ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(31), 2016, с. 267–286

УДК 004.89:004.912

С. Р. Егикян, Е. А. Сулейманова

Модальность достоверности в задаче извлечения фактографической информации из текстов на естественном языке

Аннотация. Данная статья посвящена проблеме анализа достоверности фактографической информации при извлечении событий из текстов.

В первой части статьи оговорены основные понятия, такие как целевая пропозиция, модальность и субъект речи. Во второй части определяется понятие «достоверность» и описана его структура. В третьей части перечислены самые типичные контексты для базового случая достоверности.

Ключевые слова и фразы: анализ естественного языка, автоматическое извлечение информации, модальность, достоверность.

Введение

Одной из важнейших задач в сфере обработки естественного языка (Natural Language Processing, NLP) является извлечение фактографической информации (information extraction, event extraction), то есть извлечение структурированной информации о ситуации заданного типа из текстов на естественном языке (в первую очередь мы ориентируемся на тексты СМИ). Структура (фрейм) извлеченной информации зависит от поставленной задачи, но в самом типичном случае извлекается упоминание о событии и атрибуты события: где произошло событие, его участники и т. п. (Подробнее об этом см. [1]) Для получения более полной картины необходимо также извлечь модальные и темпоральные характеристики события. Настоящая статья посвящена анализу модального аспекта.

Прежде всего следует определить, под каким углом рассматривать понятие «модальность», какие модальные значения следует выделить и как обозначить их для дальнейшей автоматической обработки или для восприятия конечным пользователем.

Работа выполнена в рамках НИР «Моделирование модально-временного аспекта описания ситуаций в задаче извлечения информации из текстов», номер гос. регистрации 01201455353.

В разработанных мировым сообществом в течение последних лет подходах можно выделить две основных тенденции. В первом случае модальность понимается узко, как отсыл к альтернативным и возможным мирам (то есть такие узкие значения как намерение, цель, разрешение и запрет и т.д.), а низкая степень уверенности субъекта речи выделена в отдельную категорию (часто без градуирования различных степеней уверенности) — такие как стандарт TimeML [2] и основанная на этом стандарте система EvITA [3]. В некоторых случаях значение «возможность» объединяется вместе с перечисленными выше узкими значениями в рамках одного атрибута «модальность», например в языке UNL [4]. Вторая тенденция заключается в том, чтобы сосредоточиться на анализе событий, поданных в тексте как возможные или вероятные. В некоторых моделях выделяется несколько значений модальности по типу соотнесенности события с действительностью — например, «фактивное», «возможное», «предполагаемое» и «условное» ([5] и другие модели со сходным набором модальных значений). Сюда же можно отнести точку зрения (которая была развита и нами в настоящей работе), которая основана на разной степени уверенности говорящего в том, что сообщаемое им имело место, при этом выделяется от трех до пяти разных степеней уверенности, каждой сопоставлены определенные маркеры и (иногда) информация о том, насколько широко распространяется сфера действия этих маркеров [6]. Недостаток описанных подходов из второй группы заключается в том, что в них либо не уделяется внимание немаркированным по признаку модальности случаям (т. н. «уверенная модальность»), либо упускаются некоторые смежные аспекты (такие как цитирование, которое в определенных случаях может снижать степень достоверности).

В этой работе предпринимается попытка разработать подход, который позволит классифицировать *всю* распознанную фактографическую информацию по степени ее достоверности, то есть по ее отношению к действительности. Этот подход рассчитан на то, что будут выработаны формальные критерии выделения описанных значений, с помощью которых станет возможна автоматическая разметка текста.

В отличие от существующих подходов к этому вопросу, особое внимание уделяется тому, как следует анализировать события, находящиеся в сфере действия разнородных маркеров модальности. Учитывается максимально широкий круг модальных модификаторов: не только модальные глаголы и вводные конструкции со значением

достоверности (которые обычно в первую очередь рассматриваются всеми авторами в теме модальность), но и показатели цитирования, фактивные и импликативные глаголы и т.п.

1. Основные понятия

Поскольку в терминологии по теме «модальность» существует множество разночтений, необходимо начать с определения основных терминов, которыми мы будем пользоваться.

1.1. Понятие пропозиции

Прежде, чем вплотную подойти к обсуждению модальности, нам необходимо упомянуть крайне важное для этой темы понятие *пропозиции* (в данном случае это лингвистический термин, не идентичный пропозиции в логике).

Пропозиция отражает **объективное** содержание предложения, то есть семантическое ядро, которое остается неизменным во всей модальной и коммуникативной парадигме конкретного языка или при переводе на другой язык. Например, высказывание *«Соединенное Королевство движется к распаду»* и его перефразировки *«СК двигалось (будет двигаться, двигалось бы, пусть двигается, должно двигаться. . .) к распаду»* имеют инвариант: *«Соединенное королество — двигаться — распад»*. В пропозиции можно выделить предикат и его актанты.

Пропозиция является лишь «материалом» для высказывания. Для того, чтобы высказывание состоялось, объективное семантическое ядро должно быть включено в соответствующие ему **субъективные** значения и соответствующую им формальную организацию (определение и примеры из [7]).

Теория пропозиции является наследником средневековой схоластической теории, которая предполагает расчленение предложения на диктум (объективную семантическую константу) и модус (субъективную переменную). В XX в. аналогичное разделение предложил Ш. Балли [8]. В теории пропозиции модусу соответствует *пропозициональное отношение*, т.е. выражение позиции субъекта речи к предмету речи.

Пропозиции выражаются предикатными группами или определенными непредикативными конструкциями (деепричастием, деепричастным оборотом, причастием, причастным оборотом, отглагольными существительными, прилагательными и др.). В следующих двух высказываниях пропозиция одинакова: «Иван Иванович — приехать» (эту часть семантики высказывания мы будем обозначать P):

Иван Иванович приехал

Вот бы Иван Иванович приехал.

В обоих случаях пропозиция выражена предикатной группой, однако она приобретает разные темпорально-модальные характеристики: в первом случае речь о свершившемся событии, во втором — значение желательности.

В высказывании предикат может совмещать себе значения, соответствующие пропозиции, и значения, соответствующие пропозициональному отношению (в первую очередь это касается предикатов, выраженных личной формой глагола). В некоторых случаях сложно разграничить эти значения, так как нет единого подхода к тому, как определить аспект, время и модальность. Мы вслед за Дж. Лайонзом [9] будем считать, что значение, выражаемое видовой формой глагола, входит в пропозицию, тогда как значение, выражаемое временной формой, не входит. Такое различение можно объяснить тем, что время, в отличие от вида, является дейктической категорией, т.е. соотнесено с субъектом речи.

Проиллюстрируем сказанное на примерах из СМИ.

Лавров и Керри обсудили ситуацию в Алеппо.

Лавров и Керри обсудят ситуацию в Алеппо.

Сергей Лавров и Джон Керри будут обсуждать всю сирийскую проблематику.

В этих предложениях можно выделить пропозицию:

<Лавров, Керри — обсудить — ситуацию в Алеппо> (для первых двух случаев) и

<Сергей Лавров, Джон Керри — обсуждать — всю сирийскую проблематику>.

Сосредоточимся на вершине этих пропозиций: «обсудить» и «обсуждать». Во всех трех случаях предикат высказывания совмещает в себе значения, соответствующие пропозиции (собственно семантика «обсудить» и «обсуждать»), и значения, соответствующие пропозициональному отношению (форма настоящего или будущего времени).

Следует также отметить, что в первых двух случаях вершина пропозиции одна и та же («обсудить»), тогда как в третьем («обсуждать») мы сталкиваемся с другой видовой формой, которая образует иную пропозицию.

Не следует смешивать утверждение или отрицание той или иной ситуации, содержащееся в высказывании, и степень достоверности этой ситуации (понятие «достоверности» будет определено в главе 2). Отрицание некоторой ситуации может входить в пропозицию. И уже будучи реализованной в высказывании, эта пропозиция принимает тот или иной статус. Поясним на примере:

Но в итоге глава «Укроборонпрома» Роман Романов сообщил, что отец Сергея Пинькаса ***не будет назначен*** *директором завода.*

В этом высказывании можно выделить несколько пропозиций, но мы рассмотрим одну из них (которая может стать предметом задачи извлечения событий): <отец Сергея Пинькаса — не назначить — директором завода> «Неназначение» указанного лица на должность директора является тем самым «семантическим ядром» (т. е. пропозицией), которое реализуется в высказывании говорящим. А то, каковы отношения между этой пропозицией и реальностью, или между этой пропозицией и говорящим, нам как раз и предстоит установить.

1.2. Целевая пропозиция

Мы будем пользоваться термином «целевая пропозиция» (ЦП). Под ЦП будет пониматься пропозиция, выражающая ситуацию, которая может стать предметом поиска алгоритма извлечения событий, например «назначения и отставки», «встречи и переговоры», «визиты» и т. п.

Председатель правительства Российской Федерации Дмитрий Медведев ***назначил*** *на должность заместителя министра промышленности и торговли РФ Василия Осьмакова.*

В этом примере целевая пропозиция: «Дмитрий Медведев — назначить — на должность заместителя министра промышленности и торговли РФ — Василий Осмаков». Внутри этой пропозиции можно выделить предикат («назначить») и его актанты (кто, кого, на какую должность).

Для оценки модальных характеристик мы отталкиваемся от предиката внутри пропозиции (в данном случае — «назначить»). В дальнейшем, используя термин «целевая пропозиция», мы будем иметь в виду предикат этой пропозиции.

1.3. Понятие модальности и модальность достоверности

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. При этом доминирующим признаком модальности часто называется отнесенность содержания высказывания к действительности [10]. Мы рассмотрим именно этот последний аспект в свете того, каким образом сам говорящей оценивает степень соответствия содержания высказывания действительности.

1.4. Понятие субъекта. Автор текста как главный субъект

Предполагается, что в каждом высказывании обязательно присутствует субъект речи. В распоряжении исследователя оказывается картина реальности, отраженная в сознании субъекта речи. Поэтому при анализе текста невозможно опираться на некое «истинное» положение вещей, а только на его «субъективный аналог — ощущение субъекта, что пропозиция P является истинной» [16]. Другими словами, простое утверждение содержит скрытый семантический компонент, что «Я знаю, что P» [16]:

Министр иностранных дел Турции Мевлют Чавушоглу ***нанес*** *незапланированный* ***визит*** *в Иран.*

Такое утверждение следует понимать как «Я (т. е. субъект речи) знаю, что министр иностранных дел Турции нанес визит. . . »

Делая утверждение, субъект одновременно выражает определенную пропозицию и выражает свое отношение к ней. Это отношение принято называть *эпистемическим обязательством.* Как поясняет Дж. Лайонз, субъект, утверждающий некоторую пропозицию, «берет на себя обязательство быть «приверженным» ей, не в том смысле, что он должен на самом деле знать или полагать, что она истинна, но в том смысле, что его последующие утверждения $<\ldots>$ должны согласовываться с мнением, что она истинна» [9].

1.5. Цитируемые источники как второстепенные субъекты

В новостных текстах часто присутствует несколько субъектов речи. Информация может приводиться со ссылкой на какой-либо источник (новостное агентство, лицо, организацию и т. п.). В этом случае автор текста является *основным* субъектом, а все источники — *цитируемым* субъектом. См. типичный пример:

Испанец Альберто Ундиано Мальенко ***назначен*** *главным арбитром матча московского ЦСКА с леверкузенским «Байером» в рамках Лиги чемпионов,* ***сообщается на сайте УЕФА****.*

В таких случаях автор снимает с себя часть ответственности за сказанное, однако не дистанцируется окончательно. Наличие цитирования обычно говорит о том, что «автор присоединяется к мнению цитируемого субъекта или по крайней мере не имеет противоречащих сведений» [11]. В некоторых случаях автор может полностью дистанцироваться от цитируемого или указать на то, что он придерживается прямо противоположного мнения. Такая позиция автора всегда маркирована — так, в следующем примере автор указывает, что не берет на себя ответственность за содержание цитируемого сообщения, с помощью частицы «якобы»:

Ранее СМИ сообщили, что Трамп ***якобы*** *планирует посетить Россию в январе после своего официального вступления в должность.*

2. Достоверность

В рамках задачи извлечения информации важно в первую очередь вычленить из текста упоминания о событиях, соответствующих действительности. Центральным понятием при таком подходе является *достоверность* информации.

Этот термин используется и в лингвистике, и в других областях знания, однако определяется разными авторами различно. Поэтому необходимо подробно оговорить, в каком смысле понятие достоверности используется в данной работе.

Для нас определить степень достоверности информации значит ответить на вопрос: насколько можно утверждать (исходя из текста), что рассматриваемое событие имело, имеет или будет иметь место?

2.1. Достоверность как субъективная категория

Чтобы ответить на этот вопрос, нужно проанализировать, каким образом информация об интересующем нас событии подана автором текста. Иначе говоря, достоверность рассматривается как *субъективная категория*. Как замечает В. З. Панфилов, достоверность следует рассматривать в привязке к субъекту речи, так как любое утверждение может быть ложным с точки зрения реального положения вещей [**12**].

Достоверность в этом смысле в разных источниках может включаться в субъективную модальность, выделяться в особый тип — персуазивную модальность, носить название эпистемической модальности или модальности истинности. Мы остановимся на термине «достоверность» (список по [**10**]).

Исходя из того, что автор берет на себя эпистемическое обязательство (см. п. 1.4), следует установить то, как сам автор оценивает степень достоверности сообщаемого. При этом мы противопоставляем информацию, поданную автором как *достоверную*, и информацию, поданную как *проблематично-достоверную* (термин А. В. Бондарко, [**10**]).

Приведем два примера высказываний, содержащих достоверную информацию (предикат целевой пропозиции выделен зеленым)[1]:

Генеральная прокуратура Украины ***возбудила*** *уголовное дело в отношении главы Минобороны России Сергея Шойгу*.

Министр иностранных дел Турции Мевлют Чавушоглу ***нанес*** *незапланированный* ***визит*** *в Иран*.

В высказываниях, содержащих проблематично-достоверную информацию, автор, как правило, эксплицитно указывает на более низкую степень достоверности с помощью лексико-граматических маркеров.

Чтобы более точно установить степень достоверности для проблематично-достоверной информации, необходимо проанализировать различные аспекты достоверности, которые мы будем называть *компоненты достоверности*.

[1]Подробнее о высказываниях, содержащих достоверную информацию, см. раздел 3

2.2. Компонент достоверности: эпистемическая оценка

Этот компонент объединяет два сходных субкомпонента:

- степень полноты знаний субъекта о сообщаемом;
- оценка субъектом вероятности того, что событие имело, имеет или будет иметь место.

Степень полноты знаний подразумевает, что субъект обладает неполной информацией о сообщаемом и предупреждает слушателя об этом (*«насколько мне известно», «пока нельзя утверждать, что...»* и пр). Этот субкомпонент относится в первую очередь к временному плану «настоящее» или «прошлое».

Оценка субъектом вероятности того, что событие имело, имеет или будет иметь место, применима для всех временных планов, однако в СМИ чаще всего встречается для ситуаций во временном плане «будущее». Этот субкомпонент также маркируется автором (*«маловероятно, что...», «кто-либо не исключает вероятности того, что...»*). Впрочем, существует маркеры, которые можно отнести и к первому, и ко второму субкомпоненту.

Компонент «эпистемическая оценка может принимать следующие значения.

2.2.1. *«Точно P»*

Это значение предполагает, что автор либо считает некоторую ситуацию высоковероятной, либо указывает на то, что имеющаяся в его распоряжении информация хоть и не полна, но, по его мнению, достаточна для того, чтобы утверждать что-либо.

Киркленд ***уверен****, что у Агуэро* ***конфликт*** *с Гвардиолой.*

К этому же значению мы относим ситуации во временном плане «будущее», не содержащие никаких дополнительных указаний на степень вероятности, кроме указаний на будущее время, так как будущее время имеет оттенок вероятности, см. пример:

Премьер-министр России Дмитрий Медведев ***посетит*** *Финляндию* ***с рабочим визитом****, где в городе Оулу встретится со своим финским коллегой Юхой Сипиля.*

2.2.2. *«Скорее P»*

В этом случае субъект считает наступление события высоковероятным, но не может утверждать с уверенностью, что оно произошло или произойдет.

Верховная Рада Украины ***должна утвердить*** *государственный бюджет на 2017 год до 22 декабря.*

2.2.3. *«P возможно»*

Это значение подразумевает, что субъект либо располагает неполной информацией о сообщаемом и на этом основании не может целиком взять на себя обязательство за его истинность, либо субъект оценивает вероятность наступления ситуации как среднюю.

В настоящее время мы активную фазу переговоров по этому вопросу, ***наверно****,* ***завершили****.*

Оганян ***может быть назначен*** *генсеком ОДКБ.*

2.2.4. *«Не исключено, что P»*

Это значение присваивается, если субъект либо берется утверждать что-либо с большой осторожностью ввиду недостатка информации, либо считает ситуацию, названную в P, маловероятной. Отношение субъекта к сообщаемому можно было бы описать так: «не исключено, что да, но скорее всего нет», например:

Пока нельзя утверждать*, что в смерти Халфана* ***виноват*** *соперник.*

Маловероятно, что *за его отставку в ближайшее время* ***проголосуют*** *в Раде.*

2.2.5. *«Исключено, что P»*

Это значение близко к «точно P» по тому, как сам автор оценивает полноту своих знаний о сообщаемом (информация, которой он обладает, неполна, но он считает ее достаточной). В то же время по субкомпоненту «оценка субъектом вероятности сообщаемого» это значение соответствует крайне низкой степени вероятности.

Сабра ***исключил любую вероятность того, что*** *оппозиция* ***согласится*** *на то, что Асад может сыграть какую-то роль в процессе политических преобразований.*

2.3. Компонент достоверности: отношение субъекта к цитируемому источнику

Следующий фактор, с помощью которого можно установить значение достоверности — это отношение субъекта к цитируемому источнику, что подразумевает разные степени доверия субъекта.

2.3.1. *Нейтральное отношение субъекта к источнику*

Важным маркером для этого значения выступают глаголы речи и ментального действия, вводящие чужую речь, и синонимичные им конструкции [**13**] (такие как *«заявил»*, *«утверждает»*, *«по словам кого-либо»*, *«как сообщает кто-либо»* и пр). Большинство глаголов речи (такие как *«рассказать»*, *«заявить»*, *«утверждать»*) лишь указывают на источник, из которого автор почерпнул приведенную информацию, то есть степень доверия автора к источнику никак не маркирована (автор «присоединяется к мнению источника», см. п. 1.5), см. пример:

***Как заявили в МИД**, Россия **не ведет** открытых **переговоров** по поводу дальнейшего восстановления военных баз на территории Кубы и Вьетнама.*

2.3.2. *Дистанцирование от того, что сообщается источником*

Определенные лексические маркеры («якобы, дескать» и некоторые другие) свидетельствуют о том, что субъект не присоединяется к содержанию цитируемой речи, а лишь передает ее читателю.

Так, в следующем примере автор указывает, что не берет на себя ответственность за содержание цитируемого сообщения, с помощью частницы «якобы»:

*Ранее СМИ сообщили, что Трамп **якобы планирует** посетить Россию в январе после своего официального вступления в должность.*

2.3.3. *Низкая степень доверия субъекта к цитируемому источнику*

Это значение может маркироваться множеством различных способов. Одним из таких способов является указание на сам источник, о котором и автору и читателю известно, что он не вызывает доверия. Самый типичный пример такого источника — «слухи».

***По слухам**, после сентябрьских выборов правительство Пермского края **отправится в отставку**.*

2.4. Компонент достоверности: оценка субъектом сообщаемого источником

Этот компонент близок к предыдущему, однако в данном случае субъект указывает не на свое отношение к самому источнику сообщаемого, а на отношение к *содержанию* сообщаемого (в какой мере он разделяет то, что содержится в речи источника). Типичный представитель этого компонента — это разные степени несогласия с тем, о чем сообщается в чужой речи.

В СМИ часто встречается комбинированный вариант субъекта речи для случаев, когда присутствует выраженная оценка того, что сообщается источником. Автор текста цитирует источник1, относительно которого он не выражает недоверия, а источник1 в свою очередь уже выражает свое несогласие с источником2. При этом источник2 часто не назван, а только подразумевается — «информация», «слухи», «версия» и пр. Маркерами несогласия выступают такие конструкции как «отрицать информацию/слухи/данные о том, что. . . », «опровергнуть» и пр.

Opel ***опроверг*** *слухи* ***о возвращении на рынок*** *России в 2019 году.*

Отметим, что несколько разных компонентов могут накладываеься друг на друга. Так, в следующем примере пропозиция «назначение Олега Ляшко на должность вицепремьере» находится в сфере действия маркеров разных компонентов: эпистемическая оценка («P возможно») и оценка сообщаемого («отрицает информацию о том, что. . . »):

В то же время председатель БПП ***отрицает информацию о том****, что на должность вице-премьера* ***может быть назначен*** *глава Радикальной партии Олег Ляшко.*

3. Средства выражения достоверной информации

Каждый из компонентов достоверности, описанных в главе 2, требует пристального изучения и выделения конкретных маркеров для каждого значения. В этой главе мы ограничимся описанием *базового*, основного случая: высокая степень достоверности информации, что подразумевает следующие значения описанных четырех компонентов: 1) эпистемическая оценка — не маркирована, 2) отношение к источнику — нейтральное, 3) оценка субъектом того, что сообщается источником — не выражена.

3.1. Немаркированный базовый случай

Самый типичный случай, который мы будем называть *немаркированным* — если ЦП

- выражена глаголом в личной форме,
- не входит в вопросительное предложение,
- не находится в синтаксически подчиненной позиции (т.е. не входит в придаточное),
- не осложнена вводными словами и

- не входит в сферу действия показателей чужой речи, например:
  Министр иностранных дел Турции Мевлют Чавушоглу ***нанес*** *незапланированный* ***визит*** *в Иран.*

3.2. Контексты, указывающие на высокую степень достоверности

Если одно из условий для немаркированного значения из предыдущего пункта нарушено, информация все-таки может иметь высокую степень достоверности. Такое возможно, если ЦП входит в один из следующих контекстов (сейчас мы рассматриваем случаи, когда ЦП выражена личной формой глагола или существительным).

3.2.1. *Вопросы местоименного типа и косвенные вопросы*

Если ЦП входит в вопрос с вопросительным местоимением, или в производный от него косвенный вопрос (вопросительное предложение, относящееся как зависимое к глаголу или глагольному слову).

Стало известно, ***когда состоится*** *экстренное* ***заседание Совбеза ООН по Сирии****.*

В этом примере зеленым выделены слова, выражающие пропозицию «заседание Совбеза», а подчинительный союз «когда» свидетельствует о том, что рассматриваемая пропозиция имеет статус «высокая степень достоверности».

Когда шли переговоры *с Ираном о снятии санкций?*

Этот пример аналогичен предыдущему, за исключением того, что «когда» в данном случае выступает в качестве вопросительного местоимения.

3.2.2. *Некоторые вводные слова и обороты*

На высокую степень достоверности ЦП могут указывать определенные вводные слова, которые указывают на то, что ассоциированная с ними пропозиция имела место [11]: *«к сожалению»*, *«честно говоря»*, *«тем не менее»* и др.

Тем не менее*, Сербия* ***не ввела санкции*** *против России.*

3.2.3. *Фазовые глаголы*

Фазовые глаголы в личной форме (т. е. глаголы, указывающие на начало, продолжение или завершение чего-либо), подчиняющие ЦП, также служат указанием на высокую степень достоверности информации. Как правило, это свойство они сохраняют и в отрицательной форме, кроме глаголов со значением «начала».

В Санкт-Петербурге ***завершилась встреча*** *президента Путина с Эрдоганом.*

3.2.4. *Глаголы с фактивной пресуппозицией*

Фактивные глаголы, т. е. глаголы, подразумевающие истинность подчиненной им пропозиции [14]. В текстах СМИ из фактивных глаголов чаще всего приходится сталкиваться с глаголами знания и ментального действия, такими как *«вспомнить»*, *«осознать»* и др; глаголами эмоциональной или этикетной реакции на событие — *«благодарить»*, *«сожалеть»* и др.

Джефф Гласс: я был просто ***поражен*** *и* ***шокирован отставкой*** *Поковича.*

Некоторые исследователи [14, 15] относят к фактивным также *глаголы сообщения*, такие как *«сообщить»*, *«информировать»*, *«предупредить»*.

Как ***сообщал*** *ONLINE.UA, ранее прокуратура* ***назначила*** *сына главы СБУ Олега Грыцака начальником отдела прокуратуры Киевской области.*

3.3. Нейтральный контекст

При анализе текста важно распознавать контексты, которые не влияют на степень достоверности ЦП, хотя и нарушают условия базового значения из п. 3.1. Другими словами, ЦП, входящую в такие нейтральные контексты, можно рассматривать как базовую.

3.3.1. *Нейтральное цитирование*

В первую очередь к таким контекстам следует отнести большинство случаев цитирования, которое вводится т. н. «глаголами говорения» (*«говорить»*, *«сказать»*, *«рассказать»*, *«утверждать»*, *«заявить»*...).

Как заявили в МИД*, Россия* ***не ведет*** *открытых* ***переговоров*** *по поводу дальнейшего восстановления военных баз на территории Кубы и Вьетнама.*

3.3.2. *Дискурсивные маркеры*

Нейтральным контекстом также являются *дискурсивные маркеры* — метатекстовые элементы, которые автор использует для структурирования своей речи.

Стоит отметить, что *ранее Сергей Лавров* ***провел*** *повторные* ***переговоры*** *со спецпосланником ООН по Сирии Стаффаном де Мистурой.*

3.3.3. *Придаточные присоединительные*

О нейтральном контексте можно говорить также в том случае, если ЦП выражена предикатом в присоединительном придаточном (в следующем примере присоединительное придаточное выделено полужирным).

Ранее стало известно, что в Женеве, ***где проходили рекордные по своей продолжительности переговоры на уровне глав внешнеполитических ведомств России и США****, Сергей Лавров накормил и напоил ожидавших делегации журналистов.*

3.4. Пример для иллюстрации контекстов для достоверной информации

Проиллюстрируем описанные контексты для высокой степени достоверности на примере, совмещающем в себе несколько контекстов. При рассмотрении предложения мы движемся «снизу вверх» по синтаксической структуре предложения, начиная от слова, выражающего ЦП — «переговоры».

Российский МИД удовлетворен продлением ***переговоров*** *в Йемене <...>, отмечается в сообщении на портале МИД РФ.*

Целевая пропозиция P: «переговоры — в Йемене» (предикатом этой пропозиции является «переговоры», именно его значение достоверности мы попытаемся определить).

Контекст, 1-й уровень: «*продлением* переговоров».

«Продление» — пропозиция, подчиняющая ЦП и выраженная существительным, образованным от фазового глагола, подразумевает высокую степень достоверности подчиненной пропозиции.

Контекст, 2-й уровень: (некто) «*удовлетворен* продлением переговоров».

«удовлетворен» — эта пропозиция описывает эмоциональную реакцию на ситуацию, также относится к фактивным предикатам, т.е. подразумевает высокую степень достоверности подчиненной пропозиции.

Контекст, 3-й уровень: «*Российский МИД* удовлетворен...».

Нейтральный контекст: указание на источник, не содержащее никаких показателей недоверия к нему.

Контекст, 4-й уровень: «... *отмечается в сообщении* на портале МИД РФ».

Нейтральный контекст: указание на источник, не содержащее никаких показателей недоверия к нему.

Таким образом, можно сделать вывод о том, что пропозиция «переговоры в Йемене» представлена субъектом как имеющая высокую степень достоверности.

Заключение

В статье предложено и определено понятие «достоверность» для задачи извлечения фактографической информации из текстов на естественном языке.

Структура достоверности представлена на основании трех типов отношений: отношение субъекта к сообщаемому, отношение субъекта к цитируемому субъекту и отношение субъекта к тому, что сообщается цитируемым субъектом.

Соответственно этим трем типам отношений выделены компоненты достоверности: эпистемический статус сообщаемого (который включает оценку субъектом речи полноты своих знаний о сообщаемом и оценку субъектом вероятности того, что сообщаемое имело, имеет или будет иметь место); степень доверия субъекта к цитируемому источнику и оценка субъектом того, что сообщается источником.

В последней главе описаны контексты, позволяющие извлечь информацию с «базовой» достоверностью, что является первостепенной задачей при извлечении фактографической информации.

В дальнейшем для развития этого подхода необходимо, во-первых, разработать принципы разметки достоверности с учетом разных значений компонентов. Во-вторых, исследовать способы выражения разных значений компонентов и разработать алгоритмы для установления связей между языковыми выражениями, маркирующими то или иное значение, и рассматриваемой пропозицией. В-третьих, разработать алгоритм автоматической оценки достоверности извлекаемой из текста фактографической информации.

Благодарности. *Авторы выражают искреннюю благодарность Игорю Владимировичу Трофимову за ценные замечания и помощь в подготовке статьи.*

Список литературы

[1] Н. А. Власова. «К проблеме разметки текстов на русском языке для задачи извлечения фактографической информации», *Программные системы: теория и приложения*, **5**:4(22) (2014), с. 67–82, URL: http://psta.psiras.ru/read/psta2014_4_67-82.pdf ↑ 267

[2] J. Pustejovsky, B. Ingria, R. Saurí, 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, pp. 545–557. ↑ 268

[3] R. Saurí, M. Verhagen, J. Pustejovsky. “Annotating and Recognizing Event Modality in Text”, *Proceedings of of the 19th International FLAIRS Conference*, FLAIRS 2006 (Melbourne Beach, Florida, USA, May 11–13, 2006), URL: https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-065.pdf ↑ 268

[4] J. Cardeñosa, A. Gelbukh, E. Tovar, *Universal Networking Language: Advances in Theory and Applications*, Research on Computing Science, vol. **12**, Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico, 2005, 443 p., URL: http://www.cicling.org/2005/UNL-book/UNL.pdf ↑ 268

[5] P. Jindal, D. Roth. “Extraction of Events and Temporal Expressions from Clinical Narratives”, *Journal of Biomedical Informatics*, **46** (2013), pp. S13–S19, URL: http://cogcomp.cs.illinois.edu/papers/JindalRo13d.pdf ↑ 268

[6] 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 ↑ 268

[7] Т. И. Краснова, *Субъективность — Модальность (материалы активной грамматики)*, Изд-во СПбГУЭФ, СПб., 2002, с. 94. ↑ 269

[8] Ш. Балли. *Общая лингвистика и вопросы французского языка*, ИЛ, М., 1955, 416 с. ↑ 269

[9] Дж. Лайонз, *Лингвистическая семантика*, Языки славянской культуры, М., 2003, с. 334, 339. ↑ 270,272

[10] А. В. Бондарко (ред.), *Теория функциональной грамматики: Темпоральность. Модальность*, Наука, Л., 1990, с. 69. ↑ 272,274

[11] Е. В. Падучева, *Семантические исследования. Семантика времени и вида в русском языке. Семантика нарратива*, Изд. 2-е, Языки славянской культуры, М., 2010, с. 326. ↑ 273,279

[12] В. З. Панфилов. «Категория модальности и ее роль в конституировании структуры предложения и суждения», *Вопросы языкознания*, 1977, №4, с. 37–48. ↑ 274

[13] С. В. Доронина. «Коммуникативы как средство выражения эпистемического смысла высказывания», *Вестник Омского ун-та*, 2013, №1, с. 76–81. ↑ 277

[14] Е. В. Падучева. *Динамические модели в семантике лексики*, Языки славянской культуры, М., 2004, 608 с. ↑ 280

[15] И. Б. Шатуновский. «Коммуникативные типы высказываний, описывающих действительность», *Логический анализ языка: истина и истинность в культуре и языке*, Наука, М., 1995, с. 158–165. ↑ 280

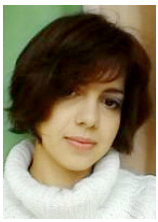
[16] А. А. Зализняк. ««Знание» и «мнение» в семантике предикатов внутреннего состояния», *Коммуникативные аспекты исследования языка*, Изд-во ин-та языкознания АН СССР, М., 1986, с. 4–15. ↑ 272

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

Пример ссылки на эту публикацию:

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

Об авторах:



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

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации

e-mail: yes@helen.botik.ru

Seda Egikian, Elena Suleymanova. *The actuality modality in the framework of the information extraction for texts written in a natural language.*

Abstract. The article deals with the "actuality" of the information extracted from the texts written in a natural language. The first part of the article is devoted to the basic notions we are using, such as proposition, modality and the speaker. In the second part the notion "actuality" is defined by describing its main components. The third part contains the list of the most important contexts for the basic case of "actuality". (*In Russian*).

Key words and phrases: natural language processing, automatics information extraction, modality, actuality.

References

[1] N. A. Vlasova. "On annotating Russian texts for information extraction task", *Programmnyye sistemy: teoriya i prilozheniya*, **5**:4(22) (2014), pp. 67–82 (in Russian), URL: http://psta.psiras.ru/read/psta2014_4_67-82.pdf

[2] J. Pustejovsky, B. Ingria, R. Saurí, 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, pp. 545–557.

[3] R. Saurí, M. Verhagen, J. Pustejovsky. "Annotating and Recognizing Event Modality in Text", *Proceedings of of the 19th International FLAIRS Conference*, FLAIRS 2006 (Melbourne Beach, Florida, USA, May 11–13, 2006), URL: https://www.aaai.org/Papers/FLAIRS/2006/Flairs06-065.pdf

[4] J. Cardeñosa, A. Gelbukh, E. Tovar, *Universal Networking Language: Advances in Theory and Applications*, Research on Computing Science, vol. **12**, Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico, 2005, 443 p., URL: http://www.cicling.org/2005/UNL-book/UNL.pdf

[5] P. Jindal, D. Roth. "Extraction of Events and Temporal Expressions from Clinical Narratives", *Journal of Biomedical Informatics*, **46** (2013), pp. S13–S19, URL: http://cogcomp.cs.illinois.edu/papers/JindalRo13d.pdf

[6] 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

[7] T. I. Krasnova, *Subjectivity — Modality (active grammar material)*, Izd-vo SPbGUEF, SPb., 2002, pp. 94 (in Russian).

[8] Ch. Bally. *Linguistique général et la linguistique française*, 2nd ed., P.U.F, 1944 (in French), 440 p.

[9] J. Lyons, *Linguistic Semantics: An Introduction*, Cambridge Approaches to Linguistics, Cambridge University Press, 1995, 376 p.

[10] A. V. Bondarko (red.), *The theory of functional grammar. Temporality. Modality*, Nauka, L., 1990, pp. 69 (in Russian).

[11] Ye. V. Paducheva, *Semantic research (Semantics of tense and aspect in Russian; Semantics of narrative)*, Izd. 2-ye, Yazyki slavyanskoy kul'tury, M., 2010, pp. 326 (in Russian).

[12] V. Z. Panfilov. "The modality category and its role in the constituating of the structure of a sentence and a proposition", *Voprosy yazykoznaniya*, 1977, no.4, pp. 37–48 (in Russian).

[13] S. V. Doronina. "The fact and opinion meanings of communicative words", *Vestnik Omskogo univesiteta*, 2013, no.1, pp. 76–81 (in Russian).

[14] Ye. V. Paducheva. *Dynamic models of lexical semantics*, Yazyki slavyanskoy kul'tury, M., 2004 (in Russian), 608 p.

[15] I. B. Shatunovskiy. "Communicative types of the reality describing utterances", *Logicheskiy analiz yazyka: istina i istinnost' v kul'ture i yazyke*, Nauka, M., 1995, pp. 158–165 (in Russian).

[16] A. A. Zaliznyak. ""Knowledge" and "belief" in the semantics of the predicates of the mental condition", *Kommunikativnyye aspekty issledovaniya yazyka*, Izd-vo in-ta yazykoznaniya AN SSSR, M., 1986, pp. 4–15 (in Russian).

Sample citation of this publication:

Seda Egikian, Elena Suleymanova. "The actuality modality in the framework of the information extraction for texts written in a natural language", *Program systems: Theory and applications*, 2016, **7**:4(31), pp. 267–286. (*In Russian*). URL: http://psta.psiras.ru/read/psta2016_4_267-286.pdf