Концептуальная неоднозначность в англоязычных текстах о терроризме: причины возникновения и методы разрешения

А. Ю. Зиновьева

***Аннотация*—Актуальные исследования в области автоматической обработки текста нередко затрагивают тему семантизации контента (в частности неструктурированных текстовых потоков), которая достигается посредством семантической разметки или ее вариации, основанной на концептуальной модели и ориентированной на ограниченную предметную область, — концептуальной разметки. В процессе автоматической концептуальной разметки возникает концептуальная неоднозначность, которая проявляется во множественных связях между лексической единицей и концептами онтологии. В статье рассматриваются причины возникновения концептуальной неоднозначности в текстах ограниченной предметной области на материале новостных сообщений о терактах на английском языке. Предлагаются и анализируются возможные количественные методы разрешения такой неоднозначности, основанные на корпусных данных. Делается предположение о пользе применения рассмотренных методов при автоматизированном снятии неоднозначности с участием человека.**

***Ключевые слова*—английский язык, концептуальная неоднозначность, концептуальная разметка, терроризм.**

I. Введение

В современных исследованиях по автоматической обработке текста на естественном языке большое внимание уделяется семантической разметке, под которой в общем виде понимается процедура обогащения контента семантической информацией. Конкретное понимание семантической разметки варьируется от определения конкретного значения многозначного слова на основе словаря или онтологии [1] до выявления универсальных семантических свойств слов на основе лексической классификации [2] и обнаружения семантических отношений в тексте [3]. Частным случаем семантической разметки является концептуальная разметка, выполняемая на основе онтологии или иной концептуальной модели, как правило, ориентированная на ограниченную предметную область.

Концептуальная разметка может выполняться вручную [4]–[6] или с использованием средств автоматизации [7] вплоть до полностью автоматической разметки [8]. Ручная разметка, обеспечивающая при корректном выполнении высокое качество результата, тем не менее требует предварительного обучения разметчиков и связана с большими затратами времени. Частично автоматизированная разметка не решает этих проблем в полной мере [9], попытки же радикальной автоматизации процесса, в свою очередь, приводят к концептуальной неоднозначности [10], т.е. возможности связывания одной лексической единицы (в виде словоформы, слова, словосочетания) с несколькими концептами онтологии. Возможным решением проблемы является автоматическая разметка текста с последующим автоматизированным снятием неоднозначности, в процессе которого компьютер предлагает наиболее вероятный вариант разрешения неоднозначности, однако окончательный выбор делает человек.

Насколько нам известно, проблеме концептуальной неоднозначности, в частности причинам ее возникновения в текстах ограниченной предметной области и возможным способам разрешения, посвящено мало научных работ (см. [11], где упомянута проблема неоднозначности концептуальных тегов в сфере информационной безопасности, а также [8], где описан алгоритм машинного обучения для снятия концептуальной неоднозначности). Цель настоящей работы — изучить причины концептуальной неоднозначности в англоязычных медиатекстах предметной области «Терроризм» и проанализировать возможные методы ее разрешения, которые могут быть полезны для автоматизированного снятия неоднозначности с участием человека.

II. Ресурсы и процедура эксперимента

A. Ресурсы и входные данные эксперимента

Эксперимент, описанный в данной статье, был проведен с привлечением следующих ресурсов:

- независимой от конкретного языка онтологии, построенной для обработки текстов о терроризме на нескольких языках (подробнее см. [12]);
- англоязычного онтолексикона (лексикона, элементы

Статья получена 11 августа 2020.

А. Ю. Зиновьева, Южно-Уральский государственный университет, Институт лингвистики и международных коммуникаций, Челябинск, Россия (email: zinovevaaiu@bk.ru).

которого связаны с концептами онтологии);
- прототипа инструмента для автоматической концептуальной разметки;
- неразмеченного англоязычного корпуса объемом 25 000 словоупотреблений, составленного из 107 новостных интернет-сообщений о терактах.

Используемая в данном исследовании предметная онтология представлена в формализме онтологии Mikro-Kosmos, основными классами которой являются OBJECT, EVENT и PROPERTY [13]. Онтология терроризма является трехуровневой и на момент эксперимента содержит 105 концептов классов OBJECT и EVENT, в том числе 23 концепта первого уровня, 49 концептов второго уровня и 33 концепта третьего уровня, а также 20 концептов класса PROPERTY. Некоторые примеры онтологических концептов первого уровня с дефинициями показаны в таблице I. Наименования концептов онтологии традиционно представлены на английском языке, при этом значение каждого концепта определяется его дефиницией и наполнением. Так, дефиниция концепта WEAPON формулируется следующим образом: ‘оружие или подобные ему объекты, используемые для совершения теракта’. На основании этой дефиниции мы можем отнести к концепту WEAPON не только очевидные лексические единицы (*assault rifle* ‘автомат’, *knife* ‘нож’), но и другие: *nail bomb* ‘самодельная бомба из гвоздей’, *explosive* ‘взрывчатка’, *truck* ‘грузовик’, *bullet* ‘пуля’.

Таблица I
Некоторые концепты онтологии с дефинициями

Концепт	Дефиниция
COUNTER-TERRORISM	Меры по противодействию терроризму и борьбе с терроризмом, а также учреждения и лица, предпринимающие эти меры
TERROR ATTACK	Нападение, совершаемое террористом или группой террористов для устрашения населения и (или) достижения каких-л. политических целей
TIME	Время, дата и сопутствующие обстоятельства теракта, указывающие на его временную соотнесенность
WEAPON	Оружие или подобные ему предметы, используемые для совершения теракта

Как было упомянуто ранее, онтология терроризма многоязычна и имеет связи с тремя онтолексиконами: на русском, английском и французском, что обеспечивает возможность обрабатывать тексты на этих языках. В качестве материала исследования был выбран англоязычный корпус с целью поиска и изучения характерных для английского языка причин концептуальной неоднозначности. Английский онтолексикон содержит лексические единицы (именные и глагольные группы, прилагательные, наречия, предложные группы) длиной от одного до десяти компонентов. Фрагмент онтолексикона представлен в таблице II, откуда видно, что некоторые лексические единицы могут относиться к нескольким концептам (например, слово *operation* может быть отнесено к концептам COUNTERTERRORISM и TERROR ATTACK) в зависимости от контекста.

Прототип инструмента концептуальной разметки основан на представленных лексических и онтологических знаниях. С его помощью была выполнена разметка англоязычного корпуса тегами концептов, после чего вручную было проведено постредактирование (снятие концептуальной неоднозначности) и получена «золотая» разметка корпуса. Затем для выявления причин концептуальной неоднозначности было проведено сравнение автоматически размеченного и «золотого» корпуса.

Таблица II
Фрагмент английского онтолексикона

Концепт	Лексические единицы
COUNTER-TERRORISM	air force night raid, all police forces, counter-terrorism squad, eradicate terrorism, operation
TERROR ATTACK	attempted hijacking, deadly shooting rampage, explosion, hostage taking, intimidation act, knife attack, mass shooting, operation
TIME	1 a.m., a few days ago, after the sunset, during the Friday prayer, on Wednesday night
WEAPON	armored car bomb, assault rifle, bomb-laden vehicle, combustible liquid, homemade mortar, incendiary mixture, lorry, vest

B. Схема и процедура разметки

Для удобства были отобраны 23 онтологических концепта первого уровня, закодированные при помощи тегов: A = TERRORISM-AGENT ‘террорист’, BW = TIME ‘время’, C = WEAPON ‘оружие’, CR = CLAIM RESPONSIBILITY ‘брать на себя ответственность’, D = DECLARE ‘заявлять’, DA = DIRECTION OF ATTACK ‘направление теракта’, E = OTHER TERRORIST ACTIVITIES ‘другая террористическая деятельность’, EW = CAUSE ‘причина’, HA = HAVE WEAPON ‘иметь оружие’, I = ASSUMPTION ‘предположение’, K = TERRORISM-AGENT’S PLAN ‘планы террористов’, L = LOCATION ‘место’, M = SCALE OF ATTACK ‘масштаб теракта’, N = NATION ‘национальность, страна’, OW = OTHER ‘другое’, P = CONSEQUENCES ‘последствия’, RW = COUNTERTERRORISM ‘контртеррористическая деятельность’, S = SOURCE ‘источник’, T = TERROR ATTACK ‘теракт’, UW = TERRORIST ORGANIZATION ‘террористическая организация’, X = GOAL OF ATTACK ‘цель теракта’, Y = REACTION ‘реакция общественности’, Z = OBJECT OF ATTACK ‘объект теракта’.

Кроме того, был введен ряд тегов для лексических единиц, не связанных с концептами онтологии по крайней мере в некоторых контекстах: различных предикатов (B, R, U), именных групп (PO), прилагательных, наречий, именованных сущностей (O), чисел (Num), детерминативов (DEF) и неизвестных лексических единиц, отсутствующих в лексиконе (UNK). Такое решение было принято, чтобы избежать присвоения концептуального тега лексической единице, нерелевантной для

предметной области в данном контексте.

По разработанным нами правилам разметки лексической единице может присваиваться либо один тег, либо несколько (мультитег), но только в том случае, если у этой единицы в конкретном контексте наблюдается синкретичность концептуального значения (концептуальная синкретичность). Под концептуальной синкретичностью мы понимаем наличие у единицы двух и более непротиворечивых концептуальных значений в контексте или вне его. Например, в предложении *At least 15 people were killed in an explosion that hit the rebel-held city of al-Bab in northern Syria* ('По крайней мере 15 человек погибли в результате взрыва в захваченном повстанцами городе Эль-Баб на севере Сирии'), разметка слова *Syria* тегами N и L оправдана, поскольку Сирия одновременно является и страной (NATION), и местом теракта (LOCATION).

Концептуальная синкретичность отличается от концептуальной неоднозначности тем, что первая не требует разрешения. Более того, концептуально синкретичные лексические единицы представляют потенциальные дочерние концепты онтологии или дают информацию о свойствах концепта. Так, в примере выше мультитег L-N указывает на то, что концепт NATION может быть связан с другим концептом (из контекста ясно, что таким концептом является TERROR ATTACK) отношением LOCATION-OF. Мультитеги S-N (*Turkish media* 'турецкие СМИ') и Z-N (*Iranians* 'иранцы') показывают, что источник сообщения (SOURCE) и объект теракта (OBJECT OF ATTACK) могут относиться к определенной стране или нации. Мультитег A-I показывает, что некто, возможно, является террористом (AGENT), но это не подтверждено (ASSUMPTION). Таким образом, концепты NATION и ASSUMPTION могут рассматриваться как атрибуты.

Можно выделить три случая, в которых имеет смысл говорить о концептуальной синкретичности:

1. Каждый компонент многокомпонентной лексической единицы может быть отнесен к отдельному концепту, при этом лексическую единицу нельзя разделить из-за тесной смысловой или структурной связи между компонентами. Например, в трехкомпонентной лексической единице *Afghan security forces* 'силы безопасности Афганистана' лексема *Afghan* соотносится с концептом NATION, а *security forces* — с концептом COUNTER-TERRORISM, в результате чего всей лексической единице присваивается синкретичный мультитег RW-N.

2. Лексическая единица содержит несколько релевантных семантических компонентов, в том числе вне указанного контекста. Примером является существительное *suspect* 'подозреваемый', которое по определению обозначает человека, подозреваемого в совершении преступления, и, следовательно, логично соотносится с концептами TERRORISM-AGENT и ASSUMPTION.

3. В данном контексте лексическая единица имеет два и более непротиворечивых концептуальных значения. В предложении *The terrorist was killed by soldiers* 'Террорист был убит солдатами' слово *killed* соотносится с концептами P (CONSEQUENCES) и RW (COUNTER-TERRORISM), так как передает информацию о контртеррористическом действии и последствиях для террориста.

Остальные случаи встречаемости мультитегов в разметке являются проявлением концептуальной неоднозначности и должны быть разрешены.

III. Результаты

A. Основные результаты анализа корпуса

В результате исследования в автоматически размеченном корпусе обнаружено 338 уникальных тегов, из которых 307 мультитегов. В «золотом» корпусе найдено 125 уникальных тегов, из которых 94 мультитега. Мы также вычислили относительную частоту лексических единиц с различными тегами и определили, что в автоматически размеченном корпусе отношение всех (концептуальных и неконцептуальных) мультитегов ко всем тегам составляет 28 %, а отношение мультитегов, содержащих хотя бы один концептуальный тег, ко всем концептуальным тегам, составляет 46 %. В «золотом» корпусе аналогичные показатели значительно ниже: 6 и 15 % соответственно, при том, что все мультитеги в нем вызваны концептуальной синкретичностью.

На рис. 1–2 показано распределение мультитегов (т.е. размеченных ими лексических единиц) в автоматически размеченном и «золотом» корпусах. За 100 % в обоих случаях принимается общее количество размеченных тегами лексических единиц.

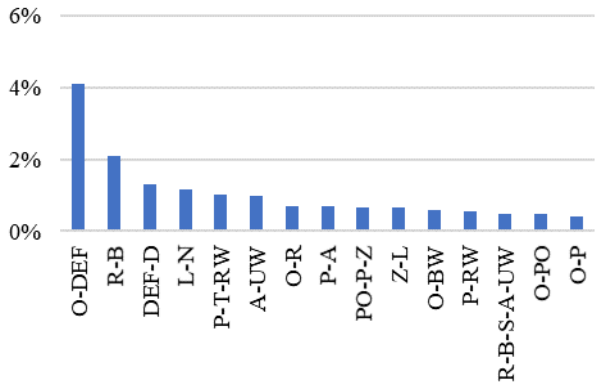

Рис. 1. Доля мультитегов в автоматическом корпусе

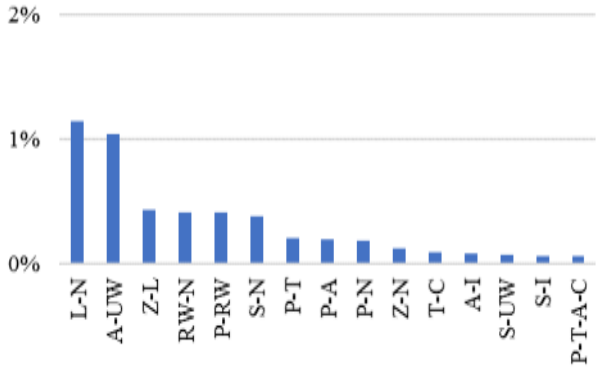

Рис. 2. Доля мультитегов в «золотом» корпусе

Некоторые мультитеги постоянно автоматически присваиваются только одной лексической единице (мультитег DEF-D всегда присваивается словоформе *said*), другие же характерны для многих (например, мультитег P-

RW в корпусе присвоен 43 различным словоформам: *arrested, injured, injuring, rescue, were rescued* и др.)

Полученные результаты свидетельствуют о том, что концептуальная неоднозначность должна рассматриваться как серьезная проблема концептуальной разметки текстов в ограниченных предметных областях, поскольку, во-первых, требующие разрешения мультитеги имеют высокую частоту встречаемости, и, во-вторых, количество уникальных мультитегов высоко, а их природа различна, что может говорить о необходимости применения различных методов разрешения неоднозначности.

B. Причины концептуальной неоднозначности

В результате сопоставительного анализа автоматически размеченного и «золотого» корпусов обнаружено четыре причины концептуальной неоднозначности.

Частеречная омонимия. При частеречной омонимии концептуальная неоднозначность возникает из-за совпадения отдельных словоформ у разных лексических единиц. Такой неоднозначности подвержены как релевантные, так и нерелевантные для предметной области однокомпонентные лексические единицы (см. таблицу III). Как видно из таблицы, дополнительным фактором появления неоднозначности является то, что используемый нами инструмент разметки не различает строчные и прописные буквы.

Таблица III
Концептуальная неоднозначность, вызванная частеречной омонимией

Лексическая единица	Часть речи	Начальная форма	Теги
act(s)	сущ.	act	T
	глагол	act	R
suspect	сущ.	suspect	A-I
	глагол	suspect	I
may	сущ.	May	BW
	глагол	may	I
us	мест.	we	O
	сущ.	US	N / RW-N / S-N / Z-N / L-N

Лексическая неоднозначность. Под лексической неоднозначностью, как правило, понимается способность лексической единицы иметь две и более интерпретации, вызванная полисемией или омонимией [14]. Например, слово *release* в рассматриваемом корпусе может иметь одно из двух значений: ‘освобождать, отпускать’ (о заложниках или террористах) или ‘выпускать, делать публичным’ (о заявлениях). Выбор одного из этих значений в контексте не представляет сложности для человека (ср. *Hostages were released* ‘заложники были освобождены’ и *A statement was released* ‘заявление было сделано’), однако при автоматической разметке возникают проблемы: в частности, слово *release* автоматически размечается мультитегом RW-D (COUNTERTERRORISM / DECLARE).

Примеров лексической неоднозначности, которая влечет за собой неоднозначность концептуальную, в рассматриваемом корпусе значительное количество. Ниже представлен далеко не исчерпывающий список:

Be directed — 1) (о теракте) ‘быть направленным на какой-л. объект’ (DIRECTION OF ATTACK); 2) (о террористе) ‘быть направляемым, получать советы, какую-л. информацию’ (OTHER TERRORIST ACTIVITIES).

Be located — 1) ‘быть расположенным’ (LOCATION); 2) (о взрывном устройстве) ‘быть обнаруженным’ (Counterterrorism).

Body — 1) ‘труп’ (CONSEQUENCES); 2) ‘организация’ (Counterterrorism).

Station — 1) ‘автобусная или железнодорожная станция’ (Location); 2) ‘радиостанция’ (SOURCE).

Underground cell — 1) ‘небольшая группировка в составе террористической организации’ (TERRORIST organization); 2) ‘небольшое помещение в тюрьме, расположенное под землей’ (COUNTERTERRORISM).

Khorasan — 1) ‘регион на Ближнем Востоке’ (LOCATION); 2) ‘ячейка Исламского государства, расположенная в Хорасане’ (TERRORIST ORGANIZATION); 3) ‘предполагаемая группа высокопоставленных членов Аль-Каиды, действующая в Сирии’ (TERRORIST ORGANIZATION).

Множественность концептуальных значений. В этом случае лексическим единицам, имеющим одинаковую форму и лексическое значение, ставятся в соответствие несколько концептов, зачастую взаимоисключающих, при этом не все из этих концептов оказываются релевантными в конкретном контексте. Более того, некоторые из таких единиц в ряде контекстов могут быть вообще не связаны с предметной областью. По существу, такие случаи не являются множественностью *концептуальных* значений, но мы относим их в эту категорию для удобства, поскольку фактически в работе рассматривается неоднозначность тегов, а неконцептуальные лексические единицы также размечаются ими.

Приведем в качестве примера множественности концептуальных значений слово *children* в предложениях:

1. Six people were killed, 47 got injured, most of whom students, *children* (= CONSEQUENCES) and women.

2. He has actively recruited Australian men, women and *children* (= TERRORISM-AGENT).

3. As well as locking down schools, police have reportedly cleared Cathedral Square, where *children* (= неконцептуальная единица) were staging a climate change protest.

4. A “henna night” was underway in the streets, attended mostly by women and *children* (= OBJECT OF ATTACK).

Неоднозначности такого типа встречаются в корпусе наиболее часто, причем среди как однокомпонентных, так и многокомпонентных лексических единиц.

Экстралингвистический контекст. Иногда концептуальная неоднозначность является следствием различия точек зрения на какой-либо вопрос, например, относится ли некая организация к террористическим. Так, словосочетание *Dogon group* автоматически размечается тегами A (TERRORISM-AGENT) и RW (COUNTERTERRORISM), так как считается, что догоны ответственны за несколь-

ко терактов против народа фулани в Мали, при этом сами фулани поддерживают исламистов, поэтому в какой-то мере действия догонов могут быть расценены как антитеррористические.

Подобные случаи имеют невысокую частоту в корпусе, но они достаточно значимы, поскольку снятие такой неоднозначности проблематично не только для компьютера, но и для человека. Решение в данном случае должно приниматься обдуманно, возможно, на основе существующего списка террористических организаций. Проблема, однако, состоит в выборе такого списка, поскольку в разных странах списки могут различаться.

Можно заключить, что проблема экстралингвистического контекста как причины концептуальной неоднозначности нетривиальна и требует обширного анализа, что не входит в задачи данного исследования.

Следует отметить, что иногда лексическая единица размечается мультитегом сразу по нескольким причинам. Например, слову *accused* присваивается мультитег R-P-A-I-D, появляющийся вследствие концептуальной синкретичности, частеречной омонимии и множественности концептуальных значений. Если *accused* — субстантивированное прилагательное, из представленных тегов релевантен только концептуально синкретичный мультитег A-I, если же это глагол в форме прошедшего времени, в зависимости от контекста необходимо выбрать один из вариантов: R, P, I или D. Случаи концептуальной неоднозначности такого рода часты в рассматриваемом корпусе. На наш взгляд, они должны разрешаться поэтапно.

C. Методы разрешения неоднозначности

В этом разделе рассмотрены три количественных метода разрешения концептуальной неоднозначности на основе корпусных данных. Применение методов иллюстрируется на примере двух предложений:

1. 85 killed in terrorist bomb attack in Iraq, 20 Iranians.

Автоматическая разметка: [85]-Num [killed]-P-T-RW [in]-O [terrorist bomb attack]-T-C [in]-O [Iraq]-L-N, [20]-Num [Iranians]-P-Z-N.

«Золотая» разметка: [85]-Num [killed]-P [in]-O [terrorist bomb attack]-T-C [in]-O [Iraq]-L-N, [20]-Num [Iranians]-P-N.

2. The other bomb was found in Giza near Al-Nahda square as security forces stopped it from detonating.

Автоматическая разметка: [The other]-DEF-P [bomb]-T-C [was found]-R-B-P-RW [in]-O [Giza]-L [near Al-Nahda square]-L, [as]-O [security forces]-Z-RW [stopped it from detonating]-RW.

«Золотая» разметка: [The other]-DEF [bomb]-C [was found]-RW [in]-O [Giza]-L [near Al-Nahda square]-L, [as]-O [security forces]-RW [stopped it from detonating]-RW.

Отметим, что оценка представленных методов не входит в задачи настоящего исследования.

Ранжирование тегов. В общих чертах ранжирование тегов похоже на метод, описанный в работе [2], где было предложено отсортировать лексические значения лексем на основе корпусных данных вместо данных словарей и таким образом установить иерархию значений для разрешения семантической неоднозначности в НКРЯ. Мы предлагаем построить аналогичную иерархию для тегов в мультитегах. Это может быть сделано двумя способами: с учетом и без учета лексической единицы, которой присваивается данный тег.

Вариант 1. Ранжирование тегов без учета лексики.

Рассмотрим этот вариант метода на примере высокочастотных мультитегов, имеющих в своем составе по крайней мере один концептуальный тег: DEF-D, L-N и P-T-RW (см. таблицу IV). Мультитег DEF-D в рассматриваемом корпусе во всех случаях разрешается в пользу тега D. Мультитег L-N в большинстве случаев проявляет синкретичный характер и, как следствие, не разрешается. Тем не менее в 9 % случаев L-N разрешается в пользу N и никогда — в пользу L. Это объясняется тем, что страна, национальность, в отличие от локации, являются неотъемлемыми атрибутами чего- или кого-либо. Мультитег P-T-RW имеет пять вариантов разрешения, из которых два — в пользу синкретичных тегов P-T и P-RW. Интересно, что P-T-RW никогда не разрешается в пользу T-RW, поскольку концепты, обозначаемые этими тегами, дизъюнктивны (некое действие не может *одновременно* быть терактом и контртеррористическим действием).

Таблица IV
Ранжирование тегов без учета лексики

Мультитег (частота)	Варианты разрешения (вероятность)
DEF-D (264)	*D:* 264 (1)
L-N (235)	*L-N:* 213 (0,906), *N:* 22 (0,093)
P-T-RW (203)	*P:* 104 (0,512), *T:* 35 (0,172), *P-T:* 27 (0,133), *P-RW:* 27 (0,133), *RW:* 10 (0,049)

Применив ранжирование тегов без учета лексики к контрольным предложениям, получаем следующий результат (в круглых скобках указана вероятность, с которой разрешение неоднозначности происходит в пользу указанного тега; жирным помечены ошибки):

[85]-Num [killed]-P *(0,512)* [in]-O [terrorist bomb attack]-T-C *(0,51)* [in]-O [Iraq]-L-N *(0,906)*, [20]-Num [Iranians]-P-N *(0,8)*.

[The other]-P *(0,65)* [bomb]-C *(0,51)* [was found]-RW *(0,857)* [in]-O [Giza]-L [near Al-Nahda square]-L, [as]-O [security forces]-RW *(0,7)* [stopped it from detonating]-RW.

В первом предложении неоднозначность снята корректно; во втором предложении мультитег DEF-P, присвоенный единице *the other*, ошибочно разрешен как P.

Вариант 2. Ранжирование тегов с учетом лексики.

В ряде случаев ранжирование тегов без учета лексики дает искаженные результаты. Если обратить внимание на лексику, которой присваиваются мультитеги, то можно заметить, что необходимость и вероятность разрешения неоднозначности в пользу того или иного тега зависит от лексемы и словоформы, в которой она представлена. Например, для лексемы *Pakistan* процент разрешения неоднозначности L-N в пользу N составляет 5 %, в то время как для *Italy* — 75 %. Мультитег P-T-RW, присваиваемый словоформам *shot* и *shooting*, в первом случае чаще всего разрешается в пользу P-RW (45 %), во втором — в пользу T (92 %).

В таблице V представлены самые высокочастотные словоформы, размечаемые мультитегами, и варианты разрешения их концептуальной неоднозначности в «золотом» корпусе.

Применение ранжирования тегов с учетом лексики к снятию неоднозначности в контрольных предложениях дает результаты, аналогичные предыдущим, при этом вероятность разрешения неоднозначности в пользу того или иного тега в некоторых случаях оказывается выше:

[85]-Num [killed]-P *(0,706)* [in]-O [terrorist bomb attack]-T-C *(1)* [in]-O [Iraq]-L-N *(1)*, [20]-Num [Iranians]-P-N *(1)*.

[The other]-P *(0,625)* [bomb]-C *(1)* [was found]-RW *(0,75)* [in]-O [Giza]-L [near Al-Nahda square]-L, [as]-O [security forces]-RW *(0,714)* [stopped it from detonating]-RW.

Таблица V
Ранжирование тегов с учетом лексики

Словоформа (частота)	Мультитег	Варианты разрешения (вероятность)
is (93)	R-B-S-A-UW	*R:* 85 (0,914), *A-UW:* 6 (0,065), *UW:* 2 (0,022)
people (81)	PO-P-S-Z	*P:* 57 (0,704), *PO:* 18 (0,222), *Z:* 5 (0,062), *S:* 1 (0,012)
killed (68)	P-T-RW	*P:* 48 (0,706), *P-T:* 10 (0,147), *P-RW:* 8 (0,118), *RW:* 2 (0,029)
police (55)	P-S-Z-RW	*RW:* 29 (0,527), *S:* 23 (0,418), *Z:* 2 (0,036), *P:* 1 (0,018)

В целом, оба варианта метода не лишены недостатков. Во-первых, значительное количество лексических единиц в корпусе имеют недостаточно высокую частоту встречаемости для исчерпывающего ранжирования тегов. Во-вторых, частоты нескольких тегов, присваиваемых одной лексической единице, могут быть одинаковы, вследствие чего предпочтение не может быть отдано одному их них. Однако применение данного метода при автоматизированном разрешении неоднозначности может упростить процесс выбора тега для разметчика.

Контекстный метод. Данный метод использует биграммную модель «золотого» корпуса без учета лексики. Все лексические единицы были предварительно удалены из корпуса с помощью регулярных выражений, и были оставлены только теги (в т.ч. синкретичные мультитеги). Кроме того, были добавлены маркеры начала и конца предложения <s> и </s>. Подсчет вероятности разрешения концептуальной неоднозначности в контрольных предложениях был произведен без сглаживания.

Для первого предложения существует 441 вариант снятия неоднозначности. Самая высокая вероятность (указана в круглых скобках) наблюдается у варианта

<s> Num P O **T** O L-N Num **P** </s> (3,19e-10).

В этом варианте предложения есть два недочета: мультитег T-C ошибочно разрешен в пользу T (разрешение неоднозначности не требуется: мультитег в данном случае синкретичен), P-Z-N ошибочно разрешен в пользу P (не хватает тега N, который никогда не должен удаляться при разрешении неоднозначности).

Для второго контрольного предложения существует 405 вариантов разрешения неоднозначности. Самая высокая вероятность обнаруживается у варианта

<s> DEF **T R** O L L O RW RW </s> (5,58e-10).

Этот вариант также содержит две ошибки: разрешение T-C в пользу T вместо C и разрешение R-B-P-RW в пользу R вместо RW.

Очевидно, что контекстный метод разрешения концептуальной неоднозначности не может основываться исключительно на статистических показателях. В некоторых случаях точность метода могли бы повысить простые правила, ограничивающие варианты разрешения неоднозначности. Так, введение правила о неудалении тега N уменьшает общее количество возможных вариантов снятия неоднозначности в первом предложении до 168, при этом наиболее вероятным становится вариант, в котором неоднозначность P-Z-N снята корректно:

<s> Num P O **T** O L-N Num P-N </s> (1,93e-11),

Однако для исправления других указанных ошибок требуются более сложные и развернутые правила. Так, для корректного снятия неоднозначности тега DEF-P, присваиваемого прилагательным *another, the other* и местоимениям *others, the other* может быть введено следующее правило: DEF-P разрешается в P только в том случае, если отмеченное им слово является местоимением и связанный с ним предикат также отмечен тегом P, например: [the other]-P [was killed]-P.

Позиционный метод. В основе данного метода лежит идея Г. Эдмундсона о выявлении ключевых слов по их расположению в тексте [15], а также некоторые соображения из стилистики медиатекста. Считается, что структура новостной статьи, как правило, представляет собой перевернутую пирамиду, в которой информация расположена по степени важности: от наиболее релевантной в начале сообщения к наименее релевантной в конце [16]. Следовательно, можно выдвинуть гипотезу, что доля концептуальных тегов максимальна в начале сообщения и снижается к концу.

Для проверки данной гипотезы был использован «золотой» корпус, из которого была удалена лексика. Все 107 новостных сообщений корпуса (длиной от 3 до 44 предложений, модальная длина сообщения — 7 предложений, медианная — 10) были разделены на предложения, после чего на их основе были сформированы позиционные подкорпусы и подсчитана частота тегов в каждом подкорпусе. Результат обработки позиционных подкорпусов показан на рис. 3.

Рис. 3. Доля концептуальных тега в предложениях сообщения (на корпусе объемом 25 тыс. словоупотреблений)

Значение синей кривой, показывающей изменение доли концептуальных тегов на рис. 3, снижается с 72 до 37 % в позиционных корпусах № 1–9, затем незначительно поднимается и вновь падает, после чего кривая становится заметно нестабильной, с наблюдаемыми всплесками, однако не поднимается выше 50 %. Эта нестабильность легко объяснима: несколько статей в корпусе имеют значительно большую длину, чем остальные; как следствие, общее количество тегов (голубая кривая на рис. 3) в подкорпусах предложений, расположенных ближе к концу текста, недостаточно для того, чтобы делать какие-либо выводы. Тем не менее полученные данные показывают, что для среднестатистического сообщения гипотеза о зависимости доли концептуальных тегов от положения в тексте подтверждается.

Этот метод также может использоваться для выбора наиболее подходящего концептуального тега. На рис. 4 представлены кривые распределения тегов RW (COUNTERTERRORISM) и Z (OBJECT OF ATTACK). Если в начале текста частота этих тегов одинакова, то ближе к концу среднестатистического новостного сообщения их частоты становятся диаметрально противоположны.

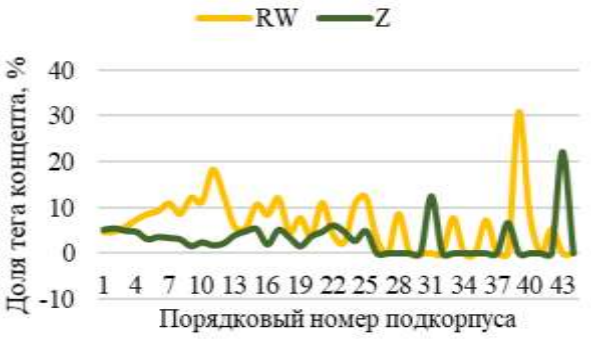

Рис. 4. Распределение тегов RW и Z

IV. ЗАКЛЮЧЕНИЕ

Корпусное исследование показало, что при концептуальной разметке англоязычных новостных сообщений о терактах имеет место концептуальная неоднозначность, вызываемая частеречной омонимией, лексической неоднозначностью, множественностью концептуальных значений, экстралингвистическим контекстом или совокупностью причин.

Рассмотренные количественные методы могут быть использованы для автоматизированного снятия концептуальной неоднозначности при участи человека. Точность получаемых результатов и, следовательно, скорость разрешения неоднозначности может быть повышена путем составления продукционных правил, разработка которых планируется на следующих этапах нашего исследования.

БИБЛИОГРАФИЯ

[1] M. Djemaa, M. Candito, Ph. Muller and L. Vieu, “Corpus annotation within the French FrameNet: a domain-by-domain methodology,” in *Proc. 10th International Conference on Language Resources and Evaluation,* Portorož, Slovenia, 2016, pp. 3794–3801.

[2] Е.В. Рахилина, Б.П. Кобрицов, Г.И. Кустова, О.Н. Ляшевская и О.Ю. Шеманаева, «Многозначность как прикладная проблема: семантическая разметка в национальном корпусе русского языка», в сборнике *Труды международной конференции «Диалог 2006»,* Москва, 2006, с. 445–450.

[3] M. Palmer, P. Gildea and P. Kingsbury, “The proposition bank: an annotated corpus of semantic roles," in *Computational Linguistics* vol. 31(1), 2005, 71–106.

[4] Ц. Линь, Д.М. Семёнова, С.Л. Пущин, Т.Г. Петров, М.Н. Бабарико и С.В. Чебанов, «Ручная разметка корпуса для изучения статистики концептов», в сборнике *Международной научной конференции «Корпусная лингвистика-2019»,* Санкт-Петербург, 2019, с. 248–257.

[5] J.D. Kim, T. Ohta and J. Tsujii, “Corpus annotation for mining biomedical events from literature” in *BMC Bioinformatics* vol. 9, 2008, pp. 9–10.

[6] М.Ю. Загорулько, И.С. Кононенко и Е.А. Сидорова, «Система семантической разметки корпуса текстов в ограниченной предметной области», в сборнике *Труды международной конференции «Диалог 2012»,* Москва, с. 674–685.

[7] D. Song, C.G. Chute and C. Tao, “Semantator: a semi-automatic semantic annotation tool for clinical narratives,” in *Proc. 10th International Semantic Web Conference,* Bonn, Germany, 2011, pp. 1–4.

[8] J.S. Viju, “Concept interpretation by semantic knowledge harvesting,” in *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* vol. 6(5), 2018, pp. 477–484.

[9] A.N. El-Ghobashy, G.M. Attiya and H.M. Kelash, “SAAT: a manual annotation tool for the Arabic content authoring,” in *International Journal of Computing and Digital Systems* vol. 4(4), 2015, pp. 1–6.

[10] S. Sheremetyeva and A. Zinoveva, “Ontological analysis of e-news: a case for terrorism domain,” in *Proc. 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction,* Ulyanovsk, 2019, pp. 130–141.

[11] А.Ю. Сиротина, Н.В. Лукашевич, «Опыт создания корпуса текстов в сфере информационной безопасности», в сборнике *Международной научной конференции «Корпусная лингвистика-2019»,* Санкт-Петербург, 2019, с. 79–85.

[12] S. Sheremetyeva, “Towards creating interoperable resources for conceptual annotation of multilingual domain corpora,” in *Proc. 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16),* Marseille, 2020, pp. 102–109.

[13] S. Nirenburg and V. Raskin, *Ontological semantics*. Cambridge: MIT Press, 2004.

[14] A. Preda, “Lexical ambiguity revisited: on homonymy and polysemy,” in *Proc. International Conference Literature, Discourse and Multicultural Dialogue. Section: Language and Discourse,* Târgu Mureș, Romania, 2013, pp. 1047–1054.

[15] H.P. Edmundson, “New methods in automatic extracting,” in *Journal of the Association for Computing Machinery* vol. 16(2), 1969, pp. 264–285.

[16] T.I. DeAngelo and N.S. Yegiyan, “Looking for efficiency: How online news structure and emotional tone influence processing time and memory,” in *Journalism & Mass Communication Quarterly* vol. 96(2), 2019, pp. 385–405.

Conceptual ambiguity in English texts on terrorism: causes and disambiguation methods

A. Yu. Zinoveva

***Abstract*—Today's natural language processing research frequently addresses the issue of content semantization (including the semantization of unstructured texts such as electronic news) by means of semantic annotation or its special case, ontology-based and domain-oriented conceptual annotation. Conceptual annotation is often complicated by conceptual ambiguity manifested in one-to-many mappings between lexical items and ontology concepts. This paper examines the causes of conceptual ambiguity in restricted domain texts, with the case study of English-language electronic news on terror attacks. Four causes of conceptual ambiguity are revealed: part-of speech homonymy, lexical ambiguity, the plurality of conceptual meanings (the most productive), and the extralinguistic context (the least productive, but the hardest to resolve). Three quantitative disambiguation methods are studied: a) tag ranking, b) a bigram-model-based contextual method, and c) a positional method. All the methods are found useful for computer-aided conceptual disambiguation, yet it is pointed out that these quantitative methods are not quite accurate when used alone and rule-based methods would be a good addition.**

***Keywords*—conceptual ambiguity, conceptual annotation, English, terrorism.**

REFERENCES

[1] M. Djemaa, M. Candito, Ph. Muller and L. Vieu, "Corpus annotation within the French FrameNet: a domain-by-domain methodology," in *Proc. 10th International Conference on Language Resources and Evaluation,* Portorož, Slovenia, 2016, pp. 3794–3801.

[2] E.V. Rakhilina, B.P. Kobritsov, G.I. Kustova, O.N. Lyashevskaya and O.J. Shemanayeva, "Semantic ambiguity as an application-oriented problem: word class tagging in the RNC," in *Proc. International Workshop Dialogue 2006*, Moscow, 2006, pp. 445–450 (in Russian).

[3] M. Palmer, P. Gildea and P. Kingsbury, "The proposition bank: an annotated corpus of semantic roles," in *Computational Linguistics* vol. 31(1), 2005, 71–106.

[4] J. Lin, D.N. Semenova, S.L. Pshchin, T.G. Petrov, M.N. Babariko and S.V. Chebanov, "Manual tagging of the corpus for studying of concept statistics", in *Proc. International Scientific Conference Corpus Linguistics 2019,* Saint Petersburg, 2019, pp. 248–257 (in Russian).

[5] J.D. Kim, T. Ohta and J. Tsujii, "Corpus annotation for mining biomedical events from literature" in *BMC Bioinformatics* vol. 9, 2008, pp. 9–10.

[6] M.J. Zagorulko, I.S. Kononenko and E.A. Sidorova, "System for semantic annotation of domain-specific text corpora," *in Proc. International Conference Dialogue 2012,* Moscow, 2012, pp. 674–685 (in Russian).

[7] D. Song, C.G. Chute and C. Tao, "Semantator: a semi-automatic semantic annotation tool for clinical narratives," in *Proc. 10th International Semantic Web Conference,* Bonn, Germany, 2011, pp. 1–4.

[8] J.S. Viju, "Concept interpretation by semantic knowledge harvesting," in *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* vol. 6(5), 2018, pp. 477–484.

[9] A.N. El-Ghobashy, G.M. Attiya and H.M. Kelash, "SAAT: a manual annotation tool for the Arabic content authoring," in *International Journal of Computing and Digital Systems* vol. 4(4), 2015, pp. 1–6.

[10] S. Sheremetyeva and A. Zinoveva, "Ontological analysis of e-news: a case for terrorism domain," in *Proc. 14th International Conference on Interactive Systems: Problems of Human-Computer Interaction,* Ulyanovsk, 2019, pp. 130–141.

[11] A.Yu. Sirotina, N.V. Loukachevich, "Towards construction of an annotated corpus in cybersecurity," in *Proc. International Scientific Conference Corpus Linguistics 2019,* Saint Petersburg, 2019, pp. 79–85 (in Russian).

[12] S. Sheremetyeva, "Towards creating interoperable resources for conceptual annotation of multilingual domain corpora," in *Proc. 16th Joint ACL-ISO Workshop Interoperable Semantic Annotation (ISA-16),* Marseille, 2020, pp. 102–109.

[13] S. Nirenburg and V. Raskin, *Ontological semantics*. Cambridge: MIT Press, 2004.

[14] A. Preda, "Lexical ambiguity revisited: on homonymy and polysemy," in *Proc. International Conference Literature, Discourse and Multicultural Dialogue. Section: Language and Discourse,* Târgu Mureș, Romania, 2013, pp. 1047–1054.

[15] H.P. Edmundson, "New methods in automatic extracting," in *Journal of the Association for Computing Machinery* vol. 16(2), 1969, pp. 264–285.

[16] T.I. DeAngelo and N.S. Yegiyan, "Looking for efficiency: How online news structure and emotional tone influence processing time and memory," in *Journalism & Mass Communication Quarterly* vol. 96(2), 2019, pp. 385–405.