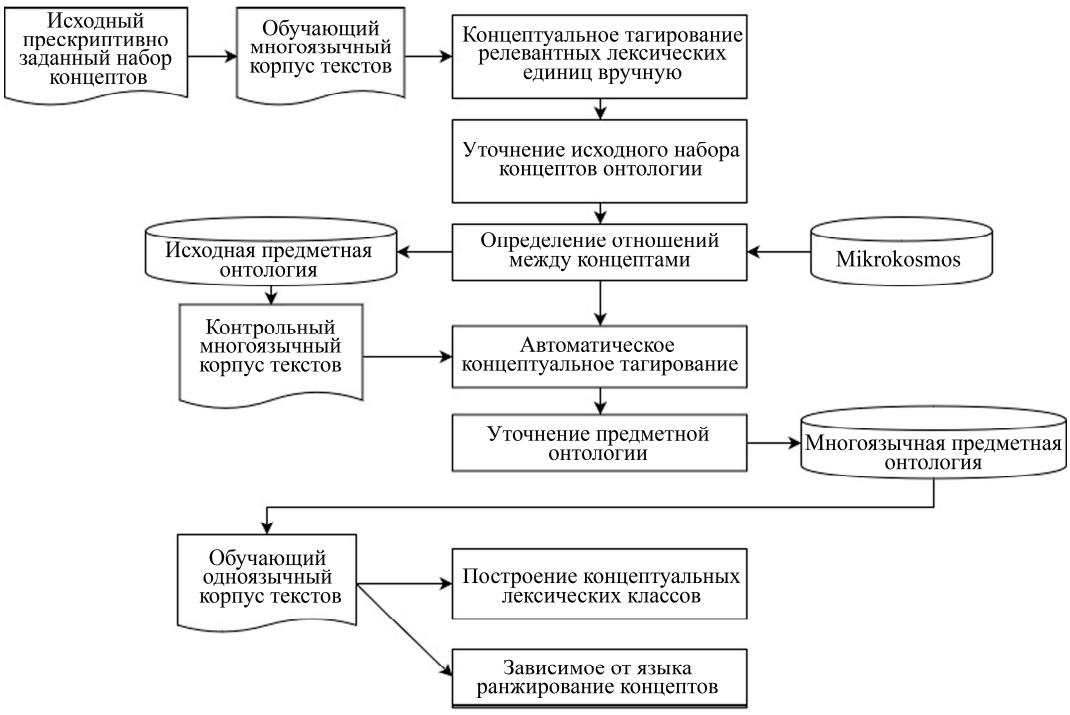

Рис. 1. Дорожная карта исследования

знаний о мире. Следовательно, предметная онтология должна быть связана с онтологией верхнего уровня.

- Пределы анализа и набор онтологических концептов и взаимосвязей между ними задаются эмпирическими данными.

2. Многоязычный онтологический ресурс

В процессе настоящего исследования создана многоязычная, т. е. не зависящая от конкретного языка, формальная онтология, ориентированная на автоматизированный анализ интернет-сообщений о террористических актах на различных языках. На первом этапе работы был задан прескриптивный исходный набор семантических (концептуальных) категорий, релевантных (по мнению исследователей) для контента новостей о террористических актах, и собраны три корпуса интернет-сообщений указанной предметной области на французском, английском и русском языках, объемом 500 000 словоупотреблений каждый. Каждый корпус был разделен на две части: обучающую и контрольную. Обучающие части корпусов текстов каждого языка были вручную размечены (протегированы) прескриптивным набором семантических (предположительно концептуальных) категорий, которые в процессе разметки уточнялись и дополнялись. Таким образом, прескриптивный метод был усилен дескриптивным.

В результате этого этапа анализа сформирован набор базовых концептов предметной области с их основными атрибутами и отношениями, множество которых затем дополнялось на основе анализа корпусов с помощью текстовых шаблонов (см. детальное описание этой процедуры в [10]). Полученные онтологические знания представлены в формализме онтологии Mikrokosmos [9], которая в нашем исследовании использована в качестве онтологии верхнего уровня. Для совместимости нашей предметной онтологии с онтологией Mikrokosmos онтологические концепты обозначены английскими словами. Отметим, что содержание концепта определяется не используемым для его обозначения словом, а его дефиницией (табл. 1). Фрагмент построенной на этом этапе исходной онтологии показан на рис. 2.

Таблица 1

Фрагмент набора концептов исходной онтологии с дефинициями

Концепт	Дефиниция
ADVERSARY'S PLANS	Деятельность по планированию теракта
AGENT	Исполнитель террористического акта
GOAL OF ATTACK	Цель совершения террористического акта
OBJECT OF ATTACK	Объект, на который направлен теракт
MEANS OF ATTACK (WEAPON)	Оружие или подобные ему предметы, используемые для совершения теракта
TERROR ATTACK	Тип теракта: взрыв, похищение, поджог и т. д.
ASSUMPTION	Предположения о террористической группе, совершившей теракт, или об организаторе теракта
CONSEQUENCES	Результаты теракта: человеческие жертвы, разрушенные объекты, исход для террористов
SOURCE	Источники сообщений о теракте: газеты, ТВ-каналы, информагентства или органы власти

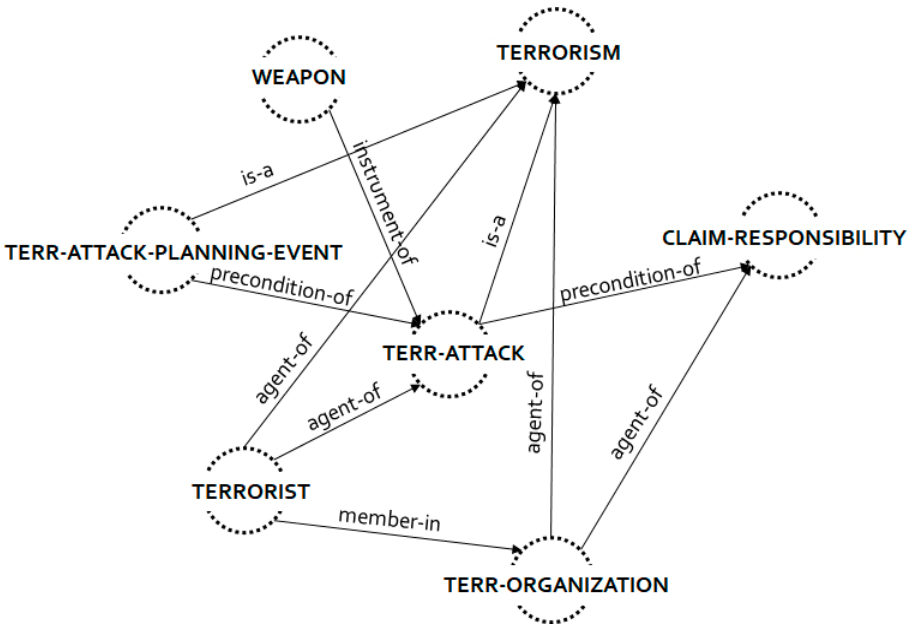

Рис. 2. Фрагмент онтологии предметной области «Терроризм»

На следующем этапе разработки онтологии из контрольных частей русского, английского и французского корпусов текстов с применением автоматического экстрактора [11] были извлечены и концептуально классифицированы многокомпонентные лексические единицы определенных частей речи (именных, глагольных и т. д. групп). При этом набор уже существующих онтологических концептов и отношений был еще раз уточнен и увеличен.

На основе полученных на этом этапе лексико-онтологических знаний разработана компьютерная платформа, с помощью которой проведено автоматическое концептуальное тегирование контрольных частей английского, французского и русского корпусов текстов с очередным уточнением онтологии. Для каждого национального языка построены частотные лексиконы одно- и многокомпонентных текстовых единиц, отображенных на концепты онтологии и, следовательно, релевантных для передачи контента интернет-сообщений о террористических актах.

3. Онтологический анализ одноязычного текста

Основой онтологического анализа является концептуальное тегирование лексических единиц текста, серьезной проблемой которого является неоднозначность. Например, в англоязычных сообщениях о терактах лексема *policeman* может отображаться на следующие концепты:

OBJECT OF ATTACK: A *policeman* was killed.

COUNTER-TERRORISM: The authorities deployed *policemen*.

AGENT: Russia's ambassador is assassinated in Ankara by a *policeman*.

Это означает, что при тегировании слову *policeman* будет приписан «мультитег», включающий три концептуальных тега. На рис. 3 представлена дистрибуция концептуальных мультитегов, соответствующих концептам верхнего уровня предметной онтологии, в англоязычном корпусе текстов. Использованы следующие теги: P = CONSEQUENCES, L = LOCATION, Z = OBJECT OF ATTACK, T = TYPE OF ATTACK, S = SOURCE, BW = TIME, D = DECLARATION, A = AGENT, RW = COUNTER-TERRORISM, UW = TERRORIST ORGANIZATION, C = MEANS OF ATTACK, CR = CLAIM RESPONSIBILITY, N = NATION, DA = DIRECTION OF ATTACK, I = ASSUMPTION, M = CHARACTER OF ATTACK, OW = OTHER, E = OTHER TERRORIST ACTIVITIES, X = GOAL OF ATTACK, HA = HAVE MEANS OF ATTACK.

Из рис. 3 видно, что уровень концептуальной многозначности текста достаточно велик, и для корректной интерпретации результатов онтологического анализа неоднозначность должна быть устранена. Одним из решений этой проблемы может быть использование контекста предложения или целого текста. Однако полный контекстный анализ, как правило, требует очень большого количества лингвистических знаний, получение которых проблематично.

Мы выдвигаем гипотезу о том, что для разрешения многозначности концептуальных тегов могут использоваться как дистрибутивные характеристики индивидуальных концептуальных тегов в одноязычном тексте, так и частотные показатели лексических единиц соответствующего языка, отраженных на концепты онтологии. На рис. 4 представлена дистрибуция концептов многоязычной онтологии (индивидуальных концептуальных тегов) в англоязычном корпусе текстов, показывающая их приоритетное использование.

Для разрешения концептуальной многозначности введены следующие параметры и их обозначения:

- *CF* (concept frequency) – частота употребления концепта (концептуального тега) в одноязычном тексте при концептуальном тегировании;

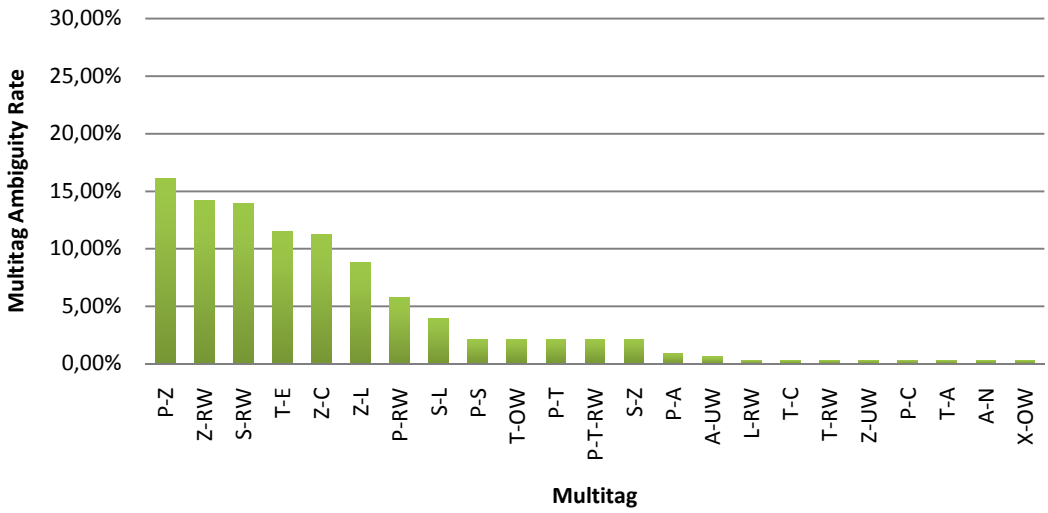

Рис. 3. Распределение концептуальных мультитегов в англоязычном корпусе текстов; за 100 % принимается общее количество мультитегов

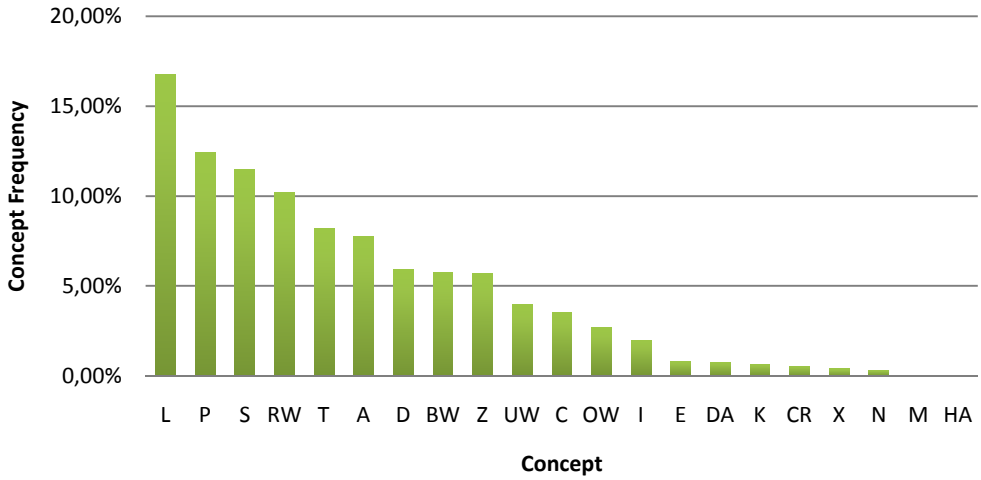

Рис. 4. Дистрибуция концептов (концептуальных тегов) в англоязычном корпусе текстов; за 100% принимается общее количество концептуальных тегов в корпусе

- *RCF* (ratio of concept fillers) – коэффициент лексического наполнения концепта, который учитывает количество единиц лексикона конкретного языка, отображаемых на концепт, и рассчитывается следующим образом:

$$RCF = n \div N,$$

где n – количество связанных с онтологией единиц одноязычного лексикона, соотнесенных с одним определенным концептом, а N – общее количество единиц в отраженном на все концепты онтологии одноязычном лексиконе;

- *CUR* (concept usage relevancy) – релевантность концепта многоязычной онтологии для онтологического анализа конкретного одноязычного текста. Эмпирическая формула, которую мы используем на данном этапе исследования, имеет вид:

$$CUR = (RCF \times 10 + CF) \div T,$$

где *CUR* – релевантность концепта, *RCF* – коэффициент лексического наполнения концепта, *CF* – частота концепта, *T* – количество словоформ в одноязычном корпусе текстов.

Ранжирование понятий онтологии в соответствии со значением параметра *CUR* может быть полезным при разработке эвристик для разрешения неоднозначности концептуального тегирования. Чем выше значение параметра *CUR*, тем более вероятен соответствующий тег. В табл. 2 приведены значения введенных параметров для концептов, теги которых включены в первые три наиболее часто встречающиеся в англоязычном корпусе интернет-сообщений о террористических актах мультитега P-Z, Z-RW, S-RW (см. рис. 3).

В соответствии с вычисленными значениями параметра *CUR* неоднозначность мультитегов P-Z и Z-RW скорее всего может быть разрешена в пользу Z, а мультитег S-RW может быть заменен тегом S. Мы осознаем, что в этом направлении необходимо провести дальнейшие исследования, но уже полученные результаты с достаточной степенью обоснованности показывают, что *CUR* может выступать в качестве одного из основных параметров разрешения концептуальной многозначности.

Заключение

В статье на примере предметной области «Терроризм» приведены предварительные результаты разработки модели онтологического анализа многоязычных электронных текстовых потоков. Исследование включает создание предметной многоязычной онтологии и разработку модели онтологического анализа. Метод создания онтологии описан на примере извлечения онтологических знаний из трех корпусов текстов интернет-сообщений о террористических актах на английском, французском и русском языках и их представления в формализме онтологии Mikrokosmos. На примере использования разработанной многоязычной онтологии для анализа англоязычных текстов предметной области рассмотрена проблема снятия концептуальной неоднозначности. Введены метрики

Таблица 2

Значения метрик RCF, CF и CUR для концептов P, Z, RW и S

Концепт	RCF	CF	CUR
P (CONSEQUENCES)	0,19	0,12	2,89
RW (COUNTER-TERRORISM)	0,22	0,10	3,28
S (SOURCE)	0,30	0,11	4,44
Z (OBJECT OF ATTACK)	0,30	0,06	4,37

ранжирования онтологических концептов. Предложены два новых количественных параметра: *коэффициент лексического наполнения концепта*, который учитывает количество единиц лексикона конкретного языка, отображаемых на конкретный концепт, и *релевантность концепта* многоязычной онтологии для онтологического анализа конкретного одноязычного текста, вычисляемого на основе эмпирической формулы. Значения введенных параметров могут служить основой для разрешения концептуальной неоднозначности, и мы, осознавая необходимость получения большего количества данных и возможного введения дополнительных параметров, считаем проблему разрешения концептуальной многозначности предметом наших дальнейших исследований. Кроме того, мы продолжим развивать структуру онтологии «вглубь» и «вширь», наращивать объемы связанных с онтологией одноязычных лексиконов, а также проводить работы по усовершенствованию модели онтологического анализа.

Литература/References

1. Berners-Lee T., Hendler J., Lassila O. The Semantic Web. *Scientific American.* 2001, 284 (5), pp. 34–43.

2. D'Aquin M., Noy N.F. Where to Publish and Find Ontologies? A Survey of Ontology Libraries. *Web Semantics: Science, Services and Agents on the World Wide Web.* 2012, no. 11, pp. 96–111.

3. Façanha R.L., Cavalcanti M.C., Campos M.L.M. A Systematic Approach to Review Legacy Schemas Based on Ontological Analysis. *Metadata and Semantic Research. MTSR* 2018. Ed. by E. Garoufallou, F. Sartori, R. Siatri, M. Zervas. (Communications in Computer and Information Science Series, vol. 846). Springer, Cham, 2019, pp. 63–75.

4. Fernández N., Fuentes D., Sánchez L., Fisteus J.A. The News ontology: Design and applications. *Expert Systems with Applications.* 2010, no. 37 (12), pp. 8694–8704. DOI: 10.1016/j.eswa.2010.06.055.

5. Green P.S., Rosemann M., Indulska M. *The Practice of Ontological Analysis.* 2005. https://pdfs.semanticscholar.org/513c/a04a8132a723cf47d9d9504983a98dd9ec08.pdf

6. Inyaem U., Haruechaiyasak Ch., Meesad Ph., Tran D. Ontology-Based Terrorism Event Extraction. *Proceedings of the 1st International Conference on Information Science and Engineering.* Nanjing, China, 2009, pp. 912–915. DOI: 10.1109/ICISE.2009.804

7. Mannes A., Golbeck J. Building a Terrorism Ontology. *ISWC Workshop on Ontology Patterns for the Semantic Web.* 2005, vol. 36. – https://pdfs.semanticscholar.org/9bcb/90e48677e39da7b84939e8c8da2b2a63cde7.pdf

8. Najgebauer A., Antkiewicz R., Antkiewicz M., Kasprzyk R. The Prediction of Terrorist Threat on the basis of Semantic Association acquisition and Complex Network Evolution. *Journal of Telecommunications and Information Technology.* 2008, no. 2, pp. 14–20.

9. Nirenburg S., Raskin V. *Ontological Semantics.* MIT Press, Cambridge, MA, 2004. xxi, 420 p.

10. Sheremetyeva S., Zinovyeva A. On Modelling Domain Ontology Knowledge for Processing Multilingual Texts of Terroristic Content. *Proceedings of DTGS 2018: Digital Transformation and Global Society. (Communications in Computer and Information Science Series, vol. 859).* Springer, Cham, 2018, pp. 368–379. DOI: 10.1007/978-3-030-02846-6_30

11. Sheremetyeva S. Automatic Extraction of Linguistic Resources in Multiple Languages. *Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012.* Wroclaw, Poland, 2012. Pp. 44–52.

12. Tenenboim L., Shapira B., Shoval P. Ontology-Based Classification of News in an Electronic Newspaper. *Proceedings of the International Conference "Intelligent Information and Engineering Systems" INFOS 2008. (Information Science and Computing Series: Advanced Research in Artificial Intelligence).* Varna, Bulgaria, 2008. Pp. 89–97.

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

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и перевода, Южно-Уральский государственный университет (Челябинск), babinaoi@susu.ru

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

Неручева Екатерина Дмитриевна, лаборант НОЦ «Лингво-инновационные технологии» института лингвистики и международных коммуникаций, Южно-Уральский государственный университет (Челябинск), neruchevaekaterina@mail.ru

Поступила в редакцию 14 октября 2019 г.

DOI: 10.14529/ling200106

ON USING MULTILINGUAL ONTOLOGY TO ANALYSE ENGLISH E-NEWS ON TERRORIST ATTACKS

S.O. Sheremetyeva, *sheremetevaso@susu.ru*
O.I. Babina, *babinaoi@susu.ru*
A.Yu. Zinoveva, *bihcwd@bk.ru*
E.D. Nerucheva, *neruchevaekaterina@mail.ru*
South Ural State University, Chelyabinsk, Russian Federation

The article explores the issues of ontological analysis of electronic text streams in various languages based on a language-independent domain-specific ontological resource. The research methodology combines descriptive and prescriptive approaches and includes several iterative steps: acquisition, analysis and formalization of textual data with subsequent identification of their conceptual structure, which, in turn, is used both to determine and improve the ontology content, and to refine the ontological analysis of specific monolingual texts. The approach to the ontological analysis is presented based on the example of its application to the English-language e-news on terrorist attacks. Particular attention is paid to the main problem of ontological analysis, which is the resolution of conceptual ambiguity in mapping lexical manifestations onto the ontological level of information representation.

Keywords: ontological analysis, multilingual subject ontology, e-news, terrorism, English.

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

Olga I. Babina, PhD, associate professor of the Department of Linguistics and Translation Studies, South Ural State University (Chelyabinsk), babinaoi@susu.ru

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

Ekaterina D. Nerucheva, laboratory assistant, Research and Education Centre of Innovative Linguistic Technologies, Institute of Linguistics and International Communications, South Ural State University (Chelyabinsk), neruchevaekaterina@mail.ru

Received 14 October 2019

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Анализ англоязычных интернет-сообщений о террористических актах на основе многоязычной онтологии / С.О. Шереметьева, О.И. Бабина, А.Ю. Зиновьева, Е.Д. Неручева // Вестник ЮУрГУ. Серия «Лингвистика». – 2020. – Т. 17, № 1. – С. 30–36. DOI: 10.14529/ling200106

FOR CITATION

Sheremetyeva S.O., Babina O.I., Zinoveva A.Yu., Nerucheva E.D. On Using Multilingual Ontology to Analyse English E-News on Terrorist Attacks. *Bulletin of the South Ural State University. Ser. Linguistics*. 2020, vol. 17, no. 1, pp. 30–36. (in Russ.). DOI: 10.14529/ling200106