Анастасия Юрьевна ЗИНОВЬЕВА[1]
Светлана Олеговна ШЕРЕМЕТЬЕВА[2]
Екатерина Дмитриевна НЕРУЧЕВА[3]

УДК 811.161.1 + 004.82

АНАЛИЗ НЕОДНОЗНАЧНОСТИ КОНЦЕПТУАЛЬНОЙ РАЗМЕТКИ РУССКОЯЗЫЧНОГО ТЕКСТА

[1] аспирант кафедры лингвистики и перевода,
Южно-Уральский государственный университет (г. Челябинск)
zinovevaaiu@bk.ru; ORCID: 0000-0002-7658-7376

[2] доктор филологических наук,
профессор кафедры лингвистики и перевода,
Южно-Уральский государственный университет (г. Челябинск)
sheremetevaso@susu.ru

[3] лаборант НОЦ «Лингво-инновационные технологии»,
Южно-Уральский государственный университет (г. Челябинск)
neruchevaekaterina@mail.ru

Аннотация

Наличие корректно размеченных (аннотированных) корпусов текстов является критически важным условием создания эффективных средств автоматизированной обработки естественного языка, обеспечивающих оперативное решение как теоретических, так и прикладных лингво-информационных задач. Одной из основных и наиболее сложных проблем корпусной разметки является разрешение неоднозначности меток на конкретном уровне реализации аннотирования (морфологическом, синтаксическом, семантическом и т. д.).

Настоящая статья посвящена проблеме неоднозначности, возникающей на концептуальном, наиболее релевантном для решения информационных задач уровне разметки текстов. Под концептуальной разметкой (аннотированием) понимается специальный тип

Цитирование: Зиновьева А. Ю. Анализ неоднозначности концептуальной разметки русскоязычного текста / А. Ю. Зиновьева, С. О. Шеремetьева, Е. Д. Неручева // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2020. Том 6. № 3 (23). С. 38-60.
DOI: 10.21684/2411-197X-2020-6-3-38-60

семантической разметки, как правило, применяемый к корпусам предметных областей для решения конкретных информационных задач (автоматической классификации, контент- и тренд-анализов, машинного обучения, машинного перевода и др.).

При концептуальной разметке корпусы текстов размечаются метками, отражающими контент конкретной предметной области, что ведет к отличному от общесемантического типу неоднозначности, который имеет как универсальные, так и зависящие от конкретного языка и предметной области характеристики. В статье проблема концептуальной неоднозначности исследуется методом кейс-стади на материале русскоязычных текстов предметной области «Терроризм».

Методология исследования сочетает автоматизированные и вручную выполненные этапы работ, включающие а) статистико-качественный анализ корпусного материала; б) использование предварительно разработанных аннотационных ресурсов (онтологии предметной области «Терроризм», русского онтолексикона и компьютерной платформы концептуального аннотирования); в) основанную на онтологическом анализе концептуальную разметку отобранного для кейс-стади корпуса; г) основанное на корпусном подходе выявление и анализ причин возникновения концептуальной неоднозначности; д) исследование статистических параметров концептуальных меток и соотнесенных с ними лексем в аннотированном корпусе; е) разработка и экспериментальная проверка возможных методов разрешения отдельных типов концептуальной неоднозначности.

В настоящем исследовании получены конкретные результаты для русскоязычных текстов, но разработанная методика концептуальной разметки и подходы к разрешению концептуальной неоднозначности применимы к текстам других предметных областей на различных языках.

Ключевые слова

Концептуальная разметка, русскоязычный корпус, концептуальная неоднозначность, метод кейс-стади, онтологический анализ, многоязычная предметная онтология, терроризм.

DOI: 10.21684/2411-197X-2020-6-3-38-60

Введение

В настоящее время в сфере автоматической обработки текстов наблюдается тенденция к семантизации текстовых метаданных посредством онтологического анализа. Под онтологическим анализом понимается процесс извлечения знаний о сущностях какой-либо предметной области [9]. Этот процесс осуществляется в два этапа: на первом этапе лексические единицы текста аннотируются посредством меток (тегов), обозначающих концепты онтологии, а на втором проводится формализация и интерпретация результатов этого аннотирования в рамках конкретной задачи. Для обозначения первого этапа исследователи используют различные термины: в некоторых работах применяются широкие термины *семантическая аннотация* или *разметка* [10, 18], которые в общем случае обозначают процедуру обогащения контента различной семантической информацией; другие авторы предпочитают более узкие термины *концептуальная*

аннотация или *концептуальная разметка* (англоязычные эквиваленты — *conceptual annotation* [15], *concept labeling* [16]).

Содержание понятия *семантическая разметка*, по-видимому, является дискуссионным: разными учеными под семантической разметкой понимается, во-первых, определение значений многозначных лексем на основе словаря или онтологии [6]; во-вторых, выявление их универсальных семантических свойств на основе какой-либо лексической классификации [3]; и, в-третьих, отображение семантических отношений между словами в тексте [12]. В свою очередь, на основе анализа работ Дж. Д. Ким и др. [10], Дж. С. Виджу [16] и М. Ю. Загорулько и др. [18] концептуальная разметка может быть определена как частный случай семантической разметки, основанный на онтологии и ориентированный на ограниченную предметную область.

Например, такие слова, как *человек* и *ребенок*, в текстах новостных интернет-сообщений о терроризме на русском языке часто свидетельствуют о последствиях террористического акта, поскольку люди (в частности, дети) могут быть убиты, ранены или взяты в заложники; при этом семантический компонент ‘последствия террористического акта’ в этих лексемах отсутствует. В настоящей работе мы используем термин *концептуальная разметка,* имея в виду отдельный тип семантической разметки текстов концептами предметной области и, таким образом, отличаем его от более общего термина *семантическая разметка*.

В данной статье рассматривается проблема концептуальной неоднозначности, возникающая при концептуальном аннотировании даже в ограниченной предметной области [14]. В ходе онтологического анализа лексической единице может быть присвоено два или более концептуальных тега. Цель исследования заключается том, чтобы изучить проблему концептуальной неоднозначности в ограниченной предметной области, определить ее источники и рассмотреть возможные способы ее разрешения на материале русскоязычного корпуса текстов интернет-сообщений о террористических актах. Отметим, что в статье мы используем принятые в русскоязычной литературе отечественные и синонимичные им заимствованные термины, такие как аннотация (разметка) аннотирование/тегирование (процесс разметки), тег (метка) и т. д.

Остальная часть статьи структурирована следующим образом. Во втором разделе описаны ресурсы, использованные в ходе исследования, в третьем — методология исследования. В четвертом и пятом разделах представлены результаты исследования (источники концептуальной неоднозначности, выявленные в русскоязычном корпусе текстов), и обсуждение возможных методов разрешение концептуальной неоднозначности. В шестом разделе приведены выводы и описаны дальнейшие перспективы исследования.

Исследовательские ресурсы

Исследование концептуальной неоднозначности русскоязычного текста является одним из этапов проекта по созданию концептуально аннотированных многоязычных корпусов текстов. В рамках исследования были использованы

предварительно созданные ресурсы концептуального аннотирования и русскоязычный корпус объемом 26 442 словоупотребления, содержащий 262 новостных интернет-сообщения о террористических актах за 2014-2019 гг. Корпус был собран из открытых источников, в частности сайтов новостных агентств РИА «Новости», ИА Regnum, ТАСС и т. д.

Одним из основных компонентов аннотационных ресурсов, использованных в настоящем исследовании, является многоязычная онтология предметной области «Терроризм», предназначенная для анализа текстов о террористической деятельности на русском, английском и французском языках [13]. Знания онтологии представлены в формализме многоязычной онтологии MikroKosmos [11], поэтому наименования концептов онтологии даны на английском языке, при этом содержание каждого концепта определяется только его дефиницией (таблица 1).

Следующий компонент аннотационных ресурсов — основанный на корпусе русскоязычный лексикон контентно-релевантных одно- и многокомпонентных лексических единиц, отображенных на концепты онтологии (далее *онтолексикон*); при этом одна и та же лексема может быть отображена на несколько концептов онтологии (см. примеры в таблице 2).

Такое неоднозначное отображение обусловлено двумя различными лингвистическими явлениями — концептуальной неоднозначностью и концептуальной синкретичностью. Концептуально неоднозначными являются лексические единицы, которые в проанализированном корпусе предметной области встречаются в нескольких концептуальных значениях, но при этом в каждом конкретном случае употребления реализуется только одно из возможных концептуальных значений,

Таблица 1

Примеры концептов онтологии предметной области «Терроризм» с дефинициями и лексическими примерами на русском языке

Table 1

Examples of the terrorism domain ontology concepts with definitions and lexical examples in Russian

Концепт	Дефиниция	Примеры
Terrorism-Agent	Исполнитель ТА, человек или группа людей, занимающиеся террористической деятельностью	*боевик, вербовщик, стрелок*
Counterterrorism	Работа по противодействию терроризму, агенты и средства контртерроризма	*операция, полиция, рейд, задержание*
Object of attack	Объект, на который направлен ТА	*авиабаза, вуз, журналист*
Terror attack	Атака, совершаемая террористом или группой террористов для устрашения населения	*акт террора, бойня, взрыв, трагедия*
Weapon	Оружие или подобные ему предметы, используемые для совершения ТА	*автомобиль, автомат, бомба, грузовик*

Таблица 2

Лексические единицы, связанные с несколькими концептами

Table 2

Lexical items mapped into several concepts

Лексическая единица	Концепт							
	A	C	L	N	P	S	T	Z
взрыв заминированного автомобиля		●					●	
водитель автомобиля	●				●			
жертва взрыва					●		●	
жилые кварталы Алеппо			●					●
Исламское государство	●					●		
нападавший	●				●			
премьер-министр					●	●		
российский турист				●	●			●

Примечания: концепты закодированы под следующими тегами: A = Terrorism-Agent, C = Weapon, L = Location, N = Nation, P = Consequences, S = Source, T = Terror attack, Z = Object of attack.

Notes: the concepts are coded with the tags: A = Terrorism-Agent, C = Weapon, L = Location, N = Nation, P = Consequences, S = Source, T = Terror attack, Z = Object of attack.

как, например, у лексемы *солдат*. Ниже приведены четыре предложения, в каждом из которых данное слово реализует одно из своих концептуальных значений:

1. «По данным агентства, исполнитель взрыва в Германии был одним из солдат (= террористический деятель) исламистов»[1].
2. «Сообщается о гибели десяти солдат (= последствия)»[2].
3. «Солдаты (= представители контртеррористической деятельности) выстрелили и нейтрализовали нападавшего»[3].
4. «По информации пресс-служб ЦАХАЛа, солдаты (= объект террористического акта) заметили террориста, бежавшего в их сторону»[4].

[1] «Исламское государство» взяло ответственность за взрыв в Ансбахе // Новая газета. 2016. URL: http://novayagazeta.ee/articles/8246/

[2] Атака на военную базу в Сомали: есть жертвы // Корреспондент. 2016. URL: https://korrespondent.net/world/3713902-ataka-na-voennuui-bazu-v-somaly-est-zhertvy

[3] Еще одна попытка теракта в Хевроне. Террористка арестована // NEWSru.co.il. 2019. URL: https://www.newsru.co.il/israel/12mar2019/hebron_002.html

[4] Попытка теракта в Хевроне, террорист ликвидирован // NEWSru.co.il. 2019. URL: https://www.newsru.co.il/israel/12mar2019/tkifa_706.html

Концептуальная синкретичность может иметь место в трех случаях:

1) если каждый компонент многокомпонентной лексической единицы относится к отдельному концепту, при этом многокомпонентная единица не может быть разделена из-за тесной смысловой связи компонентов;
2) если лексическая единица содержит несколько релевантных семантических компонентов, в том числе вне указанного контекста;
3) если в данном контексте лексическая единица обладает двумя концептуальными значениями, которые не противоречат друг другу.

Все три случая концептуального синкретизма можно наблюдать в предложении «*Правоохранительные органы Колумбии задержали подозреваемого в организации теракта 39-летнего Рикардо Андреса Карвахаля*»[1]. Трехкомпонентная единица *правоохранительные органы Колумбии* содержит компоненты *правоохранительные органы* и *Колумбия*, первый из которых относится к концепту COUNTERTERRORISM, а второй — к концепту NATION (случай 1). Лексическая единица *подозреваемый* содержит семантические компоненты ‘действующее лицо’ и ’подозрение’, поскольку *подозреваемый* — это, по определению, тот, кто предполагается виновным в совершении преступления, а значит, относится одновременно к концептам TERRORISM-AGENT и ASSUMPTION (случай 2). В свою очередь, глагол *задержать* в данном контексте обладает двумя релевантными концептуальными значениями и связывается с концептами CONSEQUENCES и COUNTERTERRORISM, поскольку передает информацию о контртеррористическом действии и последствиях для террориста (случай 3).

К предварительно разработанным аннотационным ресурсам относятся также следующие: модель концептуального аннотирования, основанная на онтологическом анализе, который представляет собой отображение лексических единиц текста на концепты онтологии и формально представлен тегированием единиц текста метками (тегами) концептов онтологии, и компьютерная платформа концептуального аннотирования, база знаний которой содержит все возможные в предметной области концептуальные значения контентно-релевантных лексем.

Методология

При изучении концептуальной неоднозначности русскоязычного текста был использован широко применяемый в прикладной лингвистике метод кейс-стади [17] с сочетанием автоматизированных и вручную выполненных этапов работ с использованием предварительно созданных ресурсов концептуального аннотирования, описанных в предыдущем разделе. В частности, методология данного исследования включает следующие конкретные этапы:

1. Выбор и кодирование представительного для кейс-стади *набора концептов*, в который было отобрано 23 наиболее часто встречающихся концепта предметной области со следующей кодировкой (тегами): A —

[1] Теракты 2019 года, список и краткое описание // Посреди России. 2019. URL: https://posredi.ru/terakty-2019-goda.html

Terrorism-agent, BW — Time, C — Weapon, CR — Claim responsibility, D — Declare, DA — Direction of attack, E — Threat, EW — Cause, HA — Have weapon, I — Assumption, K — Adversary's plans, L — Location, M — Scale of attack, N — Nation, OW — Other-terrorism, P — Consequences, RW — Counterterrorism, S — Source, T — Terror attack, UW — Terrorist organization, X — Goal of attack, Y — Reaction, Z — Object of attack. Кроме того, был введен ряд неконцептуальных тегов для лексических единиц, нерелевантных для предметной области: разного рода неконцептуальных предикатов (R, B, U), именных групп (PO), чисел (Num), детерминативов (DEF), прочих частей речи (O), а также неизвестных единиц, отсутствующих в лексиконе (UNK).

2. *Основанная на онтологическом анализе концептуальная разметка* отобранного для кейс-стади корпуса, которая осуществлялась в два этапа. На первом этапе корпус был размечен автоматически при помощи разработанной платформы аннотирования с присвоением лексемам текста тегов всех связанных с ними концептов онтологии, что отражает все возможные в предметной области концептуальные значения лексем. На втором этапе результаты автоматической разметки были вручную отредактированы экспертами. Таким образом был получен «золотой» корпус с идеальной разметкой. Примеры автоматической и «золотой» разметок представлены в таблице 3.

Сочетание нескольких тегов, присваиваемых одной лексической единице, далее будет называться *мультитегом*. Очевидно, что мультитеги обусловлены либо концептуальной неоднозначностью, либо концептуальной синкретичностью. Напомним, что под последней мы понимаем одновременную реализацию нескольких не противоречащих друг другу концептуальных значений в одной лексической единице.

Мультитеги, реализующие синкретичные концептуальные значения лексем текста, в настоящем исследовании не считаются подлежащими разрешению. В остальных случаях появления мультитегов необходимо разрешение неоднозначности. Следовательно, требуется дальнейший анализ, который позволил бы выявить, во-первых, источники концептуальной неоднозначности, и, во-вторых, способы ее разрешения.

1. Выявление и анализ причин возникновения концептуальной неоднозначности, основанные на корпусном подходе.
2. Исследование статистических параметров концептуальных меток и соотнесенных с ними лексем в аннотированном корпусе.
3. Разработка и экспериментальная проверка методов разрешения отдельных типов концептуальной неоднозначности

Этапы 1-2 обеспечивают получение данных кейс-стади, анализ которых на этапах 3-5 обеспечивает получение результатов настоящего исследования, которые описаны в следующем разделе.

Таблица 3

Автоматическая и «золотая» разметка новостного сообщения

Table 3

Automatic and “golden” annotation of a news article

Автоматическая разметка	«Золотая» разметка
{В}$^{\sim O}$ {Нигере}$^{\sim L\sim N}$ {при}$^{\sim O}$ {атаке}$^{\sim T}$ {на}$^{\sim O}$ {военный пост}$^{\sim Z}$ {погибли}$^{\sim P}$ {89}$^{\sim Num}$ {человек}$^{\sim PO\sim P\sim Z\sim A}$	{В}$^{\sim O}$ {Нигере}$^{\sim L\sim N}$ {при}$^{\sim O}$ {атаке}$^{\sim T}$ {на}$^{\sim O}$ {военный пост}$^{\sim Z}$ {погибли}$^{\sim P}$ {89}$^{\sim Num}$ {человек}$^{\sim P}$
{89}$^{\sim Num}$ {военнослужащих}$^{\sim P\sim Z\sim RW}$ {стали жертвами}$^{\sim P}$ {атаки}$^{\sim T}$ {неизвестных}$^{\sim R\sim A}$ {на военную базу}$^{\sim L}$ {в}$^{\sim O}$ {Республике Нигер}$^{\sim L\sim N}$, {пишет}$^{\sim D}$ {Reuters}$^{\sim S}$.	{89}$^{\sim Num}$ {военнослужащих}$^{\sim P}$ {стали жертвами}$^{\sim P}$ {атаки}$^{\sim T}$ {неизвестных}$^{\sim A}$ {на военную базу}$^{\sim L}$ {в}$^{\sim O}$ {Республике Нигер}$^{\sim L\sim N}$, {пишет}$^{\sim D}$ {Reuters}$^{\sim S}$
{Инцидент}$^{\sim PO\sim T}$ {случился}$^{\sim R}$ {в субботу}$^{\sim BW}$ {в}$^{\sim O}$ {населенном пункте}$^{\sim Z\sim L}$ {Шинегодар}$^{\sim L}$, {расположенном}$^{\sim R}$ {на западе государства}$^{\sim L}$.	{Инцидент}$^{\sim T}$ {случился}$^{\sim R}$ {в субботу}$^{\sim BW}$ {в}$^{\sim O}$ {населенном пункте}$^{\sim L}$ {Шинегодар}$^{\sim L}$, {расположенном}$^{\sim R}$ {на западе государства}$^{\sim L}$.
{Нападавшие}$^{\sim P\sim A}$ {приехали}$^{\sim R}$ {к}$^{\sim O}$ {посту}$^{\sim PO\sim Z}$ {на}$^{\sim O}$ {автомобилях}$^{\sim PO\sim P\sim Z\sim C}$ {и}$^{\sim O}$ {мотоциклах}$^{\sim PO\sim C}$ {и}$^{\sim O}$ {открыли огонь}$^{\sim T}$.	{Нападавшие}$^{\sim A}$ {приехали}$^{\sim R}$ {к}$^{\sim O}$ {посту}$^{\sim Z}$ {на}$^{\sim O}$ {автомобилях}$^{\sim PO}$ {и}$^{\sim O}$ {мотоциклах}$^{\sim PO}$ {и}$^{\sim O}$ {открыли огонь}$^{\sim T}$.
{По данным издания}$^{\sim S}$, {в настоящее время}$^{\sim BW}$ {против}$^{\sim O\sim DA}$ {вооруженных сил республики}$^{\sim Z}$ {действуют}$^{\sim O\sim R\sim OW}$ {боевики}$^{\sim P\sim A\sim RW}$, {связанные с}$^{\sim B\sim OW}$ {запрещенной в РФ международной террористической организацией Аль-Каида}$^{\sim UW}$ {и}$^{\sim O}$ {радикальной исламистской группировкой Боко Харам}$^{\sim UW}$.	{По данным издания}$^{\sim S}$, {в настоящее время}$^{\sim BW}$ {против}$^{\sim DA}$ {вооруженных сил республики}$^{\sim Z}$ {действуют}$^{\sim OW}$ {боевики}$^{\sim A}$, {связанные с}$^{\sim OW}$ {запрещенной в РФ международной террористической организацией Аль-Каида}$^{\sim UW}$ {и}$^{\sim O}$ {радикальной исламистской группировкой Боко Харам}$^{\sim UW}$.

Результаты анализа концептуальной неоднозначности

В результате анализа «золотого» корпуса было выделено пять источников концептуальной неоднозначности.

Частеречная омонимия

Концептуальная неоднозначность, вызванная частеречной омонимией, возникает при совпадении по крайней мере одной словоформы у разных лексических единиц. Этот тип неоднозначности встречается как у релевантных для предметной области лексических единиц, так и у нерелевантных. Для русскоязычного корпуса частеречная омонимия как источник концептуальной неоднозначности непродуктивна; немногочисленные примеры представлены в таблице 4.

Таблица 4

Концептуальная неоднозначность, вызванная частеречной омонимией

Table 4

Conceptual ambiguity caused by part-of-speech homonymy

Лексическая единица	Часть речи	Начальная форма	Соотнесенность с концептом онтологии
после	сущ.	посол	Object of attack
			Source
			нет
	предлог	после	Time
			нет
кто	сущ.	КТО	Counterterrorism
	мест.	кто	нет
так	сущ.	ТАК	Terrorist organization
	наречие	так	нет
военный	сущ.	военный	Object of attack
			Consequences
			Counterterrorism
	прил.	военный	нет
задержанный	субст. прил.	задержанный	Terrorism-Agent
			Consequences
	причастие	задержанный	Consequences
			Counterterrorism

Примечания: в первой колонке единица указана в той форме, которой в тексте присваиваются соответствующие теги.

Notes: in the first column, the item is given in the form which is assigned the tag in the text.

Лексическая неоднозначность

Под лексической неоднозначностью, как правило, понимается способность лексической единицы обладать двумя или более интерпретациями в контексте, вызываемая полисемией или омонимией [2].

В исследуемом корпусе лексическая неоднозначность крайне редко является источником концептуальной неоднозначности. Одним из таких немногочисленных примеров является слово *огонь*, которое может употребляться в значении ‘стрельба’ (и тогда соотносится с концептом Terror attack) или

в значении ‘процесс горения’ (Other-terrorism), поэтому данной лексической единице при автоматическом аннотировании присваивается мультитег T-OW. Есть и другие примеры:

движение (неконцептуальное существительное / Terrorism-agent) —
1) ‘перемещение в определенном направлении’;
2) ‘общественная деятельность, преследующая определенные цели’.

следствие (Cause / Counterterrorism) —
1) ‘результат чего-н.’;
2) ‘расследование обстоятельств преступления’.

Синтаксическая неоднозначность

Под синтаксической неоднозначностью традиционно понимается неоднозначность синтаксической структуры словосочетания, вызванная синтаксическими средствами языка, а именно различным делением высказывания на синтагмы, порядком слов и валентностью [1].

В анализируемом корпусе мы обнаружили только один пример синтаксической неоднозначности, который приводит к концептуальной неоднозначности, — словосочетание *обстрел террористов*, автоматически размечаемое мультитегом T-RW. В данном примере неоднозначность возникает за счет неопределенности типа связи между компонентами словосочетания, поскольку здесь возможна как субъектная, так и объектная связь. Субъектная связь, в частности, реализуется в следующем предложении: *«Расположенные рядом с демилитаризованной зоной населенные пункты на севере Сирии практически ежедневно подвергаются минометным обстрелам террористов»*[1] — в то время как в предложении *«Украинские силовики начали авиационный и артиллерийский обстрел террористов»*[2] обнаруживается объектная связь. В случае субъектной связи между компонентами данного словосочетания необходимо разрешение концептуальной неоднозначности в пользу тега T, иначе — в пользу тега RW.

В тексте встречаются и другие словосочетания, построенные по этой модели и являющиеся, таким образом, синтаксически неоднозначными: например, *обстрел бандформирований, контроль боевиков*. Однако при автоматическом аннотировании этим словосочетаниям присваивается единичный тег — по всей видимости, результат того, что они не встречались в ином концептуальном значении в текстах предметной области, использованных нами для создания лексикона. В перспективе, по мере увеличения лексикона, такие словосочетания также могут стать источником концептуальной неоднозначности.

---

1 СМИ: в сирийской Хаме из-за обстрела террористов погиб мирный житель // РИА «Новости». 2019. URL: https://ria.ru/20190406/1552443035.html

2 Украинские силовики начали авиационный и артиллерийский обстрел террористов — ИС // НВ. 2014. URL: https://nv.ua/ukraine/ukrainskie-siloviki-nachali-aviacionnyy-i-artilleriyskiy-obstrel-terroristov-is-1886.html

Множественность концептуальных значений

Лексические единицы, идентичные по форме и словарному значению, могут соотноситься с разными концептами онтологии в зависимости от контекста. В этом случае можно говорить о множественности концептуальных значений, примером которой может послужить слово *водитель*, которое в разных контекстах может быть отнесено к концептам Terrorism-agent, Consequences или Object of attack:

1. «По данным агентства, взрывное устройство было заложено в автомобиле, водитель (= Terrorism-agent) которого протаранил здание Министерства труда и социальных вопросов, расположенное в административном центре Могадишо»[1].
2. «Водитель (= Consequences) и несколько сотрудников академии погибли на месте»[2].
3. «По данным полиции, подозреваемый затем приблизился к другому водителю (= Object of attack) в красном автомобиле марки Prius, стреляя в него»[3].

Чаще всего случаи концептуальной неоднозначности в русскоязычном корпусе обусловлены именно множественностью концептуальных значений: помимо слова *водитель*, примерами также являются слова *полиция*, *человек*, *мужчина*, *женщина*, *солдат*, *пассажир*, *автомобиль* и др.

Указанные лексические единицы также могут встречаться вне террористического контекста; в этом случае они должны размечаться как единицы, не связанные с концептами онтологии. Определение концептуальности лексической единицы само по себе является нетривиальной задачей. Для удобства такие случаи мы также относим к множественности концептуальных значений.

Экстралингвистический контекст

Разрешение концептуальной неоднозначности, вызванной экстралингвистическими факторами, проблематично не только для компьютерной программы, но и для человека-эксперта. Такая неоднозначность обусловлена различиями экстралингвистического характера, например, разными точками зрения двух сторон на ситуацию.

Например, слово *боевик* соотносится с концептами Terrorism-agent (A) и Counterterrorism (RW), причем частота его встречаемости в связи с первым значительно выше. Однако в предложении *«На территории Донецкой области,*

---

[1] Боевики взорвали здание Министерства труда и соцвопросов в Могадишо // ИА REGNUM. 2019. URL: https://regnum.ru/news/accidents/2597241.html

[2] Колумбия скорбит по жертвам теракта в Боготе. Погиб 21 человек // ТАСС. 2019. URL: https://tass.ru/proisshestviya/6013393

[3] Два человека убиты в результате стрельбы в американском Сиэтле // 112 Украина. 2019. URL: https://112.ua/mir/dvoe-chelovek-ubity-v-rezultate-strelby-v-amerikanskom-sietle-485681.html

оккупированной пророссийскими боевиками, был сбит пассажирский Boeing 777»[1] нельзя сделать однозначный выбор между тегами A и RW, поскольку существует по крайней мере две точки зрения на эту ситуацию: украинская сторона обвиняет в крушении самолета российских военных; российская сторона, в свою очередь, указывает на украинских военных как возможных виновников катастрофы.

Аналогичная ситуация наблюдается с лексической единицей *КСС (курдские силы самообороны)*, которая может быть связана с концептами Terrorist organization (UW) и Counterterrorism (RW), т. к., с одной стороны, КСС выступают против Исламского государства, и следовательно их действия можно считать контртеррористическими; но с другой стороны, они совершают террористические акты против турецких военных, из-за чего их можно считать террористической организацией.

Все обнаруженные нами случаи концептуальной неоднозначности, вызванной экстралингвистическими факторами, связаны с разделением организаций на террористические и нетеррористические, а также отдельных лиц на террористов и представителей контртеррористической деятельности.

Проблема экстралингвистического контекста, таким образом, видится нам довольно серьезной, поскольку каждый такой случай требует тщательного рассмотрения. В частности, при выборе концепта необходимо решить, на каком основании мы определяем организацию как террористическую или нетеррористическую, а лицо — как террориста или нетеррориста. Данное решение может быть принято на основе контекста или какого-либо списка, но оба варианта влекут за собой некоторые проблемы.

В первом случае точка зрения издания, в котором опубликовано новостное сообщение, может не совпадать с мнением мирового сообщества, и тогда организация, являющаяся террористической, может быть размечена как нетеррористическая. Во втором случае затруднения вызывает выбор списка террористических организаций, на который следует ориентироваться при разметке, учитывая, что в разных странах эти списки могут быть различны. Например, в России ХАМАС не признан террористической организацией [4] в отличие от Евросоюза, где организация на данный момент включена в соответствующий список [8]. Кроме того, эти списки периодически изменяются, что указывает на необходимость использования динамических ресурсов для разрешения концептуальной неоднозначности такого типа.

Можно заключить, что проблема концептуальной неоднозначности, проистекающей из экстралингвистического контекста, изучена недостаточно и требует дальнейшего анализа.

Смешанные источники

Важно отметить, что концептуальная неоднозначность лексической единицы может проистекать одновременно из нескольких источников. Например, слово

[1] Самые кровавые теракты за последние 15 лет. Инфографика // ТСН. 2015. URL: https://tsn.ua/ru/svit/samye-krovavye-terakty-za-poslednie-15-let-infografika-525118.html

операция, автоматически размечаемое тегами P-RW-OW, имеет два источника концептуальной неоднозначности: лексическую неоднозначность и множественность концептуальных значений.

В исследуемом корпусе слово *операция* может обладать одним из двух лексических значений:

1) ‘ряд военных действий, объединенных одной целью’;

2) ‘хирургическая лечебная помощь’.

В первом значении слово употребляется довольно часто, причем при описании действий как представителей контртеррористической деятельности, так и террористов, т. е. в этом лексическом значении это слово имеет также два концептуальных значения. Например, в предложении «*В ответ на активизацию боевиков армия Буркина-Фасо проводит операции по выявлению и ликвидации баз террористов*»[1] слову *операция* должен быть присвоен тег RW, в то время как в предложении «*В начале месяца в МИД РФ предупредили, что террористы готовят военную операцию в сирийском Идлибе*»[2] — тег OW, указывающий на другую террористическую деятельность.

Во втором лексическом значении слово *операция* употребляется редко и связывается только с тегом P, обозначающим последствия террористического акта: «*Пострадавший доставлен в больницу „Шаарей-Цедек“ в Иерусалиме. Ему предстоит пройти операцию*»[3].

Другим примером концептуальной неоднозначности со смешанным источником является упомянутая ранее словоформа *после*, неоднозначность которой обусловлена множественностью концептуальных значений в совокупности с частеречной омонимией. Если *по́сле* — предлог, то он может быть либо неконцептуальным (при использовании в нетеррористическом контексте), либо связанным с концептом Time (тег BW). Если же *после́* — это форма слова *посол* в предложном падеже, то она может быть соотнесена с одним из концептов — Object of attack или Source — или же может представлять собой неконцептуальное существительное.

На наш взгляд, концептуальную неоднозначность в подобных примерах следует разрешать поэтапно.

Обсуждение подходов к разрешению концептуальной неоднозначности

В рамках настоящего исследования нами были проанализированы три возможных количественных подхода к разрешению концептуальной неоднозначности на основе корпусных данных (ранжирование тегов, совместная встречаемость тегов, позиционный подход).

---

1 AFP: в Буркина-Фасо при нападении на церковь погибли 14 человек // ТАСС. 2019. URL: https://tass.ru/mezhdunarodnaya-panorama/7239641

2 В результате двух взрывов в Идлибе погибло не менее 15 человек // LIFE. 2019. URL: https://life.ru/p/1194727

3 Теракт в поселении Бейт-Эль: 7-летний мальчик получил огнестрельное ранение // Вести. 2019. https://www.vesty.co.il/articles/0,7340,L-5484404,00.html

Ранжирование тегов

Данный подход к разрешению концептуальной неоднозначности аналогичен описанному Е. В. Рахилиной и др. [3], согласно которому предложено изменить порядок значений лексемы, основываясь на корпусных данных в противоположность данным словарей, таким образом создав иерархию значений для разрешения семантической неоднозначности в Национальном корпусе русского языка. Мы, в свою очередь, предлагаем установить иерархию тегов (как концептуальных, так и неконцептуальных) для высокочастотных лексических единиц, для которых такая иерархия может быть построена на основе корпусных данных.

Для иллюстрации данного метода мы отобрали три высокочастотных единицы, которые встречаются в «золотом» корпусе с каждым из автоматически присваиваемых тегов (выполнение этого условия необходимо для построения полной иерархии): *человек*, *полиция* и *автомобиль*.

Совокупная частота слова *человек*, которому может быть присвоен один из тегов PO-P-Z-A (неконцептуальное существительное / Object of attack / Terrorism-agent / Consequences) во всех его словоформах, составляет 186. Из них 176 относятся к Consequences (тег P), 7 — к концепту Object of attack (Z), 1 — к концепту Terrorism-agent (A), 2 — к неконцептуальным существительным (PO). Таким образом, на основе этих данных иерархия тегов для слова *человек* будет выглядеть следующим образом: P, Z, PO, A. Аналогично, для слова *полиция* теги будут ранжированы в порядке RW (Counterterrorism, частота = 60), S (Source, 15), Z (Object of attack, 5); для слова *автомобиль* — C (Weapon, 33), Z (Object of attack, 11), P (Consequences, 4), PO (неконцептуальное существительное, 1).

Рассмотренный метод имеет два очевидных недостатка. Во-первых, значительное количество лексических единиц в корпусе имеет недостаточную частоту встречаемости, чтобы построить иерархию тегов. Во-вторых, частоты некоторых тегов, присвоенных одной лексической единице, либо равны, либо приблизительно равны, и по этой причине ни одному из этих тегов нельзя отдать предпочтение. Таким образом, метод подходит только для ограниченного числа лексических единиц, у которых наблюдается значительный перевес в частоте одного из тегов (где можно определить основное концептуальное значение). В остальных случаях положение тега в иерархии для конкретной единицы может выступать дополнительным индикатором разрешения концептуальной неоднозначности в пользу одного из тегов.

Совместная встречаемость тегов

С целью проверки данного метода на основе «золотого» корпуса были построены конкордансы для всех единичных и множественных тегов. Рассмотрим применение метода для разрешения неоднозначности мультитега S-Z-RW, присваиваемого слову *полиция* в результате автоматического аннотирования в следующих предложениях:

1. {Полиция}$^{\sim S\sim Z\sim RW}$ {начала расследование}$^{\sim RW}$ {после}$^{\sim O\sim PO\sim S\sim Z\sim BW}$ {того}$^{\sim O}$, {как}$^{\sim O}$ {в}$^{\sim O}$ {четырех мечетях Бирмингема}$^{\sim L}$ {были выбиты}$^{\sim P}$ {окна}$^{\sim PO\sim P}$.[1]
2. {Полиция}$^{\sim S\sim Z\sim RW}$ {сообщила}$^{\sim D}$ {о}$^{\sim O}$ {стрельбе}$^{\sim T\sim M}$ {в}$^{\sim O}$ {Лондондерри}$^{\sim L}$ {в}$^{\sim O}$ {Северной Ирландии}$^{\sim L\sim N}$.[2]
3. {Кумар}$^{\sim UNK}$ {добавил}$^{\sim D}$, {что}$^{\sim O}$ {это}$^{\sim O\sim DEF}$ {самое кровавое нападение}$^{\sim T}$ {на}$^{\sim O}$ {полицию}$^{\sim S\sim Z\sim RW}$ {в}$^{\sim O}$ {Кашмире}$^{\sim L}$.[3]

В поисках возможных индикаторов для разрешения неоднозначности мультитега S-Z-RW было изучено его ближайшее окружение на два шага влево и два шага вправо. В первом предложении ближайшим концептуальным мультитегом к исходному мультитегу является RW в правом контексте, во втором — D в правом контексте, в третьем — T в левом контексте и L в правом. В таблице 5 показаны частоты последовательностей указанных тегов и одного из тегов S-Z-RW в конкордансе, по которым можно сделать следующие выводы: в первом предложении концептуальная неоднозначность с большей долей вероятности может быть разрешена в пользу RW, во втором — в пользу S, в третьем — в пользу Z.

Таблица 5

Частота последовательностей тегов в конкордансе

Table 5

Frequency of the tag sequences in the concordance

Пример	**1**			**2**			**3**					
Теги	S RW	Z RW	RW RW	S D	Z D	RW D	T O S	T O Z	T O RW	S O L	Z O L	RW O L
Частота	23	15	*83*	*137*	15	31	9	*106*	9	13	*54*	13

Итак, несмотря на то, что результаты применения этого метода могут быть не всегда точными (в частности, из-за одинаковой частоты совместной встречаемости тегов), он может быть использован в качестве одной из метрик для вычисления вероятности разрешения неоднозначности в пользу того или иного тега.

Позиционный подход

Новостные интернет-сообщения могут иметь различную структуру, но принято считать, что наиболее эффективным и частым способом представления

[1] Неизвестные выбили окна в четырех мечетях в Бирмингеме // ТАСС. 2019. URL: https://tass.ru/proisshestviya/6244720

[2] Полиция сообщила о стрельбе в Лондондерри в Северной Ирландии // ТАСС. 2019. URL: https://tass.ru/proisshestviya/6070869

[3] При взрыве самодельной бомбы в Индии погибли 40 полицейских // Газета.ru. 2019. URL: https://www.gazeta.ru/social/news/2019/02/14/n_12640261.shtml

информации в таких текстах является перевернутая пирамида [5], в которой наиболее важные данные расположены в начале сообщения, а наименее важные — в конце. На основе этого положения нами была выдвинута гипотеза, что в новостных сообщениях, имеющих структуру перевернутой пирамиды, предложения, расположенные ближе к началу сообщения, содержат бóльшую долю концептуальных тегов, чем предложения, расположенные ближе к концу. Данный метод близок классическому позиционному методу для определения ключевых слов, предложенному Г. Эдмундсоном [7]; концептуальные теги в этом случае можно рассматривать как ключевые слова.

В целях проверки гипотезы «золотой» корпус новостных интернет-сообщений (длиной от 3 до 27 предложений, модальная и медианная длина — 8 и 9 предложений соответственно) был разделен на предложения, после чего были сформированы позиционные подкорпусы предложений. Для каждого подкорпуса была подсчитана относительная частота концептуальных тегов и построен линейный график (рис. 1), отражающий изменение доли концептуальных тегов в зависимости от позиции предложения в сообщении.

Можно заметить, что кривая непрерывно снижается с отметки 77 до 56% в предложениях 1-7, затем наблюдается незначительный рост и последующее снижение вплоть до 42% в предложении 18, после чего кривая становится нестабильной. Это можно объяснить тем, что несколько сообщений в корпусе значительно длиннее остальных, и данные для предложений 19-27 построены на основе малого количества сообщений (рис. 2). Очевидно, данных для последних девяти предложений недостаточно, чтобы делать статистические выводы. Эта часть графика выделена на рис. 1 и 3 пунктиром.

Следует также отметить, что сообщения были взяты из разных источников, что, безусловно, отражается на их длине и структуре. Так, в последних абзацах

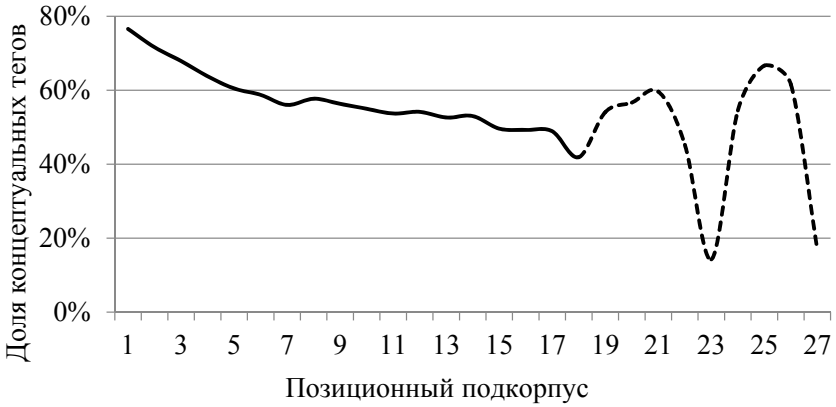

Рис. 1. Относительная частота концептуальных тегов (ось x) в зависимости от позиции предложения в тексте (ось y)

Fig. 1. Relative frequency of concept tags (axis x) depending on sentence position in the text (axis y)

некоторых сообщений присутствует информация о террористических актах, которые не имеют прямого отношения к основной теме сообщения, отсюда идет повышение кривой относительной частоты концептуальных тегов.

Проведенный эксперимент в целом подтверждает гипотезу о зависимости доли концептуальных тегов от позиции в сообщении, и следовательно полученные данные могут быть использованы для разрешения неоднозначности мультитегов, в состав которых входят и концептуальные, и неконцептуальные теги.

В ходе эксперимента были также получены дополнительные данные об изменении доли отдельных тегов в зависимости от позиции в сообщении. Например, тег T чаще встречается в начале текста, и затем его доля постепенно снижается, в то время как доля тега RW растет (рис. 3). Это объясняется тем, что информация о террористическом акте первостепенна для такого типа новостных

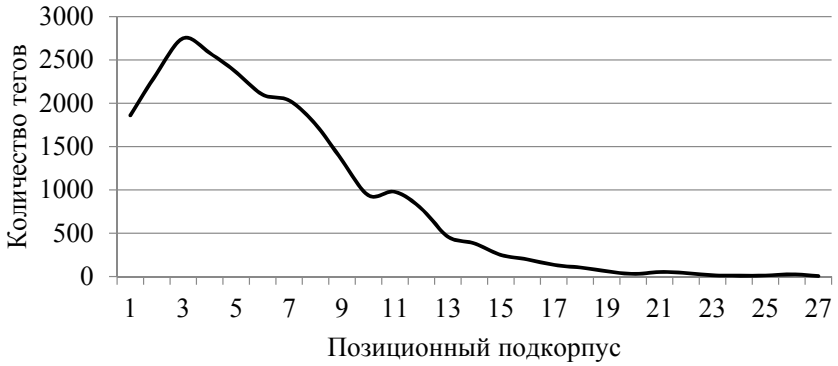

Рис. 2. Общее количество тегов (ось x) в позиционном подкорпусе (ось y)

Fig. 2. Total number of tags (axis x) in a positional subcorpus (axis y)

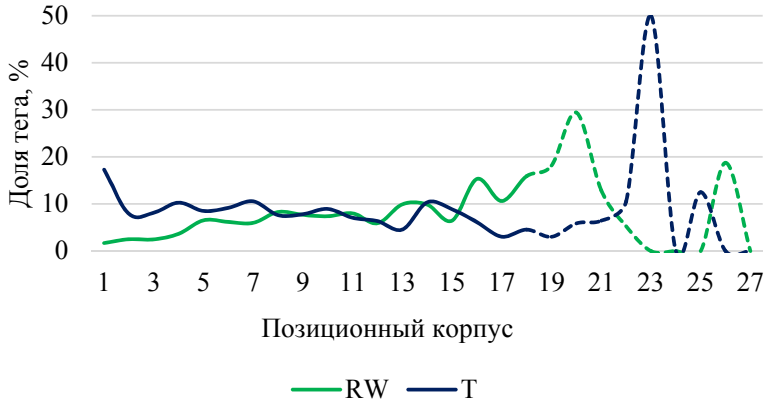

Рис. 3. Относительная частота тегов T и RW (ось x) в зависимости от позиции предложения в тексте (ось y)

Fig. 3. Relative frequency of the tags T and RW (axis x) depending on sentence position in the text (axis y)

сообщений, а информация о контртеррористических мерах является дополнительной, вследствие чего ее размещают ближе к концу текста. Данные о разнице в распределении тегов T и RW могут быть полезны для разрешения концептуальной неоднозначности таких слов, как *рейд*, *операция*, *стрельба* и т. д.

Заключение

Концептуальное аннотирование — это сложный процесс, одна из проблем которого связана с отображением лексической единицы на несколько онтологических концептов. Соответственно, перед исследователями стоит важная задача — определить, в каких случаях необходимо разрешение неоднозначности и какие методы для этого требуются.

В результате проведенного исследования было выявлено, что в зависимости от необходимости разрешения неоднозначности явление отображения лексической единицы на несколько концептов делится на случаи концептуальной синкретичности и концептуальной неоднозначности. Концептуальная синкретичность проявляется в следующих случаях:

1) если каждый компонент многокомпонентной лексической единицы относится к отдельному концепту;
2) если лексическая единица содержит несколько релевантных семантических компонентов;
3) если лексическая единица в данном контексте обладает двумя концептуальными значениями, которые не противоречат друг другу.

Прочие случаи присвоения одной лексической единицы нескольким концептам следует считать проявлением концептуальной неоднозначности, требующей разрешения.

Дальнейший анализ помог выявить пять источников концептуальной неоднозначности: частеречная омонимия, лексическая и синтаксическая неоднозначность, множественность концептуальных значений и экстралингвистический контекст. При этом концептуальная неоднозначность может проистекать из нескольких источников одновременно. На основе анализа текстов новостных интернет-сообщений о террористических актах на русском языке было установлено, что концептуальная неоднозначность чаще всего вызывается множественностью концептуальных значений. Согласно результатам анализа, особого внимания заслуживают случаи, в которых источником концептуальной неоднозначности является экстралингвистический контекст, т. к. их проблематично концептуально аннотировать не только автоматически, но и вручную, а следовательно, такие явления требуют более детального рассмотрения.

Кроме того, в ходе исследования были проанализированы три метода разрешения концептуальной неоднозначности на основе корпусных данных: метод ранжирования тегов, метод совстречаемости тегов и позиционный метод. Несмотря на наличие у каждого метода недостатков, их можно использовать для разрешения определенных типов концептуальной неоднозначности. В то же время их комплексное применение может привести к более точным результатам.

В дальнейшем мы рассчитываем изучить проблему концептуальной неоднозначности более подробно. В частности, планируется провести сопоставительный анализ применения данных методов для разрешения концептуальной неоднозначности на всех языках проекта (русском, английском и французском) и определить, являются ли методы лингвоспецифическими или универсальными.

СПИСОК ЛИТЕРАТУРЫ

1. Иорданская Л. Н. Автоматический синтаксический анализ / Л. Н. Иорданская. Новосибирск: Наука Сиб. отд-ние, 1967. Том 1. 231 с.
2. Поляков В. Н. Использование технологий, ориентированных на лексическое значение, в задачах поиска и классификации / В. Н. Поляков // Проблемы прикладной лингвистики: сборник статей. 2004. Вып. 2. С. 101-117.
3. Рахилина Е. В. Многозначность как прикладная проблема: семантическая разметка в национальном корпусе русского языка / Е. В. Рахилина, Б. П. Кобрицов, Г. И. Кустова, О. Н. Ляшевская, О. Ю. Шеманаева // Труды международной конференции «Диалог 2006». 2006. С. 445-450.
4. Федеральная служба безопасности РФ. Единый федеральный список организаций, в том числе иностранных и международных организаций, признанных в соответствии с законодательством Российской Федерации террористическими (на 5 июля 2019 г.). URL: http://www.fsb.ru/fsb/npd/terror.htm (дата обращения: 22.09.2020).
5. DeAngelo T. I. Looking for efficiency: how online news structure and emotional tone influence processing time and memory / T. I. DeAngelo, N. S. Yegiyan // Journalism and Mass Communication Quarterly. 2019. No. 96 (2). Pp. 385-405.
6. Djemaa M. Corpus annotation within the french framenet: a domain-by-domain methodology / M. Djemaa, M. Candito, Ph. Muller, L. Vieu // Proceedings of the 10th International Conference on Language Resources and Evaluation. 2016. Pp. 3794-3801.
7. Edmundson H. P. New methods in automatic extracting / H. P. Edmundson // Journal of the Association for Computing Machinery. 1969. No. 16 (2). Pp. 264-285.
8. European Council. Council Decision (CFSP) 2019/1341 of 8 August 2019. URL: https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:32019D1341&from=en (дата обращения: 22.09.2020).
9. Guarino N. Introduction to Applied Ontology and Ontological Analysis / N. Guarino. 2012. URL: https://iaoa.org/isc2012/docs/AppliedOntology_OntologicalAnalysis.pdf (дата обращения: 22.09.2020).
10. Kim J. D. Corpus annotation for mining biomedical events from literature / J. D. Kim, T. Ohta, J. Tsujii // BMC Bioinformatics. 2008. No. 9. Pp. 9-10.
11. Nirenburg S. Ontological Semantics / S. Nirenburg, V. Raskin. Cambridge: MIT Press, 2004. 440 pp.
12. Palmer M. The proposition bank: an annotated corpus of semantic roles / M. Palmer, P. Gildea, P. Kingsbury // Computational Linguistics. 2005. No. 31 (1). Pp. 71-106.
13. Sheremetyeva S. On modelling domain ontology knowledge for processing multilingual texts of terroristic content / S. Sheremetyeva, A. Zinovyeva // Communications in Computer and Information Science. 2018. No. 859. Pp. 368-379.

14. Sheremetyeva S. Ontological analysis of e-news: a case for terrorism domain / S. Sheremetyeva, A. Zinoveva // Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction. 2019. Pp. 130-141.
15. Sheremetyeva S. Towards creating interoperable resources for conceptual annotation of multilingual domain corpora / S. Sheremetyeva // Proceedings of the 16th Joint ACL — ISO Workshop on Interoperable Semantic Annotation (ISA-16). 2020. Pp. 102-109.
16. Viju J. S. Concept interpretation by semantic knowledge harvesting / J. S. Viju // International Journal for Research in Applied Science and Engineering Technology (IJRASET). 2018. No. 6 (5). Pp. 477-484.
17. Wu H. Scientific impact at the topic level: a case study in computational linguistics / H. Wu, J. He, Y. Pei // Journal of the American Society for Information Science and Technology. 2010. Vol. 61. No. 11. Pp. 2274-2287.
18. Zagorulko M. J. System for semantic annotation of domain-specific text corpora / M. J. Zagorulko, I. S. Kononenko, E. A. Sidorova // Proceedings of the Annual International Conference “Dialogue” 2012. No. 11 (1). Pp. 674-685.

Anastasiia Yu. ZINOVEVA[1]
Svetlana O. SHEREMETYEVA[2]
Ekaterina D. NERUCHEVA[3]

UDC 811.161.1 + 004.82

THE ANALYSIS OF AMBIGUITY IN CONCEPTUAL ANNOTATION OF RUSSIAN TEXTS

[1] Postgraduate Student,
Department of Linguistics and Translation Studies,
South Ural State University (Chelyabinsk)
zinovevaaiu@bk.ru; ORCID: 0000-0002-7658-7376

[2] Dr. Sci. (Philol.), Professor
of the Department of Linguistics and Translation Studies,
South Ural State University (Chelyabinsk)
sheremetevaso@susu.ru

[3] Laboratory Assistant, Research and Education Centre
of Innovative Linguistic Technologies,
South Ural State University (Chelyabinsk)
neruchevaekaterina@mail.ru

Abstract

Properly annotated text corpora are an essential condition in constructing effective and efficient tools for natural language processing (NLP), which provide an operational solution to both theoretical and applied linguistic and informational problems. One of the main and the most complex problems of corpus annotation is resolving tag ambiguities on a specific level of annotation (morphological, syntactic, semantic, etc.).

This paper addresses the issue of ambiguity that emerges on the conceptual level, which is the most relevant text annotation level for solving informational tasks. Conceptual annotation is a special type of semantic annotation usually applied to domain corpora to address specific

Citation: Zinoveva A. Yu., Sheremetyeva S. O., Nerucheva E. D. 2020. "The Analysis of Ambiguity in Conceptual Annotation of Russian Texts". Tyumen State University Herald. Humanities Research. Humanitates, vol. 6, no. 3 (23), pp. 38-60.
DOI: 10.21684/2411-197X-2020-6-3-38-60

informational problems such as automatic classification, content and trend analyses, machine learning, machine translation, etc.

In conceptual annotation, text corpora are annotated with tags reflecting the content of a certain domain, which leads to a type of ambiguity that is different from general semantic ambiguity. It has both universal and language- and domain-specific peculiarities. This paper investigates conceptual ambiguity in a case study of a Russian-language corpus on terror attacks.

The research methodology combines automated and manual steps, comprising a) statistical and qualitative corpus analysis, b) the use of pre-developed annotation resources (a terrorism domain ontology, a Russian ontolexicon and a computer platform for conceptual annotation), c) ontological-analysis-based conceptual annotation of the corpus chosen for the case study, d) corpus-based detection and investigation of conceptual ambiguity causes, e) development and experimental study of possible disambiguation methods for some types of conceptual ambiguity.

The findings obtained in this study are specific for Russian-language terrorism domain texts, but the conceptual annotation technique and approaches to conceptual disambiguation developed are applicable to other domains and languages.

Keywords

Conceptual annotation, conceptual tagging, Russian corpus, conceptual ambiguity, case study, ontological analysis, multilingual domain ontology, terrorism.

DOI: 10.21684/2411-197X-2020-6-3-38-60

REFERENCES

1. Iordanskaya L. N. 1967. Automatic Syntactic Analysis. Vol. 1. Novosibirsk: Nauka. 231 pp. [In Russian]
2. Polyakov V. N. 2004. "Using lexical meaning-oriented technologies in search and classification tasks". Problemy prikladnoy lingvistiki. Sbornik statey, no. 2, pp. 101-117. [In Russian]
3. Rakhilina E. V., Kobritsov B. P., Kustova G. I., Lyashevskaya O. N., Shemanayeva O. J. 2006. "Semantic Ambiguity as an Application-Oriented Problem: Word Class Tagging in the RNC". Computational Linguistics and Intellectual Technologies. Proceedings of the International Workshop Dialogue 2006 (Moscow), pp. 445-450 [In Russian]
4. Federal Security Service of the Russian Federation. 2019. The Combined Federal List of Organizations, including Foreign and International Organizations, Recognized as Terrorist in accordance with the Law of the Russian Federation. Accessed 22 September 2020. http://www.fsb.ru/fsb/npd/terror.htm [In Russian]
5. DeAngelo T. I., Yegiyan N. S. 2009. "Looking for efficiency: how online news structure and emotional tone influence processing time and memory". Journalism and Mass Communication Quarterly, no. 96 (2), pp. 385-405.
6. Djemaa M., Candito M., Muller Ph., Vieu L. 2016. "Corpus annotation within the french framenet: a domain-by-domain methodology". Proceedings of the 10th International Conference on Language Resources and Evaluation, pp. 3794-3801.

7. Edmundson H. P. 1969. “New methods in automatic extracting”. Journal of the Association for Computing Machinery, no. 16 (2), pp. 264-285.
8. European Council. 2019. Council Decision (CFSP) 2019/1341 of 8 August 2019. Accessed 22 September 2020. https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:32019D1341&from=en
9. Guarino N. 2012. Introduction to Applied Ontology and Ontological Analysis. Accessed 22 September 2020. https://iaoa.org/isc2012/docs/AppliedOntology_OntologicalAnalysis.pdf
10. Kim J. D., Ohta T., Tsujii L. 2008. “Corpus annotation for mining biomedical events from literature”. BMC Bioinformatics, no. 9, pp. 9-10.
11. Nirenburg S., Raskin V. 2004. Ontological Semantics. Cambridge: MIT Press. 440 pp.
12. Palmer M., Gildea P., Kingsbury P. 2005. “The proposition bank: an annotated corpus of semantic roles”. Computational Linguistics, no. 31 (1), pp. 71-106.
13. Sheremetyeva S., Zinovyeva A. 2018. “On modelling domain ontology knowledge for processing multilingual texts of terroristic content”. Communications in Computer and Information Science, no. 859, pp. 368-379.
14. Sheremetyeva S., Zinoveva A. 2019. “Ontological analysis of e-news: a case for terrorism domain”. Proceedings of the 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction, pp. 130-141.
15. Sheremetyeva S. 2020. “Towards creating interoperable resources for conceptual annotation of multilingual domain corpora”. Proceedings of the 16th Joint ACL — ISO Workshop on Interoperable Semantic Annotation (ISA-16), pp. 102-109.
16. Viju J. S. 2018. “Concept interpretation by semantic knowledge harvesting”. International Journal for Research in Applied Science and Engineering Technology (IJRASET), no. 6 (5), pp. 477-484.
17. Wu H., He. J, Pei Y. 2010. “Scientific impact at the topic level: a case study in computational linguistics”. Journal of the American Society for Information Science and Technology, vol. 61, no. 11, pp. 2274-2287.
18. Zagorulko M. J., Kononenko I. S., Sidorova E. A. 2012. “System for semantic annotation of domain-specific text corpora”. Proceedings of the Annual International Conference “Dialogue”, no. 11 (1), pp. 674-685.