ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №3(34), 2017, с. 133–167

УДК 004.89:004.912

С. Р. Егикян

Современные методы анализа модальности в текстах на естественном языке

Аннотация. Статья содержит обзор современных подходов к разметке и распознаванию модальной информации в текстах на естественном языке. Широко распространенные точки зрения представлены в их разнообразии — как те, которые нацелены на обработку модальности в широком смысле (включая смежные характеристики, такие как временной план, эвиденциальность и пр), так и те, которые предназначены для отделения модализованной информации от немодализованной.

Ключевые слова и фразы: NLP, natural language processing, анализ естественного языка, автоматическое извлечение информации, извлечение событий, модальность, эвиденциальность, уверенность, спекулятивная информация.

Введение

В последние десятилетия резко увеличился цифровой информационной поток, в значительной степени состоящий из текстов — таких как новостные сообщения, юридические документы, судебные иски и решения, истории болезней, сообщений в социальных сетях. В связи с этим все чаще возникает необходимость выработать эффективные методы обработки этой неструктурированной информации. В сфере компьютерной лингвистики и автоматической обработки естественного языка эта проблема решается в рамках направления Information Extraction (IE) — извлечение информации. Типичной задачей извлечения информации является вычленение заданного события в имеющемся корпусе текстов определенного дискурса (например, тексты СМИ или научные тексты). Результатом анализа неструктурированных текстов является информация о событии, представленная в структурированном виде — фрейме. Фрейм состоит из заранее определенного набора слотов, которые необходимо заполнить (например, для события «теракт» слотами

Работа выполнена в рамках НИР «Исследование и разработка методов автоматического извлечения событийно-темпоральной информации из текстов», номер гос. регистрации 0077-2016-0001.

DOI: 10.25209/2079-3316-2017-8-3-133-167

могут быть «тип теракта», «место», «время», «количество погибших», «количество раненых»). Следует отличать извлечение информации от поиска информации (Information Retrieval) — последнее является более простой задачей и сводится к тому, чтобы из ограниченного количества документов выбрать те, которые соответствуют поисковому запросу, и ранжировать их по степени релевантности [1].

Важными (хотя и необязательными) атрибутами при извлечении информации являются темпоральные и модальные характеристики события (о разметке темпоральной информации см. [2–7]. Существуют различные подходы к вопросу о том, что именно следует понимать под «модальностью» в приложении к извлечению информации. Нельзя говорить о существовании единого стандарта в этой области. Отчасти это объясняется тем, что каждая предметная область навязывает свои приоритеты при определении понятия «модальность». Как правило, большинство подходов разрабатываются либо для обработки новостных текстов (в этом случае акцент делается на степени соответствия фактографической информации действительности), либо для обработки научных текстов (в этом случае необходимо в первую очередь отделить информацию, поданную как объективную, от мнения автора и его умозаключений).

Направление извлечения информации активно разрабатывается мировым сообществом с 80-х гг XX века, когда были проведены первые соревнования между заинтересованными группами исследователей и конференции по итогам этих соревнований. Впервые организованные мероприятия по изучению проблем извлечения информации были проведены в США. При поддержке американского агентства DARPA (Defense Advanced Research Projects Agency) с 1987 по 1997 гг был проведены семь конференций MUC (Message Understanding Conference). Позже при поддержке Национального института стандартов и технологий (NIST) в США проходили соревнования Automatic Content Extraction (ACE, с 1999 по 2008 год) и Text Analysis Conference (TAC, с 2008 года). В настоящее время основной площадкой для исследования методов извлечения информации является конференции SemEval (с 2007 года состоялось 8 конференций, еще одна планируется в текущем 2017 году). В России самые современные методы извлечения информации представляются в рамках международной научной конференции по компьютерной лингвистике «Диалог» (проходит ежегодно с 1995 года).

1. Основные понятия

1.1. Понятие события

Как было указано выше, одним из центральных понятий в IE является выделение событий. Следует учитывать, что в данном случае термин «событие» употребляется в метаязыковом смысле и охватывает серию непредметных (событийных) значений, таких как процесс, ситуация, состояние, действие, изменение, положение дел и т.п. В рамках задач IE могут извлекаться такие события как встреча, назначения, взрывы, вспышка эпидемии, нелегальные демонстрации и т. п. [8, 9].

В рамках программы ACE использовалась иная терминология: предметом автоматического анализа были *отношения* (relations) между именованными сущностями, и эти отношения определялись как предикация. Кроме того, в ходе ACE были разработаны понятия триггер (trigger или anchor) — основное слово, указывающие на событие в тексте, аргумент — упоминание именованной сущности, численное значение или темпоральное выражение (атрибуты события); упоминание события (event mention) — отрезок текста, содержащий триггер, упоминания сущностей и другие атрибуты ранее обнаруженного события [10].

1.2. Понятие модальности

Первые исследования в области автоматической обработки языка были посвящены пропозициональному аспекту значения — например, для вопросно-ответных систем или для извлечения информации по шаблону «кто, что, когда и где сделал». Однако для более полного анализа текста необходимо учитывать и *экстрапропозициональный* аспект, который может включать в себя степень достоверности информации, субъективное отношение говорящего, неуверенность и многое другое.

Две крупнейших задачи, которые пытается решить мировое сообщество в сфере анализа экстрапропозиционального аспекта значения — это, во-первых, обнаружение модальности и отрицания в разных проявлениях и, во-вторых, определение сферы действия маркеров модальности и отрицания в высказывании.

В лингвистике термин «модальность» объединяет разнородные языковые явления, так или иначе обозначающие отношение говорящего к сообщаемому или сообщаемого к действительности. Возможные

значения, объединяемые понятием «модальность», включают оценку говорящим пропозиции с точки зрения реальности/нереальности; оценку ситуации с точки зрения возможности, необходимости или желательности; целевую установку говорящего; эмоциональную оценку говорящим ситуации и т.п. Соответственно, и в сфере автоматической обработки текстов не существует единого мнения относительно того, что следует понимать под «модальными характеристиками». Объем этого термина в значительной мере диктуется конкретной задачей, на которую ориентирована та или иная система IE.

2. Разнообразие подходов к автоматическому извлечению модальности

При извлечении событий из текстов на естественном языке встает вопрос о корректной обработке модальности того или иного события, наряду с аргументной структурой предложения и темпоральными характеристиками. Как отмечают в своей работе С. Мацуёси et al. [11], подход, используемый в лингвистике и логике, не может быть напрямую применим к таким прикладным задачам, как автоматическое извлечение информации. Это объясняется в частности тем, что в лингвистике в первую очередь рассматривается модальность основной пропозиции в предложении, а в сфере извлечения информации важна модальность не только основной пропозиции, но и всех событий, включенных в нее. Немаловажно и то, что лингвистика предлагает слишком детальную классификацию возможных случаев модальности — автоматизация такого рода классификаций была бы слишком трудоемка.

Мы рассмотрим самые популярные из существующих подходов в этой теме, разделив их на две большие группы.

В первую группу можно объединить все точки зрения, которые нацелены на распознавание модальности в широком смысле — другими словами, распознавание нескольких смежных между собой элементов экстрапропозиционального значения. Эти элементы могут быть выбраны и названы очень различно у разных авторов и могут включать, в числе прочих, «перспективу» (подается ли информация с точки зрения автора текста или другого источника), характер информации (факт или мнение), степень уверенности автора в достоверности информации, наличие отрицания, время и т. п.

Вторую группу составляют точки зрения, которые фокусируются на каком-либо одном способе модализации информации. Эти подходы могут быть объединены в одну группу с некоторой долей условности, так как в каждом случае авторы по-своему определяют, что именно является предметом разметки (или извлечения). Это может быть

- либо эпистемическая модальность (когда говорящий дает понять, что у него недостаточно знаний для утверждения о чем-либо),
- либо отделение модализованных высказываний (требований, пожеланий, целевых высказываний и пр) от немодализованных,
- либо отделение спекулятивных высказываний (умозаключений, мнений) от неспекулятивных.

2.1. Расширенное понимание модальности

2.1.1. *TimeML и инструменты на его основе (EvITA, SlinkET)*

TimeML как один из самых популярных языков разметки

В течение более чем десяти лет для разметки текстов на естественном языке активно используется язык TimeML, разработанный в 2004 году группой ученых во главе с Джеймсом Пустейовски (James Pustejovsky) из Университета Брендайса в рамках научной программы TARSQI (Temporal Awareness and Reasoning Systems for Question Interpretation).

В 2009 году TimeML лег в основу международного стандарта для разметки времени и событий в тексте — ISO-TimeML, согласно решению Международной организации по стандартизации (International Organization for Standardization, ISO). Хотя основной задачей TimeML является разметка событий и временнòй информации, этот язык также обладает инструментарием для разметки модальной информации [**12**]. Модальность в этом случае понимается как степень «приверженности» говорящего к событиям, упоминаемым в тексте [**13**].

Способ разметки модальности в TimeML

В TimeML «событием» (event) считается выражение, обозначающее некоторую динамическую ситуацию или состояние, которые могут быть соотнесены с определенной точкой на временнòй оси. Такие события могут быть выражены конструкцией с личными или неличными формами глагола, именными группами, на вершине которых находятся номинализации, существительные, обозначающие события, прилагательные.

Для разметки событий используются метки EVENT (событие) MAKEINSTANCE (экземпляр события). Метка EVENT в атрибутах содержит свой идентификационный номер (атрибут EID), а метка MAKEINSTANCE помимо собственного номера (EIID) ещё и номер события, к которому привязан этот «экземпляр».

Основная информация о модальности записывается в метку SLINK, которая отражает подчинительные связи между событиями. Ее атрибуты: EVENTINSTANCEID, содержащий ссылку на вершину события в подчиняющем предложении, SUBORDINATEDEVENTINSTANCE, содержащий ссылку на событие в подчиненном предложении и RELTYPE, который указывает на модальное значение, проецируемое подчиняющей клаузой на событие в подчиненной клаузе.

Возможные значения атрибута RELTYPE таковы:

FACTIVE (*фактивный*) — если глагол в подчиняющей клаузе вводит пресуппозицию существования для события-аргумента.

COUNTERFACTIVE (*контрфактивный*) — обратный случай: если глагол предполагает контрфактивность события-аргумента.

EVIDENTIAL (*основанный на показаниях или ощущениях*) — вводится предикатами цитирования или восприятия.

NEGATIVE EVIDENTIAL — аналогично предыдущему, но с отрицательной частицей.

MODAL (*модальный*) — для событий, принадлежащих «возможным мирам», также для событий, отражающих разного рода намерения (вводится глаголами со значением «пытаться», «просить», «обещать» пр).

CONDITIONAL (*условный*) — для условных конструкций.

Основанием для установления связи типа SLINK могут служить либо лексические средства, либо структура предложения.

Лексическое основание подразумевает, что одно из событий, участвующих в создании SLINK, относится к какому-либо из заранее определенных классов. Класс события — это обязательный атрибут каждого события. Для SLINK необходимы события следующих классов:

I_ACTION — события, охарактеризованные в документации к TimeML как «интенсиональные». Они вводят событие-аргумент, о котором нельзя сказать, имело ли оно место в действительности. К классу I_ACTION относятся события, выражаемые глаголами *attempt, try, investigate, delay, postpone, avoid, prevent, ask, urge, promise,*

decide, swear, name, appoint, claim (*пытаться, пробовать, расследовать, отсрочивать, откладывать, избегать, предотвращать, просить, призывать, решать, клясться, называть, назначать, утверждать*).

I_STATE — этот класс близок по своему значению к предыдущему. Им следует маркировать события, которые переносят событие-аргумент в «альтернативный мир». Такую метку получают, в частности, события, выраженные глаголами *believe, think, suspect, want, like, hope, fear, need, be ready, be able и др* (*верить, думать, подозревать, хотеть, нравиться, надеяться, опасаться, нуждаться, быть готовым, быть способным*).

PERCEPTION — в этот класс попадают все глаголы восприятия, такие как *see, watch, listen и др.* (*видеть, наблюдать, слышать*).

REPORTING — события этого класса описывают ситуацию, в которой некое лицо или организация сообщает что-либо, заявляет что-либо и т. п.: *say, report, explain...* (*сказать, сообщить, пояснить...*)

Можно добавить, что события классов I_ACTION и I_STATE устанавливают со своими событиями-аргументами связи SLINK со значениями MODAL, FACTIVE или COUNTERFACTIVE. События класса PERCEPTION вводят SLINK со значением EVIDENTIAL или NEGATIVE EVIDENTIAL, тогда как события класса REPORTING могут вводить SLINK с любым значением.

Кроме лексического, возможно также **структурное основание** SLINK, которое имеет место в двух случаях:

Во-первых, в предложениях с придаточными цели, где SLINK устанавливается между событием в главном предложении и событием в придаточном цели и всегда имеет значение MODAL.

Во-вторых, в конструкциях со значением условия связь SLINK связывает события в антецеденте и консеквенте и всегда имеет значение CONDITIONAL.

Приведем примеры разметки предложений с SLINK.

Пример 1. *The Human Rights Committee regretted that discrimination against women persisted in practice.*

TimeML

```
The Human Rights Committee
<EVENT EID="E1" CLASS="I.ACTION"> regretted </EVENT>
that discrimination against women
<EVENT EID="E2" CLASS="ASPECTUAL"> persisted </EVENT>
in practice.
<SLINK EVENTINSTANCEID="E1" SUBORDINATEDEVENTINSTANCE="E2"
    RELTYPE="FACTIVE"/>
```

Пример 2. *A Time magazine reporter avoided jail at the last minute.*

TimeML

```
A Time magazine reporter
<EVENT EID="E1" CLASS="I.ACTION"> avoided </ EVENT>
<EVENT EID="E2" CLASS="STATE"> jail </ EVENT> at the last minute
<SLINK EVENTINSTANCEID="E1" SUBORDINATEDEVENTINSTANCE="E2"
    RELTYPE="COUNTERFACTIVE" />
```

Пример 3. *Uri Lubrani also suggested Israel was willing to withdraw from southern Lebanon.*

TimeML

```
Uri Lubrani also
<EVENT EID="E1" CLASS="I.ACTION"> suggested </ EVENT> Israel was
<EVENT EID="E2" CLASS="I.STATE"> willing </ EVENT> to
<EVENT EID="E3" CLASS="OCCURENCE"> withdraw </ EVENT> from southern
 Lebanon.
<SLINK EVENTINSTANCEID="E1" SUBORDINATEDEVENTINSTANCE="E2"
RELTYPE="MODAL"/>
<SLINK EVENTINSTANCEID="E2" SUBORDINATEDEVENTINSTANCE="E3"
RELTYPE="MODAL"/>
```

Инструменты на базе TimeML: EvITA и SlinkET

Описанный выше язык TimeML используется в двух системах для автоматической разметки модальности — EviTA и SlinkET. Рассмотрим их подробнее.

EviTA (‘Events in Text Analyzer’) нацелена на подробную разметку всех событий в тексте, причем она не ограничена определенным типом событий или определенной предметной областью. EviTA выполняет две основные задачи: распознавание событий (в том смысле, как они были определены выше) и частичный анализ грамматических характеристик этих событий (анализу подвергаются характеристики, необходимые для извлечения временно́й информации). В качестве событий рассматриваются только лексические элементы — такие как глаголы, существительные и прилагательные. Прилагательные помечаются как событие только в том случае, если они упоминаются в качестве событий в корпусе TimeBank и при этом входят в предикат [14].

Ниже представлен пример грамматического правила в EvITA.

Пример 4. *Participants will have to be working on the same topics.*

EvITA

```
[FORM IN FUTUREFORM],
[FORM == 'HAVE'],
[FORM == 'TO', POS == 'TO'],
[FORM == 'BE'], [POS == 'VBG']
⇒
[TENSE = 'FUTURE'
ASPECT = 'PROGRESSIVE'
NF.MORPH = 'NONE']
```

Модальность и полярность для каждого события извлекаются с помощью шаблонов.

Оценка эффективности EvITA проводилась в сравнении с корпусом Time Bank, и EvITA показала довольно высокий результат: точность 74,55%, полнота 78,61%, F1-мера 76,53%.

Если EvITA нацелена на распознавание событий и извлечение всей необходимой информации (грамматическое время, вид, модальность, полярность) на лексическом уровне, то для анализа сложных *синтаксических* конструкций на базе TimeML был разработан инструмент SlinkET [**15**]. SlinkET автоматически извлекает отношения подчиненности, подразумевающие определенное модальное значение — отношения, которые в TimeML размечаются как SLINK (см. выше).

Триггером для обнаружения отношений SLINK является предикат из заданного списка вербальных, номинальных или адъективных предикатов (такие как *regret, promise, be capable*). Для каждого такого предиката заданы (по результатам анализа корпуса текстов) возможные подчиненные контексты и возможные типы связей (SLINK), которые он проецирует на них. Событие, подчиненное предикату из списка, помечается с помощью готового синтаксического модуля. Кроме того, метку SLINK получают придаточные предложения цели.

Например, предикат *investigate* может подчинять придаточное с союзами *if/whether* — в этом случае тип связи SLINK определяется как MODAL, или именную группу, обозначающую событие — в этом случае тип связи SLINK получает значение FACTIVE.

(*1*) *Officials are investigating whether Rudolph participated in all three attacks.* (MODAL)

(*2*) *Officials are investigating all three attacks.* (FACTIVE)

Алгоритм SlinkET основан на правилах и в настоящее время ведутся работы по его адаптации к машинному обучению. Эффективность этого инструмента измерялась на 10% TimeBank, в который вошли 218 связей типа SLINK и 681 событие. SlinkET продемонстрировал точность на уровне 92%, полноту — 56%, а F1-мера составляет 70%.

2.1.2. *Разметка модальности и отрицания по 5 параметрам (отрицание, точка зрения, степень уверенности, модальность в узком смысле, условность)*

Р. Моранте et al [**16**] предложили схему аннотирования, которая предназначена для оценки эффективности систем машинного обучения. Схема предполагает оценку каждого события с точки зрения каждого из следующих аспектов экстрапропозиционального значения:

отрицание — если в тексте присутствует частица «not» или другие маркеры отрицания,

точка зрения — по умолчанию информация подается с точки зрения автора текста, но в некоторых случаях автор может цитировать кого-либо еще,

степень уверенности — события могут быть упомянуты с разной степенью уверенности в их достоверности, включая случаи, в которых степень уверенности нельзя установить. Данный подход относит в категорию «неопределенные» (uncertain) все те события, которые не поданы как достоверные, не делая различий в том, насколько вероятно то или иное событие,

модальность (в узком смысле, подробнее см. ниже),

условность — является ли событие условием для другого события или само зависит от каких-либо условий

Модальность (в узком смысле) авторы подхода разделяют на несколько возможных значений.

немодальное событие — в эту категорию по умолчанию попадают все события, не имеющие каких-либо показателей модальности,

событие-цель — в том случае, если событие представлено как цель, намерение, стремление и т.п.,

событие-необходимость — если событие выражает необходимость или требование,

событие-обязательство,

желательное событие — в эту категорию попадают все желания, намерения и планы.

По умолчанию считается, что каждое событие не имеет отрицания, представлено с точки зрения говорящего, достоверно, не является условием для чего-либо и само не зависит от каких-либо условий. Если же удалось обнаружить показатели, свидетельствующие о наличии одного из этих параметров, то каждый их этих параметров маркируется с помощью следующих кодов:

отрицание: NEG;

точка зрения, отличная от точки зрения автора: PERS;

неопределенность: UNCERT;

модальность:

нет модальных значений: MOD-NON,
цель: MOD-PURP,
необходимость: MOD-NEED,
обязательство: MOD-MUST,
желательность: MOD-WANT;

условность:

является условием для другого события: COND,
зависит от другого события: COND-BY.

2.1.2.1. Сокращенная реализация описанной схемы

Существует и сокращенная версия описанной выше схемы разметки, также разработанная P. Моранте et al. [17]. Этот подход предполагает создание аннотации для каждого события, причем под событиями понимаются только глаголы (в то время как в предыдущем варианте событиями могли выступать и отглагольные существительные). Разметка нацелена на выделение модальности и отрицания как двух основных элементов экстрапропозиционального значения. Событие рассматривается как модализованное во всех случаях, когда оно не представлено как достоверное или фактивное (показатели модализованного события см. ниже).

Возможная классификация событий предполагает только четыре варианта. События, представленные как достоверные, маркируются меткой NONE, если они имели место, или NEG, если событие не произошло (т.е. в предложение присутствует отрицание). Модализованное событие получает метку MOD, если при нем событии присутствует отрицание — то NEGMOD.

Маркерами модальности и отрицания выступают в первую очередь лексические единицы и некоторые синтаксические конструкции.

Для отрицания это могут быть:

существительные (такие как inability),

глаголы (prevents *people from sleeping*),

предлоги (without *providing*),

наречия (*was never tendered as expert*),

частицы (*have no experience*),

местоимения (*none of these measures*),

префиксы (*un*solved *problems*),

союзы (*neither the decision nor the changes*).

Кроме того, можно говорить о том, что событие не произошло, если предложение содержит:

условную конструкцию, относящуюся к прошлому (*If matter and antimatter were truly symmetrical, then they would have come into existence in equal amounts during the Big Bang*) или

некоторые глаголы в определенной форме (*The process to determine the nominee was supposed to have ended four years ago*).

Маркерами модализованного события могут выступать:

модальные глаголы (*may never come back, could improve*),

эпистемические прилагательные (*something is possible*),

эпистемические наречия (*it will probably never generate*),

эпистемические существительные (*possibility*),

глаголы и прилагательные, выражающие пропозициональную установку, например глаголы ментального действия (*we do not believe these attacks breached the servers, we hope to unveil it, the ECB was considering writing down the value of its bonds...*),

обобщения (*American universities are usually happy to accept good students*),

условные конструкции (*If you are highly motivated to minimise your taxes, you can hunt for every possible deduction for which you're eligible*),

выражения цели (*Europe has set a goal of reducing emissions by 80–95% by 2050*),

выражения необходимости (*China has less urgent need to bolster growth*),

выражения обязательства (*Rich countries should cut the most*),

выражения желательности (*They want it raised to 30%*),

эпистемические глаголы, выражающие суждение (*suggesting, assume*),

эпистемические глаголы, выражающие умозаключение (*deduce*).

Ниже можно видеть пример текста, размеченного по этой схеме:

TimeML

```
  INPUT
Some <EVENT ID ="1"> deduce </EVENT> from the overall picture that as
China and other authoritarian states <EVENT ID ="2"> get </EVENT> more
educated and richer, their people will <EVENT ID ="3"> agitate </EVENT>
for greater political freedom, <EVENT ID ="4"> culminating </EVENT>
in a shift to a more democratic form of government.

  OUTPUT
e1=NONE e2=MOD e3=MOD e4=MOD
```

Самый высокий результат, который удалось получить с помощью этого подхода к разметке событий, составил F1 = 0,6368.

2.1.2.2. Дополнение к схеме: введение меток confidenceMod и confidenceNeg

С. Лана-Серрано et al. [18] дополнили и реализовали описанный выше подход. Они разработали модуль, который опирается на описанную выше классификацию значений.

Сначала он размечает нефинитные глаголы и глагольные группы, а также глаголы и глагольные группы в индикативе (на этой стадии также анализируется грамматическое время глагола, активный и пассивный залог и т. п.).

Затем происходит первичная разметка модальности и отрицания, которая заключается в разметке маркеров отрицания и модальности (наречия, предлоги, союзы, местоимения, существительные, префиксы, условные структуры и подчиненные предложения), а также значений отрицания и модальности в семантике глаголов. На этом этапе правила обращаются к ресурсу знаний, который, в частности, содержит информацию о «типе достоверности» (factual type) для глаголов. Под типами достоверности понимаются такие элементы семантики, как неуверенность, уверенность, желание, ожидание, предположение и пр.

На последнем этапе происходит окончательная разметка модальности и отрицания. На этом этапе формируются аннотации EVENT (для собственно события) и MODNEG_CONTEXT (для модализирующего контекста). Аннотации EVENT содержит атрибут MODALITY, значение которого может быть MOD, NEG, MODNEG, NONE (аналогично со схемой разметки, предложенной Моранте et al.) Кроме того, вводятся атрибуты CONFIDENCEMOD и CONFIDENCENEG. Они содержат информацию о том, насколько высока вероятность того, что значение модальности (MOD, NEG, MODNEG или NONE), было присвоено верно. Оба атрибута могут получить одно из следующих значений:

100 — значение по умолчанию. Это значит, что в данном случае не было обнаружено никаких показателей модальности или отрицания;

0 — самая высокая степень уверенности. Вербальная группа содержит показатель модальности или отрицания в себе или эти показатели примыкают к вербальной группе;

1 — показатель модальности или отрицания находится на расстоянии до 5 токенов в пределах знаков препинания;

2 — самая низкая степень уверенности.

Показатель модальности или отрицания находится на расстоянии до 25 токенов, включая знаки препинания.

Оценка этого подхода проводилась в несколько заходов. Для разметки с самой высокой степенью уверенности F1-мера составила 0,6551, для вариантов со средней и с самой низкой степенью уверенности — 0,6342 и 0,6125 соответственно.

2.1.3. *Распознавание неопределенности в новостных текстах по 4 измерениям (перспектива, фокус, время, уверенность)*

Следующие несколько схем, которые мы обсудим, также охватывают широкий круг смежных с модальностью явлений, однако теперь иначе сформулирована цель такой разметки. По мнению некоторых исследователей, первоочередная задача при автоматическом анализе текстов (в первую очередь, текстов новостей) — определить степень уверенности того или иного высказывания.

В. Рубин et al. [19] предложили модель для анализа степени определенности (certainty) высказываний на примере текстов новостей, которая предполагает распознавание четырех «измерений», отражающих степень определенности той или иной информации. Под определенностью в данном случае понимается «качество или состояние, характеризующееся отсутствием сомнения, особенно на основе данных о прошлом, настоящем или о фактической или абстрактной информации в будущем, выраженное пишущим или передаваемое пишущим о других лицах, прямо или косвенно вовлеченных в повествование».

Согласно этой схеме разметки, каждое предложение может быть охарактеризовано по следующим четырем измерениям.

Перспектива — здесь разделяются точка зрения автора и цитируемая (reported) точка зрения. Последняя, в свою очередь, подразделяется на точку зрения лиц, прямо вовлеченных в событие (жертвы, свидетели и пр) и косвенно вовлеченных (эксперты, официальные лица и пр).

Фокус — может быть разделен на абстрактную информацию (мнения, суждения, эмоции, моральные принципы) и фактическую информацию (события, состояния, факты).

Временные рамки (timeline) — прошедшее время включает все завершенные или недавние события или состояния, настоящее — текущие, непосредственно происходящие и незавершенные события, а будущее — прогнозы, планы, предупреждения и возможные действия.

Уровень уверенности , который разделен на четыре возможных градации: абсолютная, высокая, средняя и низкая.

Б. Гужон [20] немного видоизменил эту модель, оставив только три возможных градации степеней уверенности (высокая, средняя, низкая) и заменив «фокус» на параметр «реальность», который может принимать два значения — «утвердительно» и «отрицательно».

Для каждого значения этих параметров описаны возможные лингвистические шаблоны (для французского языка). Например, прилагательные *douteux, incertain, peu probable* («сомнительно, неясно, маловероятно») маркируют низкую степень уверенности, *présumé, supposé* («предположительно, допустимо») — среднюю, *vraisemblable, probable, possible, envisagé* («вероятно, возможно, ожидаемо») — высокую. Обороты *selon, d'après* («согласно, по словам») — цитирование

автором чьей-либо точки зрения, оборот *aller + Inf* — будущее время и т.п.

Б. Гужон применил эту модель для системы извлечения событий таким образом, чтобы для каждого события определять степень уверенности, а также реальность или нереальность события (так как в тексте иногда упоминаются события, которые никогда не происходили). Практически это означает, что первым шагом определяется степень уверенности и реальности для части предложения или всего предложения. После этого запускается модуль, извлекающий события из текста. Затем событию присваиваются все характеристики уверенности (certainty characteristics), ранее определенные для предложения, в котором было обнаружено это событие.

Подход, предложенный Б. Гужоном, показал высокий результат в том, что касается правильной разметки времени и реальности (F-мера 1 и 0,94 соотвественно). Разметка точки зрения и степени уверенности получила F-меру 0,83 и 0,69.

2.1.4. *Модальность, оцениваемая по 7 параметрам и включающая оценку*

Схожим образом к проблеме подошли С. Мацуёси et al. [**11**]. Авторы этого подхода предложили схему разметки текста, применимую к любому языку, которая предполагает для каждого события в тексте определить «расширенную модальность».

«Расширенная модальность» включает в себя следующее:

Источник (S): Событие должно получить особую метку, если в тексте эксплицитно указан источник высказывания, отличный от автора текста, причем предлагается две различные метки — для случаев, когда источник назван в предложении (“WR_(THE AGENT)"), и для случаев, где он обозначен местоимением (“WR_OT") В примере выделено событие, модальные характеристики которого мы оцениваем.

Taro said he wanted to go home soon.

He said he suffered from the illness.

Кроме того, сюда следует отнести случаи, когда источник не назван, но говорящий подает высказывание не от своего имени, например в случае передачи чужих слов (“WR_ARB").

I hear that the mediciation is continued for regulating functions of three semicircular canals.

Если источником высказывания является сам автор текста или говорящий, параметр S получает значение “WR".

Время (T): Этот параметр подразумевает *относительное* время, то есть время события относительно того момента, когда говорящий высказывает свое отношение к нему. Он может принимать два значения — «будущее» и “не-будущее". В приведенном примере выделенное событие относится к «будущему».

It is only a matter of time before progress of cloning technology leads to producing artificial organs.

Условность (C): По этому параметру событие может получить значение CONDITIONAL, если оно является часть условного придаточного, например:

If it is nice out tomorrow, I will go fishing in that lake.

Если событие находится в главном предложении, имеющее условное придаточное, то это для этого события параметру C присваивается значению HASCONDITION, например:

If it is nice out tomorrow, I will go fishing in that lake.

Если же ни один из указанных случаев не применим, то этот параметр получает значение NOTCONDITIONAL.

Основной тип модальности (P): По этому параметру событие охарактеризовано по своему модальному значению (в узком смысле) — как «утверждение», «желание», «побуждение», «разрешение», «вопрос» и т. п.

He said he suffered from symptoms due to stopping steroid medicines at that time. (P = утверждение)

You may use a larger desk. (P = разрешение)

Реальность (A): Этот параметр характеризует событие по двум осям — эпистемической модальности и полярности (утверждение/отрицание).

Эта идея основана на принципе разметки, использованной в корпусе FactBank [21]. В FactBank событие оценивается по оси эпистемической модальности как достоверное (certain или CT), вероятное (probable или PR), возможное (possible или PS, меньшая степень вероятности, чем PR), неуточненное (U), а по оси полярности — как утвердительное (+), отрицательное (-) и неизвестно

(u). Таким образом, событие получает метку CT+, если достоверно известно, что событие произошло или произойдет. Метка PR означает, что событие, вероятно, не произошло или не произойдет, согласно автору текста.

Авторы описываемого подхода предлагают сокращенную шкалу вероятности, объединив PR и PS (вероятность и возможность) в одну метку probable. Кроме того, они вводят особый способ помечать события, которые начали происходить или завершились: "certain $-\rightarrow+$", "certain $+\rightarrow-$", "probable $-\rightarrow+$", "probable $+\rightarrow-$".

Например, "certain $-\rightarrow+$" означает, что означенное событие перешло из категории «доподлинно известно, что оно не происходило» в категорию «достоверно происходит», как в следующем примере:

So, Taro began to use the toothpaste.

Оценка (E): Этот параметр фиксирует субъективную полярность, то есть то, подает ли источник высказывания данное событие как утвердительное или как отрицательное. Этот атрибут имеет три возможных значения: «утвердительное», «отрицательное», «нейтральное». В следующих примерах параметр E = «утвердительное»:

Taro said he wanted to go home soon.

If it is nice out tomorrow, I will go fishing in that lake.

Feel for yourself the effects of restoration water!

You should have told the truth to her at that time.

Следующие три примера иллюстрируют случаи, когда с точки зрения источника высказывания событие является «отрицательным»:

A person with no patience had better not try it.

Jim decided to stop buying the weekly magazine.

If I had known he would come to the party, I would not have been there.

Если в предложении нет оценки события с точки зрения источника высказывания, E = «нейтральное».

Этот параметр может быть важен для таких сфер, как анализ мнений и анализ тональности текста.

Focus (F): Последний параметр должен указывать на ту часть предложения, к которой относится отрицание, вывод (inference) или вопрос.

Например, в следующем примере для события «(he) stayed» F = negation (for you). Это означает, что отрицание распространяется только на фразу “for you":

It was not for you that he stayed.

Ниже можно видеть примеры для ограничения фокуса вывода или вопроса.

I guess he has been on steroids since a month or two after birth.

F=inference(a month or two after birth)

Then, how is xylitol effective?

F=interrogative(how)

Таким образом, в предложенной схеме разметки каждое событие (т.е. главный предикат предложения) может быть охарактеризовано по 7 параметрам. Полная разметка выглядит так (выделено событие, описываемое всеми параметрами):

Taro said he wanted to go home soon.

S=wr Taro, T=future, C=notConditional, P=wish, A=unknown, E=positive, F=no

Is it because Taro received appropriate nutrition that he recovered?

S=wr, T=notFuture, C=notConditional, P=assertion, A=certain+, E=neutral, F=interrogative(because; received)

2.1.5. *Разметка спекулятивных высказываний в научных текстах*

П. Томпсон et al. [22] описали схему аннотирования, предназначенную для научных текстов (на примере статей по биомедицине). Сам процесс разметки относительно прост, так как, по словам авторов этого подхода, в 85% спекулятивные высказывания (содержащие умозаключения, выводы или предположения) маркированы лексическими средствами. По мнению авторов этого подхода, для каждого высказывания в этой сфере важно различать следующие 3 параметра:

Тип знания. Три возможных значения этого параметра заимствованы из работ Ф. Палмера [23], а именно:

(*a*) спекулятивный — маркерами этого типа являются такие слова и фразы как *predict, prediction, hypothisize, hypothesis, view, notion, conceivable, in theory* и т.п.

(*b*) дедуктивный — дедуктивное высказывание можно распознать по таким показателям как *interpret, indication, deduce*.

(*c*) сенсорный — в том случае, если используются такие лексические показатели, как *observation, see, appear*.

И еще один возможный тип знания введен авторами подхода:

(*d*) описательный — для описания результатов опытов или наблюдений, маркерами таких высказываний являются слова *show, reveal, demonstrate, confirmation*.

Степень уверенности. Существуют различные варианты шкалы степени уверенности, в данном случае авторы остановились на четырех градациях: абсолютная уверенность, высокая, средняя и низкая. Для каждой из возможных градаций выделены определенные лексические маркеры (для высокой уверенности — *probable, likely, clearly, normally, generally*; для средней — *possible, perhaps, may, could*; для низкой — *unlikely* и т.п.

Точка зрения. Здесь речь идет о том, с чьей точки зрения описывается то или иное событие. Точка зрения определяется по подстрокам, таким как «*we*», «*our*» (эти маркеры соответствуют точке зрения автора), или различные формы цитат. Предложенная схема разметки предлагает два варианта заполнения этого атрибута: «автор» и «другое».

2.2. Узкое понимание модальности: обнаружение высказываний, отличных от фактуальных

Во вторую большую группу можно объединить те точки зрения, которые предполагают отделение модализированной (или, в некоторых случаях, спекулятивной) информации от немодализированной. В них основное внимание уделяется различению степени уверенности говорящего в том, насколько достоверно его высказывание, и обнаружению умозаключений (в противоположность констатированию фактов).

2.2.1. *Язык UNL*

Язык UNL (Universal Networking Language) предназначен для формального отображения семантики текста на естественном языке. Он разрабатывается с 1996 года, его создателем считается Хироси Учида из университета ООН, в настоящее время этот в этом проекте

участвуют исследователи из 15 стран — Бразилии, Германии, Китая, России, Франции, Японии и др. [24].

Предполагается, что UNL позволяет передать содержание любого текста на естественном языке, представляя каждое предложение в виде гиперграфа. Впоследствии этот граф можно деконвертировать обратно в текст (на исходном или любом другом естественном языке) с помощью специального ПО.

Узлы графа — это так называемые «универсальные слова» (Universal Words, UW), представляющие собой концепты. Ребра графа отражают отношения между «универсальными словами». Дуги графа обозначают семантические отношения (Universal Relations), например, agent (деятель), object (объект), time (время), place (место), instrument (инструмент), mode (образ действия) и др.

Каждое универсальное слово может быть снабжено атрибутом (Universal Attribute) — в атрибутах содержится семантикограмматическая информация о лексической категории, роде существительного, грамматическом времени, полярности, модальности и др.

Атрибут «модальность» охватывает различные значения субъективной модальности — просьба, побуждение, способность, разрешение, ожидание, намерение, необходимость (для этого значения есть несколько меток — необходимость, обязательство, желательность), возможность (для возможности также существует несколько меток с разными оттенками — неизбежность, вероятность и т.п.) Каждому значению соответствует особая метка — @REQUEST, @IMPERATIVE, @GRANT, @GRANT-NOT, @NEED, @OBLIGATION, @OBLIGATIONNOT, @MAY, @POSSIBLE, @PROBABLE, @RARE и др.

2.2.2. *Автоматическая разметка фактивности*

Кэтрин Бейкер et al. [25] разработали схему разметки модальности, модальный лексикон и два автоматических разметчика модальности, которые опираются на этот лексикон.

Подход ориентирован на распознавание модальных значений, отражающих степень достоверности событий (не затрагивая другие смежные элементы значений, такие как эвиденциальность или эмоциональный фон). Под «значениями, связанными с достоверностью» авторы понимают следующие значения (P — пропозиция, H — носитель, или субъект, модальности):

требование: H требует P?

разрешение: H разрешает P?

успех: H удалось P?

усилия: H пытается P?

намерение: H намеревается P?

способность: H может P?

желание: H хочет P?

вера: насколько сильно H верит в P?

Поскольку необходимо учитывать взаимосвязь этих модальных значений с отрицанием, модальность маркируется одним из следующих 13 вариантов:

H требует чтобы P было истинно или ложно,

H разрешает чтобы P было истинно или ложно,

H преуспевает в том чтобы P было истинно или ложно,

H не преуспевает в том чтобы P было истинно или ложно,

H пытается сделать P истинным или ложным,

H не пытается сделать P истинным или ложным,

H намеревается сделать P истинным или ложным,

H не намеревается сделать P истинным или ложным,

H способен сделать P истинным или ложным,

H не способен сделать P истинным или ложным,

H хочет чтобы P было истинным или ложным,

H твердо уверен что P истинно или ложно,

H считает что P истинно или ложно.

Для того, чтобы правильно разметить каждое из указанных значений, K. Бейкер et al. используют следующую схему анализа.

В каждом предложении, выражающем модальность, выделяется три компонента:

триггер — слово или последовательность слов, которые непосредственно выражают то или иное модальное значение;

цель или предмет модальности (target) — событие, состояние или отношение, на которые распространяется модальность;

носитель — субъект модальности (experiencer or cognizer).

Триггером выступают такие слова как *should, try, able, likely, want*. Кроме того, частью триггера может быть отрицательная частица *not*. Часто модальность выражается без использования лексического триггера. Например, обычное утвердительное предложение имеет модальность твердой уверенности, но не содержит лексического триггера.

Для распознавания описанных выше значений был разработан лексикон, содержащий возможные подстроки, выражающие то или иное значение. В каждом пункте лексикона содержится определенная грамматическая информация — часть речи, возможная синтаксическая сочетаемость и т. п.

Ниже можно видеть фрагмент лексикона[1], описывающий глагол *need*:

ModalityLexicon

```
    String: Need
    Pos: VB
    Modality: Require
    Trigger: Need
    Subcat: V3-passive-basic -- The government is needed to buy tents.
    Subcat: V3-I3-basic -- The government will need to work continuously
for at least a year. We will need them to work continuously.
    Subcat: T1-monotransitive-for-V3-verbs -- We need a Sir Sayyed again
to maintain this sentiment.
    Subcat: T1-passive-for-V3-verb -- Tents are needed.
    Subcat: Modal-auxiliary-basic -- He need not go.
```

На основании этого лексикона работают два разметочных модуля — для лексического и для синтаксического уровня. Из трех компонентов, о которых говорилось выше — триггера, цели и носителя модальности — разметочные модули распознают только триггер и цель, так как носитель модальности, как правило, эксплицитно не выражен.

Первый разметочный модуль, который работает на лексическом уровне, обрабатывает текст с уже размеченными частями речи. Он маркирует слова или фразы, точно совпадающие с триггерами, зафиксированными в описанном выше лексиконе.

[1] Сам лексикон доступен по ссылке: http://www.umiacs.umd.edu/~bonnie/ModalityLexicon.txt.

Второй разметочный модуль работает на основе правил, которым на вход подается текст, прошедшего частичный синтаксический разбор (с помощью модуля [**26**]). Каждый шаблон должен определить, где в предложении находится триггер и цель модальности, после чего им присваивается соответствующая метка — например, TRIGREQUIRE и TARGREQUIRE для модальности «требование». Ниже можно видеть пример результаты работы этого разметочного модуля.

Пример 5. *Pakistan which could not reach semi-final, in a match against South African team for the fifth position Pakistan defeated South Africa by 41 runs.*

ModalityLexicon

```
(TOP
(S
(NP
(NNP Pakistan)
(SBAR (WDT which)
(S (MD TrigAble could)
(RB TrigNegation not)
(VB B TargAble TrigSucceed TargNegation reach)
(ADJP (JJ TargSucceed semi-final))
(,,)
(PP (IN in) (DT a)
(NN match) (PP (IN against)
(ADJP (JJ South) (JJ African)) (NN team))
(PP (IN for) (DT the)
(JJ fifth) (NN position))
(NP (NNP Pakistan))))))
(VB D defeated)
(NP (NNP South) (NNP Africa))
(PP (IN by) (CD 41) (NNS runs)) (..)))
```

Эффективность этого метода была измерена на базе 249 предложений на английском языке, для которых предварительно были вручную размечены модальные значения. 86,3% случаев, обработанных автоматически, оказались правильными.

2.2.3. *Обнаружение сферы действия (scope) маркеров спекулятивных высказываний*

Говоря о маркерах модализованной или спекулятивной информации, мы всегда сталкиваемся с проблемой определения сферы действия маркеров модальности или неопределенности. Этой проблеме были посвящены описанные ниже исследования.

Р. Моранте et al. [27] рассматривают этот вопрос на примере статей по биомедицине. Система, выделяющая маркеры неопределенности, была разработана на основании корпуса BioScope [28], который содержит тексты по медицине и биологии с размеченными показателями отрицания и предположения.

Необходимо определить, на какую часть предложения распространяется действие этих показателей. Для каждого токена в предложении определяется, является ли он первым токеном в области действия (тогда он маркируется как F-SCOPE), последним (L-SCOPE) или не входит в область действия (NONE). Для каждого обнаруженного фрагмента предложения устанавливается связь с конкретным маркером неопределенности, выделенным в предыдущей фазе. Это осуществляется и три классификатора были обучены с использованием следующих методов машинного обучения: обучение на основе памяти, метод Support Vector Machine и Conditional Random Fields.

Показатели эффективности этого метода таковы. В подзадаче, выделяющей маркеры неопределенности, F-мера составила 84,77%. В подзадаче по определению сферы действия этих маркеров Pcs% (spam precision rate) составил 66,07%.

Аналогичную задачу поставили перед собой А. Озгюр и Д. Радев [29]. Как и в предыдущем подходе, они поставили перед собой задачу обнаружить спекулятивные высказывания в научных текстах и выделить фрагмент предложения, на который распространяется действие показателя неуверенности, предположения и т. п.

При этом они учитывали, что ключевые слова, маркирующие спекулятивные высказывания, не всегда употребляются в спекулятивном контексте. Следующие два примера иллюстрируют спекулятивное (первый пример) и неспекулятивное употребление слова *appears*.

(*1*) Thus, it appears that the T-cell-specific activation of the proenkephalin promoter is mediated by NF-kappa B.

(*2*) Differentiation assays using water soluble phorbol esters reveal that differentiation becomes irreversible soon after AP-1 appears.

Отличить спекулятивное употребление от неспекулятивного можно с помощью следующих приемов:

- установлено, что каждое ключевое слово с разной вероятностью употребляется в спекулятивном контексте. Например, *appears* является показателем неуверенности в 86% случаев, тогда как *can* – только в 12%;

- синтаксическое окружение тоже играет определенную роль. Некоторые ключевые слова являются показателями спекулятивного высказывания, если употребляются с придаточным предложением или инфинитивным оборотом (например, *appears*);
- некоторые слова являются спекулятивными только в тех случаях, если их сопровождает отрицание. Например, *know, evidence, proof* выражают уверенность, если употреблены без отрицания, и становятся спекулятивными, если употреблены с отрицанием — *not known, no evidence, no proof*;
- вспомогательные глаголы могут указывать на спекулятивное употребление смыслового глагола (*may indicate* — модальный глагол *may* указывает на спекулятивное употребление *indicate*);
- в некоторых случаях на спекулятивное употребление указывает ближайший контекст — три ближайших слова до и после слова, которые мы рассматриваем. Например, слово *indicate* в спекулятивном контексте часто употребляется во фразе «*our findings may indicate the presence of*»;
- определенные слова, маркирующие спекулятивные высказывания, часто встречаются одновременно в одном предложении, например *whether* и *could*.

Когда все маркеры спекулятивности определены, необходимо перейти к определению сферы их действия в предложении. Метод определения этой сферы действия зависит от того, к какой части речи принадлежит обнаруженный маркер спекулятивности:

- Для **союзов** (*or, and/or, vs*) сферой действия является синтаксическая фраза, которую он присоединяет.
- Для **модальных глаголов** (*may, might, could*) сферой действия является глагольная группа (VP), в которую он входит.
- Сфера действия **прилагательного** или **наречия** начинается с самого этого слово и заканчивается на последнем токене, входящем в именную группу (NP) самого высокого уровня, которая подчиняет себе это прилагательное или наречие.
- Сфера действия **глагола, за которым следует инфинитивный оборот**, распространяется на весь этот оборот.
- Сфера действия **глагола в страдательном залоге** распространяется на все предложение.

Если ни один из указанных принципов не применим, сфера действия ключевого слова начинается с самого этого слова и заканчивается вместе с предложением.

Этот метод показал высокую точность в задаче определения ключевых слов, маркирующих спекулятивные высказывания (F- мера = 91,69). Для задачи определения сферы действия этих ключевых слов результат составил F = 79,89 для аннотаций статей и F = 61,13 для текстов.

3. Поиск решений для русского языка

В русскоязычном научном сообществе пока не разработано схем и инструментов для автоматического анализа модальности, которые давали бы как минимум удовлетворительный результат. Также относительно мало́ количество исследований на тему того, насколько применимы к русскому языку те подходы, которые были предложены зарубежными авторами и, по их мнению, являются универсальными. Мы перечислим те публикации, которые затрагивают тему автоматической обработки модальной информации или же посвящены смежным темам.

С анализом понятия достоверности в рамках задачи автоматического извлечения фактографической информации можно ознакомиться в [**30**]. В [**31**] поднимается вопрос о необходимости учитывать модальность события при извлечении его из текста, однако не поясняется, как именно это может быть реализовано в автоматической системе извлечения информации. Следует также упомянуть [**2**], в этой работе содержится оценка возможности применения языка TimeML к русскоязычным текстам (впрочем, в данном случае упор сделан на разметку временны́х выражений).

Потенциально полезную теоретическую базу для будущих работ в сфере автоматического распознавания модальности представляет из себя классификация модальных значений, составленная Е. В. Падучевой [**32**]. Е. В. Падучева описала такие аспекты модальности как «возможность» и «необходимость», в каждом из которых выделены подтипы: внутренняя, внешняя, деонтическая и эпистемическая. Для каждого значения описаны особенности взаимодействия с отрицанием и глагольным видом.

Заключение

Мы перечислили самые распространенные подходы к проблеме распознавания и извлечения модальной информации из текстов на естественных языках. Как можно видеть, они очень различны между собой и могут быть классифицированы по нескольким признакам:

Во-первых, по предметной области, на которую они рассчитаны. Две самых частых ситуации, в которых возникает необходимость в автоматической обработке модальности — это тексты СМИ или же научные тексты. Предметная область отчасти определяет то, каким образом следует подойти к решению поставленной задачи. Для текстов СМИ необходимо в первую очередь отделить достоверную (часто ее называют фактивной) информацию от модализованной информации (то есть содержащую элемент необходимости, цели, желания и пр). В то же время для научных текстов первоочередная задача — отделение информации, поданной как факт, от разного рода спекулятивных высказываний (предположений, умозаключений и т. п.). В соответствии с этим следует опираться на разные типы маркеров. Если для газетных текстов в первую очередь маркерами модальности являются (для английского языка) модальные глаголы типа *may, must, should* и др., то в научных текстах этот маркер играет менее важную роль, поэтому необходимо ориентироваться на показатели «спекулятивности».

Во-вторых, можно разделить все рассмотренные подходы по используемому определению модальности. Зачастую предметом анализа становится «расширенная модальность», включающая в разных комбинациях отрицание, время или временной план, перспективу (подана ли информация с точки зрения автора), степень уверенности и т. п. Реже используется более узкое понимание модальности, когда необходимо отделить модализованную информацию от немодализованной (сюда же можно отнести смежную задачу по отделению спекулятивной информации).

В-третьих, по способу реализации. Некоторые методы реализованы с помощью машинного обучения, другие основаны на правилах. Нередко публикуются предложения по схеме аннотирования текстов, не поддержанные никакой работающей системой.

Благодарности. *Автор благодарит Елену Анатольевну Сулейманову за помощь в подготовке статьи.*

Список литературы

[1] Th. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (eds.), *Multi-source, multilingual information extraction and summarization*, Theory and Applications of Natural Language Processing, Springer-Verlag, Berlin–Heidelberg, 2013, XX+324 p. ↑ 134

[2] Н. С. Ландо. «TimeML для разметки русскоязычных текстов. Оценка перспектив», *Программные системы: теория и приложения*, **7**:4(31) (2016), с. 249–265, URL: http://psta.psiras.ru/read/psta2016_4_249-265.pdf ↑ 134,159

[3] Е. А. Сулейманова. «О двух видах текстовых временных координат», *Программные системы: теория и приложения*, **7**:4(31) (2016), с. 209–229, URL: http://psta.psiras.ru/read/psta2016_4_209-229.pdf ↑ 134

[4] Н. С. Ландо. «Современные методы автоматического анализа темпоральных выражений в текстах на естественном языке», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 419–439, URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf ↑ 134

[5] Ю. П. Сердюк. «Базовая архитектура, методы и алгоритмы системы извлечения темпоральной информации из текстов на естественном языке», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 401–418, URL: http://psta.psiras.ru/read/psta2015_4_401-418.pdf ↑ 134

[6] Е. А. Сулейманова. «Семантический анализ контекстных дат», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 367–399, URL: http://psta.psiras.ru/read/psta2015_4_367-399.pdf ↑ 134

[7] Е. А. Сулейманова. «Лингвистическое моделирование темпорального адвербиала со значением локализации события», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 209–225, URL: http://psta.psiras.ru/read/psta2015_4_209-225.pdf ↑ 134

[8] Н. Д. Арутюнова. *Типы языковых значений. Оценка. Событие. Факт*, Наука, М., 1988. ↑ 135

[9] J. Kim. “Events and their descriptions: some considerations”, *Essays in honor of C. G. Hampel*, Synthese Library, vol. **24**, Dordrecht, 1969. ↑ 135

[10] R. Grishman. “Information extraction: capabilities and challenges”, International Winter School in Language and Speech Technologies WSLST 2012 (Rovira i Virgili University, Tarragona, Spain, January 23–27, 2012), 41 p., URL: http://cs.nyu.edu/grishman/tarragona.pdf ↑ 135

[11] S. Matsuyoshi, M. Egushi, C. Sao, K. Murakami, K. Inui, Y. Matsumoto. “Annotating event mentions in text with modality, focus, and source information”, *Proceedings of LREC 2010* (Valletta, Malta, May 17–23, 2010), pp. 1456–1463, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/682_Paper.pdf ↑ 136,148

[12] J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, L. Mani. “The Specification Language TimeML”, *The Language of Time: A Reader*, eds. L. Mani, J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, URL: http://www.timeml.org/timeMLdocs/timeMLspec.pdf ↑ 137

[13] R. Saurí, M. Verhagen, J. Pustejovsky. “Annotating and recognizing event modality in text”, *Proceedings of FLAIRS 2006* (Melbourne Beach, Florida, USA, May 11–13, 2006), pp. 333–338, URL: https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-065.pdf ↑ 137

[14] J. Pustejovsky et al. “The TIME-BANK Corpus”, *Proceedings of Corpus Linguistics 2003* (Lancaster University, UK, 28–31 March, 2003), pp. 647–656. ↑ 140

[15] R. Sauri, M. Verhagen, J. Pustejovsky. “SlinkET: A partial model parser for events”, *Proceedings of LREC 2006* (Genoa, Italy, 22–28 May, 2006), pp. 1332–1337, URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/716_pdf.pdf ↑ 141

[16] R. Morante, W. Daelemans. “Annotating modality and negation for a machine reading evaluation”, CLEF 2012 (Rome, Italy, 17–20 September, 2012), CEUR Workshop Proceedings, vol. **1178**, 11 p., URL: http://ceur-ws.org/Vol-1178/CLEF2012wn-QA4MRE-MoranteEt2012b.pdf ↑ 142

[17] R. Morante, W. Daelemans. “Annotating modality and negation for a machine reading evaluation”, CLEF 2011 (University of Amsterdam, Netherlands, 19–22 September, 2011), CEUR Workshop Proceedings, vol. **1177**, 14 p., URL: http://ceur-ws.org/Vol-1177/CLEF2011wn-QA4MRE-MoranteEt2011.pdf ↑ 143

[18] S. Lana-Serrano, D. Sanchez-Cisneros, P. Martinez-Fernandez, A. Moreno-Sandoval, L. Campillo-Llanos. “An approach for detecting modality and negation in texts by using rule-based techniques”, CLEF 2012 (Rome, Italy, 17–20 September, 2012), CEUR Workshop Proceedings, vol. **1178**, 13 p., URL: http://ceur-ws.org/Vol-1178/CLEF2012wn-QA4MRE-SerranoEt2012.pdf ↑ 145

[19] V. L. Rubin, N. Kando, E. D. Liddy. “Certainty categorization model”, *Exploring attitude and affect in text: theories and applications*, 2004 AAAI Spring Symposium (Stanford University in Palo Alto, California, USA, March 22–24, 2004), pp. 118–123. ↑ 147

[20] B. Goujon. “Uncertainty detection for information extraction”, *Recent Advances in Natural Language Processing*, International Conference RANLP 2009 (Borovets, Bulgaria, September 14–16, 2009), pp. 118–122, URL: http://www.aclweb.org/anthology/R09-1023 ↑ 147

[21] R. Saurí, J. Pustejovsky. “Factbank: a corpus annotated with event factuality”, *Language Resources and Evaluation*, **43**:3 (2009), 227. ↑ 149

[22] P. Thompson, G. Venturi, J. McNaught, S. Montemagni, S. Ananiadou. “Categorising modality in biomedical texts”, *Building and evaluating resources for biomedical text mining*, LREC 2008 Workshop (Marrakech, Morocco, May 26, 2008), pp. 27–34, URL: http://www.lrec-conf.org/proceedings/lrec2008/ ↑ 151

[23] F. Palmer. *Mood and modality*, Cambridge University Press, Cambridge, 1986. ↑ 151

[24] J. Cardeñosa, A. Gelbukh, E. Tovar (eds.). *Universal Networking Language: Advances in Theory and Applications*, Instituto Politécnico Nacional, Centro de Investigación en Computación, México, 2005. ↑ 153

[25] K. Baker, M. Bloodgood, B. J. Dorr, N. W. Filardo, L. Levin, C. Piatko. “A modality lexicon and its use in automatic tagging”, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, May 17–23, 2010), pp. 1402–1407, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/446_Paper.pdf ↑ 153

[26] S. Miller, H. Fox, L. Ramshaw, R. Weischedel et al. “Algorithms that learn to extract information BBN: description of the SIFT system as used for MUC-7”, Seventh Message Understanding Conference (MUC-7) (Fairfax, Virginia, April 29–May 1, 1998), 17 p., URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/bbn_muc7.pdf ↑ 156

[27] R. Morante, W. Daelemans. “Learning the scope of hedge cues in biomedical texts”, *Proceedings of the Workshop on BioNLP 2009* (Boulder, Colorado, USA, June 4–5, 2009), pp. 28–36, URL: http://anthology.aclweb.org/W/W09/W09-13.pdf#page=40 ↑ 157

[28] G. Szarvas, V. Vincze, R. Farkas, J. Csirik. “The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts”, *Proceedings of the Workshop on BioNLP 2008* (Columbus, Ohio, USA, June 19–20, 2008), pp. 38–45, URL: http://www.aclweb.org/anthology/W08-0606 ↑ 157

[29] A. Özgür, D. Radev. “Detecting speculations and their scopes in scientific text”, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 6–7 August, 2009), pp. 1398–1407. ↑ 157

[30] С. Р. Егикян, Е. А. Сулейманова. «Модальность достоверности в задаче извлечения фактографической информации из текстов на естественном языке», *Программные системы: теория и приложения*, **7**:4(31) (2016), с. 267–286, URL: http://psta.psiras.ru/read/psta2016_4_267-286.pdf ↑ 159

[31] В. В. Данилова, С. В. Попова. *Извлечения событий из неструктурированного текста для задач интернет-социологии*, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Российская Академия Народного Хозяйства и Государственной Службы при Президенте Российской Федерации», М., 2009. ↑ 159

[32] Е. В. Падучева. «Вид, модальность и отрицание: корпусное исследование», Конференция «Ассерция и негация». Проблемная группа «Логический анализ языка» (Институт языкознания РАН, Москва, Россия, 28–30 мая 2007), 5 с., URL: http://lexicograph.ruslang.ru/TextPdf2/arut_mod_asp_negation.pdf ↑ 159

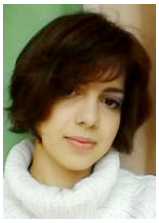
Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Пример ссылки на эту публикацию:

С. Р. Егикян. «Современные методы анализа модальности в текстах на естественном языке», *Программные системы: теория и приложения*, 2017, **8**:3(34), с. 133–167.

URL: http://psta.psiras.ru/read/psta2017_3_133-167.pdf

Об авторе:



Седа Рубеновна Егикян

Инженер Исследовательского центра искусственного интеллекта ИПС им. А.К.Айламазяна РАН. Область научных интересов: компьютерная лингвистика, теоретическая лингвистика, автоматическая обработка естественного языка

e-mail: seda.egikian@gmail.com

Seda Egikyan. *Up-to-date methods of the modality analysis in natural language texts.*

Abstract. The article reviews the up-to-date methods of recognizing and annotating modal characteristics in the natural language texts. Widespread views are presented in their diversity, including those concerned with the extended modality and those aiming at the differentiation between moralized and unmodalized information. (*In Russian*).

Key words and phrases: NLP, natural language processing, natural langage text mining, information extractiong, event extraction, modality, evidentiality, certainty, speculation.

References

[1] Th. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (eds.), *Multi-source, multilingual information extraction and summarization*, Theory and Applications of Natural Language Processing, Springer-Verlag, Berlin–Heidelberg, 2013, XX+324 p.

[2] N. S. Lando. "TimeML markup language for Russian. Future outlook", *Program Systems: Theory and Applications*, **7**:4(31) (2016), pp. 249–265, URL: http://psta.psiras.ru/read/psta2016_4_249-265.pdf

[3] Ye. A. Suleymanova. "On two types of time-referring expressions", *Program Systems: Theory and Applications*, **7**:4(31) (2016), pp. 209–229, URL: http://psta.psiras.ru/read/psta2016_4_209-229.pdf

[4] N. S. Lando. "Up-to-date methods of automatic time expression resolution in natural language texts", *Program Systems: Theory and Applications*, **6**:4(27) (2015), pp. 419–439, URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf

[5] Yu. P. Serdyuk. "Basic architecture, methods and algorithms of system for temporal information extraction from natural language texts", *Program Systems: Theory and Applications*, **6**:4(27) (2015), pp. 401–418, URL: http://psta.psiras.ru/read/psta2015_4_401-418.pdf

[6] Ye. A. Suleymanova. "Semantic analysis of contextual dates", *Program Systems: Theory and Applications*, **6**:4(27) (2015), pp. 367–399, URL: http://psta.psiras.ru/read/psta2015_4_367-399.pdf

[7] Ye. A. Suleymanova. "Linguistic modeling of temporal adverbial localizer", *Program Systems: Theory and Applications*, **6**:4(27) (2015), pp. 209–225, URL: http://psta.psiras.ru/read/psta2015_4_209-225.pdf

[8] N. D. Arutyunova. *Types of language meanings. Evaluation. Event. Fact*, Nauka, M., 1988.

[9] J. Kim. "Events and their descriptions: some considerations", *Essays in honor of C. G. Hampel*, Synthese Library, vol. **24**, Dordrecht, 1969.

[10] R. Grishman. "Information extraction: capabilities and challenges", International Winter School in Language and Speech Technologies WSLST 2012 (Rovira i Virgili University, Tarragona, Spain, January 23–27, 2012), 41 p., URL: http://cs.nyu.edu/grishman/tarragona.pdf

DOI: 10.25209/2079-3316-2017-8-3-133-167

[11] S. Matsuyoshi, M. Egushi, C. Sao, K. Murakami, K. Inui, Y. Matsumoto. "Annotating event mentions in text with modality, focus, and source information", *Proceedings of LREC 2010* (Valletta, Malta, May 17–23, 2010), pp. 1456–1463, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/682_Paper.pdf

[12] J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, L. Mani. "The Specification Language TimeML", *The Language of Time: A Reader*, eds. L. Mani, J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, URL: http://www.timeml.org/timeMLdocs/timeMLspec.pdf

[13] R. Saurí, M. Verhagen, J. Pustejovsky. "Annotating and recognizing event modality in text", *Proceedings of FLAIRS 2006* (Melbourne Beach, Florida, USA, May 11–13, 2006), pp. 333–338, URL: https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-065.pdf

[14] J. Pustejovsky et al. "The TIME-BANK Corpus", *Proceedings of Corpus Linguistics 2003* (Lancaster University, UK, 28–31 March, 2003), pp. 647–656, URL: https://pdfs.semanticscholar.org/4213/4679e402211284a1b0f7f956fc424d5f2eec.pdf

[15] R. Sauri, M. Verhagen, J. Pustejovsky. "SlinkET: A partial model parser for events", *Proceedings of LREC 2006* (Genoa, Italy, 22–28 May, 2006), pp. 1332–1337, URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/716_pdf.pdf

[16] R. Morante, W. Daelemans. "Annotating modality and negation for a machine reading evaluation", CLEF 2012 (Rome, Italy, 17–20 September, 2012), CEUR Workshop Proceedings, vol. **1178**, 11 p., URL: http://ceur-ws.org/Vol-1178/CLEF2012wn-QA4MRE-MoranteEt2012b.pdf

[17] R. Morante, W. Daelemans. "Annotating modality and negation for a machine reading evaluation", CLEF 2011 (University of Amsterdam, Netherlands, 19–22 September, 2011), CEUR Workshop Proceedings, vol. **1177**, 14 p., URL: http://ceur-ws.org/Vol-1177/CLEF2011wn-QA4MRE-MoranteEt2011.pdf

[18] S. Lana-Serrano, D. Sanchez-Cisneros, P. Martinez-Fernandez, A. Moreno-Sandoval, L. Campillo-Llanos. "An approach for detecting modality and negation in texts by using rule-based techniques", CLEF 2012 (Rome, Italy, 17–20 September, 2012), CEUR Workshop Proceedings, vol. **1178**, 13 p., URL: http://ceur-ws.org/Vol-1178/CLEF2012wn-QA4MRE-SerranoEt2012.pdf

[19] V. L. Rubin, N. Kando, E. D. Liddy. "Certainty categorization model", *Exploring attitude and affect in text: theories and applications*, 2004 AAAI Spring Symposium (Stanford University in Palo Alto, California, USA, March 22–24, 2004), pp. 118–123.

[20] B. Goujon. "Uncertainty detection for information extraction", *Recent Advances in Natural Language Processing*, International Conference RANLP 2009 (Borovets, Bulgaria, September 14–16, 2009), pp. 118–122, URL: http://www.aclweb.org/anthology/R09-1023

[21] R. Saurí, J. Pustejovsky. "Factbank: a corpus annotated with event factuality", *Language Resources and Evaluation*, **43**:3 (2009), 227.

[22] P. Thompson, G. Venturi, J. McNaught, S. Montemagni, S. Ananiadou. "Categorising modality in biomedical texts", *Building and evaluating resources for biomedical text mining*, LREC 2008 Workshop (Marrakech, Morocco, May 26, 2008), pp. 27–34, URL: http://www.lrec-conf.org/proceedings/lrec2008/

[23] F. Palmer. *Mood and modality*, Cambridge University Press, Cambridge, 1986.

[24] J. Cardeñosa, A. Gelbukh, E. Tovar (eds.). *Universal Networking Language: Advances in Theory and Applications*, Instituto Politécnico Nacional, Centro de Investigación en Computación, México, 2005.

[25] K. Baker, M. Bloodgood, B. J. Dorr, N. W. Filardo, L. Levin, C. Piatko. "A modality lexicon and its use in automatic tagging", *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (Valletta, Malta, May 17–23, 2010), pp. 1402–1407, URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/446_Paper.pdf

[26] S. Miller, H. Fox, L. Ramshaw, R. Weischedel et al. "SIFT: Statistically-derived information from text", Seventh Message Understanding Conference (MUC-7) (Fairfax, Virginia, April 29–May 1, 1998), 17 p., URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/bbn_muc7.pdf

[27] R. Morante, W. Daelemans. "Learning the scope of hedge cues in biomedical texts", *Proceedings of the Workshop on BioNLP 2009* (Boulder, Colorado, USA, June 4–5, 2009), pp. 28–36, URL: http://anthology.aclweb.org/W/W09/W09-13.pdf#page=40

[28] G. Szarvas, V. Vincze, R. Farkas, J. Csirik. "The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts", *Proceedings of the Workshop on BioNLP 2008* (Columbus, Ohio, USA, June 19–20, 2008), pp. 38–45, URL: http://www.aclweb.org/anthology/W08-0606

[29] A. Özgür, D. Radev. "Detecting speculations and their scopes in scientific text", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, 6–7 August, 2009), pp. 1398–1407.

[30] C. R. Egikyan, Ye. A. Suleymanova. "The actuality modality in the framework of the information extraction for texts written in a natural language", *Program Systems: Theory and Applications*, **7**:4(31) (2016), pp. 267–286, URL: http://psta.psiras.ru/read/psta2016_4_267-286.pdf

[31] V. V. Danilova, S. V. Popova. *Extracting events from unstructured text for the tasks of Internet sociology*, Federal'noye gosudarstvennoye byudzhetnoye obrazovatel'noye uchrezhdeniye vysshego professional'nogo obrazovaniya ≪Rossiyskaya Akademiya Narodnogo Khozyaystva i Gosudarstvennoy Sluzhby pri Prezidente Rossiyskoy Federatsii", M., 2009.

[32] Ye. V. Paducheva. "Type, modality and negation: case study", Konferentsiya ≪Assertsiya i negatsiya". Problemnaya gruppa ≪Logicheskiy analiz yazyka" (Institut yazykoznaniya RAN, Moskva, Rossiya, 28–30 maya 2007), 5 p., URL: http://lexicograph.ruslang.ru/TextPdf2/arut_mod_asp_negation.pdf

Sample citation of this publication:

Seda Egikyan. "Up-to-date methods of the modality analysis in natural language texts", *Program systems: Theory and applications*, 2017, **8**:3(34), pp. 133–167. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2017_3_133-167.pdf