ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(31), 2016, с. 209–228

УДК 004.89:004.912

Е. А. Сулейманова

О двух видах текстовых временных координат

Аннотация. В статье предложена систематизация текстовых выражений, которые в рамках задачи автоматического извлечения темпоральной информации из текста принято рассматривать в качестве временных координат событий. Выделены два типа показателей времени, различающихся способом референции к временным сущностям. Исследована специфика нормализации (определения абсолютного значения) выражений каждого типа

Ключевые слова и фразы: автоматическая обработка текста, анализ темпоральной информации, нормализация контекстно-зависимых показателей времени.

Введение

Автоматическая обработка содержащейся в тексте темпоральной информации является неотъемлемой частью извлечения из текста фактической информации. В этом контексте обработка темпоральной информации рассматривается как одна из задач извлечения (распознавания и нормализации) именованных сущностей. Объектом извлечения выступают т. наз. темпоральные выражения (temporal expressions) — лингвистические конструкции (как правило, именные группы, предложные группы, наречия), сообщающие, когда нечто имело место, как долго нечто продолжалось или как часто нечто имеет место [1].

Обычно выделяют три базовых типа темпоральных выражений:

- Временная координата — выражение, которое соотносится с конкретной отметкой на одной из календарных шкал; альтернативные термины — *Date* [2], *time point* или *time-referring expression* [3], *coordinate* [4].

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

- Длительность — *Duration*, *period-referring expression* [3], *quantity* [4]; различают локализованные (anchored) и нелокализованные (unanchored) длительности [1].
- Множество — *Set* [2], *set expression* [3], *enumeration* [4].

1. Абсолютные и контекстно-зависимые указания на время: возможные классификации

Нормализация темпорального выражения типа «временная координата» — это представление его абсолютного значения в соответствии с некоторым требуемым форматом. Поскольку стандартное представление значения выражений типа *1 января 2001 года*, *01.01.2001*, *2001 год* и т. п. представляет собой техническую задачу, тема нормализации обычно затрагивается в связи с выражениями, абсолютное значение которых не столь очевидно (*вчера*, *в прошлый понедельник*, *три года назад*). При этом под нормализацией часто понимают не столько приведение значения временного выражения к стандартной форме записи, сколько его интерпретацию (вычисление абсолютного значения).

Множество темпоральных выражений, относимых к временным координатам (первый тип в приведенной выше классификации), довольно разнородно. В публикациях приводятся различные способы их систематизации; поскольку авторы не придерживаются единой терминологии, одно и то же явление при разных подходах нередко именуется по-разному, а один и тот же термин используется для обозначения разных категорий. Пожалуй, единственное, в чем единодушны все авторы, — это в том, что на верхнем уровне таксономии темпоральные выражения следует разделить на две категории.

К первой относятся указания на время, интерпретация которых не требует обращения к контексту: каждое такое выражение содержит в явном виде всю информацию, необходимую для идентификации его абсолютного календарного значения. Такие выражения называют *абсолютными* или *полностью определенными* (*absolute*, *fully-specified* [2, 5, 6]), а также эксплицитными (*explicit* [7–9]). К этой группе относят разнообразные по масштабу и формату записи полные даты. Такие выражения успешно нормализуются стандартными методами извлечения информации.

Темпоральные выражения второй категории объединяет одно общее свойство: вне контекста их однозначная интерпретация невозможна. Это выражения следующего вида:

(*1*) *yesterday*, *today*, *tomorrow*;

(*2*) *next week*, *last year*, *the following year*;
(*3*) *this Friday*, *last Tuesday*;
(*4*) *a year ago*, *three months later*;
(*5*) *two days before departure*;
(*6*) *June 24*, *Monday*;

Самое общее название для этой категории временных показателей — контекстно-зависимые (*context-dependent* [**5**]), но многие авторы используют альтернативную терминологию: *индексальные* (*indexical* [**8**]), *имплицитные* (*implicit* [**9**]), *относительные* (*relative* [**6**]), *недоопределенные* или (*контекстно*)-*недоопределенные выражения* (*underspecified (contextually underspecified) temporal expressions* [**2**, **7**, **10**]).

Многие из этих терминов неоднозначны и могут использоваться для обозначения отдельных групп внутри категории контекстнозависимых выражений.

Довольно распространена точка зрения, противопоставляющая группу случаев типа (*6*) всем прочим видам контекстно-зависимых выражений. При таких подходах подкатегорию, объединяющую случаи (*6*) — неполные (*incomplete* [**11**]), недоопределенные «в узком смысле» (*underspecified)* [**3**]), частично-определенные (*partially specified* [**10**, **12**, **13**]) выражения, — не принято классифицировать далее. Случаи с (*1*) по (*5*) объединяют в одну подкатегорию — *относительные выражения* (*relative* [**10**, **11**, **14**]) или *выражения «со смещением»* (*offset expressions* [**3**]). Категория выражений «со смещением», или «функциональных» [**3**], названа так, поскольку значение таких выражений может быть описано функцией прибавления или вычитания (смещение может быть и нулевым). Внутри этой подкатегории авторы различают выражения *дейктические*, *анафорические* (и те, и другие есть среди примеров (*1*)–(*4*)) и *привязанные к событиям* (*event-based* — пример (*5*)).

Согласно другой точке зрения [**5**], все контекстно-зависимые показатели признаются выражениями с разного рода смещением (характер отсылки к времени отсчета не учитывается):

(*1*) выражения «с эксплицитным смещением» относительно времени отсчета (Explicit offsets from reference time) — *yesterday*, *today*, *tomorrow*,
(*2*) выражения «с позиционным смещением» относительно времени отсчета (Positional offsets from reference time) — *next month*, *last year*, *this coming Thursday*,

(*3*) выражения «с имплицитным смещением» (Implicit offsets) — *Thursday*, *February*.

(Отметим в скобках, что семантическое значение неполных выражений далеко не всегда может быть описано как смещение, пусть и нулевое).

Еще один подход [7] использует в качестве основания для классификации характер отсылки и, независимо от поверхностной полноты и семантики, противопоставляет дейктические (*deictic*) и относительные (*relative*) выражения. Дейктические определяются как находящиеся в специфическом отношении с моментом речи (*tomorrow*, *last year*), тогда как относительные (заметим, что это уже третье значение слова *relative*) — с текущим темпоральным фокусом дискурса, т.е. анафорические; при этом в качестве примера относительных выражений авторы приводят неполное *on Friday*. Если учесть, что а) неполные и неоднозначные выражения способны употребляться как с дейктической, так и с анафорической отсылкой (*on Friday*, *the following year*, ср. *в пятницу*, *в следующем году*, *через неделю*), б) некоторые «однозначно» дейктические выражения (напр., выражения с *ago*) в дискурсе допускают переориентацию с момента речи на другой временной «якорь» [**14**], то разделение показателей времени на дейктические и относительные скорее является классификацией их возможных употреблений.

В заключение, в связи с темой настоящей статьи, заметим, что ни одна из рассмотренных классификаций не дифференцирует выражения вроде *last year*, *next month*, с одной стороны, и *a year* (*x years*) *ago*, *a month* (*x months*) *later* — с другой (хотя некоторая специфика случаев второго рода обсуждается в связи с задачей нормализации, о чем будет упомянуто в соответствующем разделе).

Предлагаемая в настоящей работе систематизация показателей времени ориентирована на задачу нормализации (понимаемую как вычисление абсолютного значения), учитывает их семантическую структуру и особенности референции. Контекстная зависимость или независимость, так же как и характер отсылки к времени отсчета, не являются классифицирующими признаками самих показателей времени, а характеризуют текстовую форму выражения их семантических компонентов.

2. Способы референции к временным сущностям: именная и адвербиальная референция

Прежде всего, нельзя не заметить, что в класс временных координат оказываются включены темпоральные выражения различной синтаксической природы.

Будем различать именные и адвербиальные темпоральные выражения.

Текстовая дата (отвечает на вопрос «что?») — это именное выражение или его буквенно-цифровой аналог, служащие для обозначения в тексте календарного интервала — деления одной из календарных шкал. Референтом (референциальным интервалом) текстовой даты, таким образом, является конкретный единичный календарный интервал некоторого масштаба. Текстовая дата может быть как абсолютной, или полной (*01.04.2001*, *апрель 2001 года*, *весна 2001 года*, *второй квартал 2001 года*, *2001 год*), так и контекстно-зависимой, или контекстной [16] (*то же число прошлого месяца*, *апрель*, *следующий год*).

Темпоральный адвербиал — локализатор (ТА-локализатор). Функцию временной локализации события в предложении выполняет не именная группа, а обстоятельство времени, отвечающее на вопрос «когда?». Если *5 («пятое») апреля* — это текстовая дата, то *5 («пятого») апреля* — это адвербиальное выражение, принадлежащее к классу темпоральных адвербиалов — локализаторов, т. е. обстоятельств времени в собственном смысле слова (*показателей времени* [**17**], *Adv temp* [**18**]). Мы придерживаемся точки зрения [**19**, **20**], согласно которой ТА-локализатор обладает собственным референциальным значением (собственным временем).

3. Семантическая структура ТА-локализатора

Собственное время ТА-локализатора всегда задается относительно некоторого фиксированного времени, что позволяет привязать референциальный интервал ТА-локализатора к временной оси (Е. В. Падучева назвала это время «своего рода опорным временным моментом» показателя времени [**17**]). Мы будем использовать в этом смысле более абстрактный термин *опорное время* ТА-локализатора. В роли опорного времени могут выступать «момент речи», календарный интервал, время события.

В семантике любого ТА-локализатора присутствуют два компонента:

(*1*) компонент, идентифицирующий опорное время адвербиала (опорный компонент) и
(*2*) компонент, позиционирующий время адвербиала относительно его опорного времени (позиционирующий компонент).

В настоящей работе рассматриваются ТА-локализаторы, используемые для датирования (календарной привязки) событий в тексте[1]. Нормализовать такой адвербиал означает идентифицировать и записать в стандартизованном виде значение его референциального интервала, который, очевидно, представляет собой отрезок на одной из календарных шкал. Поскольку формат записи для нас сейчас не важен, под *нормализацией* в настоящей работе мы будем понимать идентификацию референциального значения.

В связи с задачей нормализации рассматриваемый класс показателей времени целесообразно разделить на две категории, условно называемые далее *ТА-локализаторами календарного и квазикалендарного типа.* Основанием для такого разделения служит наше представление о том, что, в силу разной семантики ТА-локализаторов этих двух категорий, их собственное время имеет несколько разную природу.

Эта классификация разделяет временные координаты вне зависимости от их абсолютного или контекстно-зависимого характера: среди ТА-локализаторов обеих категорий есть как абсолютные, так и те, чья интерпретация возможна только в контексте (хотя квазикалендарные ТА больше тяготеют к контекстной зависимости в той или иной форме).

4. ТА-локализаторы календарного типа

К этой категории относятся выражения *15-го декабря 2010 года*, *в 1991 году*, *на прошлой неделе*, *в следующем квартале*, *в будущем году*, *весной*. Опорный компонент выражен текстовой датой, поэтому опорное время представляет собой календарный интервал[2]. Время самого адвербиала позиционируется как «совпадающее с опорным временем». Таким образом, собственное время такого адвербиала

[1]Адвербиалы с неопределенным референциальным интервалом «сейчас», «в настоящее время», «в прошлом», «раньше» и т. п. в качестве средств календарной привязки не рассматриваем.

[2]Здесь и далее под «календарным интервалом» мы понимаем единичный календарный интервал, т. е. единичный отрезок календарной шкалы некоторого масштаба.

совпадает с точным календарным интервалом, упомянутым посредством текстовой даты. Эту категорию адвербиалов мы называем *ТА-локализаторами календарного типа* (далее КТА).

КТА может быть как контекстно-независимым (если входящая в его состав текстовая дата полностью определена), так и контекстно-зависимым (если опорный компонент выражен контекстной датой). Так, время контекстно-независимого адвербиала *в 2015 году* всегда совпадает с календарным интервалом масштаба «год», имеющим идентификатор «2015». Время контекстно-зависимого КТА в прошлом году — с интервалом масштаба «год», значение которого может быть установлено только с учетом контекста конкретного употребления адвербиала (*в году, предшествующем году, содержащему «момент речи»*).

Дейктические и анафорические наречия *сегодня*, *вчера*, *позавчера*, *завтра*, *послезавтра*, *накануне*, *назавтра* мы относим к контекстно-зависимым КТА— считаем, что опорное время в них выражено посредством «встроенной» текстовой даты масштаба «день» (сегодня = ⟨в *день, содержащий «момент речи»*⟩, завтра = ⟨в *день, следующий за днем, содержащим «момент речи»*⟩, назавтра — ⟨в *день, следующий за упомянутым временем*⟩).

Нормализация КТА состоит в идентификации календарного интервала его текстовой даты, которая в случае контекстнозависимого КТА является контекстной.

4.1. Виды контекстных дат

4.1.1. *Контекстные даты с отсылочными показателями*

Это текстовые даты с индексальным адъективом *этот* (*же*), *прошедший*, *ближайший*, *позапрошлый* и т. п. Считаем, что в наречия *сегодня*, *вчера*, *накануне* и т. п. «встроены» контекстные даты масштаба «день» с отсылочным показателем.

Значение контекстной даты с отсылочным показателем может быть представлено как некоторая функция от времени отсчета, а отсылочная лексема указывает на тип этой функции и содержит имплицитную отсылку к времени отсчета. Эта отсылка может быть анафорической (к времени, упомянутому в предшествующем тексте) или дейктической (к «моменту речи»), например: прошлый месяц = ⟨*месяц, предшествующий месяцу, содержащему «момент речи»*⟩.

4.1.2. *Неполные контекстные даты*

В контекстных датах возможны два типа неполноты, различающихся механизмом восстановления: неполнота-пропуск и неполнота-умолчание.

Неполнота-пропуск имеет место в тех случаях, когда в неполном упоминании опускается повторяющаяся часть, общая для него и для другого упоминания, например: *в 2001 году* рост составил 5%, *в 2002 [году]* — 15%.

Для *неполноты-умолчания* характерно наличие некоторого смыслового отношения между неполным упоминанием и коммуникативной ситуацией или ситуацией, описываемой в самом тексте. Неполные даты, в которых имеет место неполнота-умолчание, аналогичны контекстным датам с отсылочными показателями. Для нормализации (восстановления неполноты) необходима идентификация отсылочной функции — отношения между референтом неполной даты и «моментом речи» (дейксис) или референтом антецедента (анафора). Например, неполная дата *апрель*, входящая в КТА *в апреле*, может содержать дейктическую отсылку — и тогда имеется в виду ⟨*апрель, ближайший к «моменту речи»*⟩. Пример анафорической отсылки: *[В этом году]. . . в апреле* (в данном случае речь идет об апреле, включенном во время антецедента).

Подчеркнем: независимо от формы контекстной зависимости, значение текстовой даты всегда представляет собой точный календарный интервал вполне определенного масштаба.

5. Квазикалендарные ТА-локализаторы

Вторая категория ТА-локализаторов включает выражения *месяц назад*, *за неделю до выборов*, *два года спустя*, *5 лет назад*, *через два месяца после этого* и т. п. Референциальный интервал такого адвербиала задается через отношение предшествования или следования с указанием величины смещения относительно опорного времени. От первой, календарной, группы эти выражения отличает одно обстоятельство, имеющее принципиальное значение при нормализации: адвербиал такого типа в явном виде содержит лишь указание на то, где на календарной шкале располагается его референциальный интервал относительно опорного времени, но умалчивает о том, что именно он собой представляет. Такие адвербиалы очень часто не предполагают референции к точному календарному интервалу. Отсюда и второе

название, которое мы используем для этой категории, — *квазикалендарные ТА-локализаторы* (ККТА). Ср. КТА *в прошлом месяце* и ККТА *месяц назад*. В первом случае время ТА совпадает с точным календарным месяцем: если сейчас декабрь, то в прошлом месяце — это однозначно в ноябре. Для второго случая, очевидно, такую интерпретацию нельзя признать корректной.

Адвербиалы этого типа, как правило, контекстно-зависимы. Указание на опорное время в большинстве случаев имплицитно (опорный компонент синтаксически не выражен, а опорное время совпадает с «моментом речи» или временем упоминанияантецедента). Если же опорный компонент выражен на поверхности, то он обычно представляет собой именную группу событийнотемпоральной семантики, которая также часто содержит элемент индексальности (отсылку к текстуальному контексту или коммуникативной ситуации), который обеспечивает опорному компоненту определенность — *за неделю до выборов / до Нового года* (см. об индексальном компоненте в обозначениях времени у Т. В. Булыгиной и А. Д. Шмелёва [**21**], с. 379). Текстовые даты в качестве поверхностного выражения опорного компонента в ККТА маловероятны, хоть и возможны (*за неделю до этого дня*).

6. Нормализация квазикалендарных ТА-локализаторов

6.1. Специфика нормализации ККТА.

Отправной масштаб

Позиционирующий компонент смысловой структуры ККТА всегда представлен тремя подкомпонентами, задающими параметры смещения относительно опорного времени. Таким образом, семантическая структура ККТА представляется следующим образом:

(*1*) опорный компонент, позволяющий идентифицировать опорное время;
(*2*) позиционирующий компонент:
- направление смещения
- единица смещения (масштаб[3] календарного интервала)

[3]Термин «масштаб» мы используем без строгого определения, понимая под ним относительную длительность календарных интервалов.

- числовая величина смещения Для использования в качестве первого аргумента функции, вычисляющей время ККТА (аргумента, к которому прибавляется или от которого отнимается величина смещения), опорное время должно быть приведено к *отправному интервалу* — календарному интервалу определенного масштаба.

Отправной интервал либо включает опорное время, либо совпадает с ним. Необходимость выбора масштаба отправного интервала (далее *отправного масштаба*) — это первое обстоятельство, отличающее нормализацию ККТА от нормализации КТА. В КТА масштаб первого аргумента функции, он же масштаб результата нормализации, предопределен — задан явно текстовой датой в самом адвербиале. *В прошлом* (*позапрошлом, позапозапрошлом*) *году* — это всегда ⟨*в году, отстоящем влево от текущего года на один год* (*два, три года*)⟩. Иная ситуация в случае с ККТА: выбор отправного масштаба для вычислений не очевиден даже в том случае, если в тексте явно указан масштаб опорного времени (отправной масштаб не обязательно совпадает с ним).

Кроме того, текстовое указание на величину смещения далеко не всегда должно пониматься буквально. *Через неделю* может означать «ровно через семь дней», но допускает также и приблизительное толкование (⟨*примерно через неделю*⟩, т.е. ⟨*через шесть–восемь дней*⟩). Второе различие между двумя рассматриваемыми типами адвербиалов состоит в следующем: КТА всегда нормализуется в точный календарный интервал[4], а время ККТА в большом числе случаев в принципе не поддается точной идентификации, поэтому и нормализация может быть только оценочной. Если использовать аналогию со статистической оценкой, то результатом нормализации ККТА может быть либо точка (точный календарный интервал), либо некоторый доверительный интервал (произвольный отрезок календарной оси), который с приемлемым уровнем доверия включает в себя референциальный интервал нормализуемого адвербиала.

6.2. Нормализация ККТА: state of the art

В инструкции по разметке в соответствии со стандартом TIMEX2 [1] выражения, относимые нами к ККТА, обсуждаются в разделе

[4]Неточная интерпретация КТА также возможна, но только при наличии специальных маркеров и только в контекстно-независимых случаях (*примерно в 1990 году*, но не *примерно в прошлом году*).

Indeterminate Precision («неопределенная точность»). По мнению авторов, в отличие от выражения *a year ago today*[5], выражения вроде *a year ago* неточны вследствие отсутствия явной точки привязки к календарной оси: неясно, что имеется в виду — «ровно год назад (если отсчитывать от сегодняшнего дня)» или просто «(в) прошлом году». Две возможных интерпретации допускаются и у выражения *in a week* (*через неделю*, с отсылкой к «моменту речи»): оно может означать «ровно через семь дней от настоящего момента» или «(в) любое время на следующей неделе»[6]. В связи с этим, при нормализации таких выражений предлагается использовать *правило «масштаба выражения»* (Expression Granularity Rule): в отсутствие в контексте явной календарной привязки определять точность значения следует по существительному-вершине[7]. Допускаются отклонения от правила: десятилетие и век не возбраняется интерпретировать как десять и сто лет.

Выражения с величиной смещения в тысячи и миллионы лет (назад) нормализуются как «геологические эры», с использованием специальной нотации (*210 million years ago* — «MA210»). Примечательно, что предложение *The king lived 4,000 years* ago приводится в инструкции в качестве примера нормализации в конкретный год до нашей эры (2001 год до н.э., если «момент речи» приходится на 1999).

Резюмируем, используя нашу терминологию: в соответствии со стандартом TIMEX2, ККТА (за исключением отсылок к доисторическому прошлому) всегда нормализуются в точный календарный интервал — либо масштаба единицы смещения, либо масштаба точки привязки (если последний упомянут явно). Следовательно, все ККТА с «моментом речи» в качестве опорного времени (*неделю назад*, *месяц назад*, *сто лет назад*) полагается нормализовать в масштабе смещения (в «неделю», «месяц», «год»). Неточная нормализация не предусматривается.

---

[5]Точным аналогом такой конструкции в русском языке было бы выражение «сегодня ровно год, как... », не являющееся ККТА (важно присутствие слова «сегодня», ср. ККТА с показателем точности «ровно год назад», которое не содержит указания на масштаб опорного времени).

[6]Unlike “a year ago today”, which we saw earlier, expressions like “a year ago” are imprecise because they lack an explicit anchor. The writer could mean “a year ago today”, (which would be 1998-07-15 if today were 1999-07-15) or just “last year” (**??**). “In a week” could mean precisely seven days from now or any time in the following week [цитируется по источнику [1] в списке литературы, с. 22].

[7]“Expression Granularity” Rule: When no explicit anchor exists in the document context, use only the head noun to determine the precision of the VAL [там же, с. 23].

Формат TIMEX3 стандарта TimeML [2] несколько отличается от TIMEX2 — эти отличия, в частности, затрагивают и способ разметки ККТА: фрагмент, соответствующий смещению, подлежит разметке как отдельное выражение типа «длительность». Но на методах нормализации (вычисления значения) это не отразилось: за масштаб результата нормализации все так же принимается либо масштаб точки привязки, либо масштаб смещения.

6.3. Режимы нормализации ККТА

Введем понятие *режима нормализации* ККТА. Режим нормализации определяется двумя составляющими: способом выбора отправного масштаба и методом оценки — точным (точечным) или неточным (интервальным).

В качестве отправного масштаба для нормализации ККТА может быть выбран:

(*1*) масштаб единицы смещения;
(*2*) масштаб, меньший, чем масштаб единицы смещения (возможность такого выбора ограничена снизу *минимальным отправным масштабом, см. далее*);
(*3*) масштаб, больший, чем масштаб единицы смещения.

6.3.1. *Минимальный масштаб ККТА*

Прежде чем перейти к обсуждению режимов нормализации, введем понятие *минимально возможного отправного масштаба для нормализации данного ККТА*, или просто минимального масштаба ККТА. У ККТА с явно упомянутым опорным временем — календарным интервалом — минимальный масштаб совпадает с масштабом опорного времени. В тех случаях, когда опорное время явно не упомянуто, минимальный масштаб ККТА считается равным (через «или»):

(*1*) масштабу «день» — для всех ККТА с «моментом речи» в качестве опорного времени (далее для краткости такие ККТА будем называть *дейктическими*), например: *две недели назад*;
(*2*) масштабу «день» — для любого ККТА, в котором единица смещения имеет масштаб «день»;
(*3*) масштабу текстовой даты — временной координаты события, время которого служит опорным временем ККТА, например: [*Это случилось в 2001 году.*] *За полвека до этого стало известно, что...* Минимальный масштаб ККТА за полвека до этого — «год»;

(*4*) масштабу единицы смещения — в тех случаях, когда опорным временем ККТА является время события без явной календарной локализации.

6.3.2. *Точная и неточная нормализация*

Результат неточной нормализации адвербиала, как уже говорилось, представляет собой некоторый интервал на календарной оси (не совпадающий с единичным календарным интервалом), с большой вероятностью содержащий время адвербиала.

Говоря о неточной нормализации ККТА, мы имеем в виду не те случаи, в которых на неточность интерпретации указывают специальные языковые средства, а ККТА с неточным значением «по умолчанию». Способ представления неточного значения, повидимому, может быть одинаков для обоих явлений — в виде центра и окрестности. При этом если маркированная неточность может быть как «двухсторонней» (*примерно*, *приблизительно три года назад*, *года три назад*), так и «односторонней» (*почти три года назад*, *менее*, *более трех лет назад*), неточность умолчательная может быть, по-видимому, только «двухсторонней» — центр и симметричная окрестность.

Аналогично, применимость точных режимов обсуждается далее по отношению к ККТА, не содержащим никаких специальных показателей точности (*ровно*, *точно*).

6.3.3. *Точная нормализация в масштабе смещения («МС-точный режим»)*

За отправной масштаб принимается масштаб единицы смещения. ККТА нормализуется в точный календарный интервал масштаба единицы смещения, отстоящий от отправного интервала (того же масштаба) на заданное число единиц смещения.

МС-точный режим применим к большинству ККТА (как тех, у которых масштаб смещения совпадает с минимальным масштабом, так и тех, у которых это не так).

Примеры. Все ККТА с единицей смещения масштаба «день» (независимо от способа выражения опорного времени и величины смещения) — *через n дней (после X)*, *n дней назад*, *за n дней (до X)*, *n дней спустя*, *n днями раньше*, *n днями позже* — вполне

корректно нормализуются в «день, отстоящий на n дней влево или вправо от отправного дня»[8].

Другие примеры (здесь и далее в примерах в квадратных скобках приводится фрагмент предшествующего контекста, задающий минимальный отправной масштаб ККТА): [*На прошлой неделе*...]. *За неделю (две, три недели, пять недель) до этого* [...]. Время ККТА — неделя, предшествующая отправной неделе (отстоящая влево от нее на две, три, пять недель).

[... в 1999 году.] *За год* (*два, четыре года, пять, девять, одиннадцать лет, тридцать три года*) до этого, *спустя год* (*два, четыре года, пять, пятнадцать лет*) — все такие случаи вполне корректно нормализуются точно в «год».

Примеры дейктических ККТА, у которых масштаб смещения больше минимального масштаба и которые также поддаются нормализации в МС-точном режиме: *через четыре недели*, *пять месяцев назад*, *через 15 лет*, *через три столетия* («неделя», «месяц», «*год*», «*век*»).

При всей универсальности МС-точного режима, есть целый ряд случаев, в которых нужно, возможно или предпочтительно использовать другие режимы нормализации. К таким случаям относятся:

- ККТА с единичной числовой величиной смещения, у которых масштаб смещения больше минимального масштаба (*неделю назад*, *год назад*) — для них нормализация в МС-точном режиме, скорее, неприемлема (см. далее);
- ККТА с круглой числовой величиной смещения (*десять лет назад*, *50 лет спустя*) допускают неточную нормализацию в масштабе смещения (в МС-неточном режиме, см. далее);
- дейктические ККТА с очень большой величиной смещения (*через пятьсот лет*, *пять тысяч лет назад*) — для них предпочтителен режим с увеличением масштаба смещения (также описан далее).

6.3.4. *Неточная нормализация в масштабе смещения («МС-неточный*

режим») Для к ККТА с круглой числовой величиной смещения МС-точный режим может оказаться излишне «прицельным».

[8] за исключением выражения «через [один] день», которое в языке, пожалуй, не может иметь интерпретации «в последующий день» (ср. «день спустя»), а парадоксальным образом имеет то же значение, что и «через два дня».

МС-неточный режим — это нормализация в интервал, полученный применением МС-точного режима, плюс-минус погрешность. Величина погрешности зависит от величины смещения, но масштаб ее (погрешности) равен масштабу смещения.

Важно: длительность интервала, в который нормализуется ККТА при МС-неточном режиме, всегда превышает длительность единицы масштаба смещения.

Примеры. *Двадцать лет назад*, *сто лет назад* нормализуется в интервал «год, отстоящий от отправного года на 20 (??) лет, плюс-минус погрешность x лет». Значение x для сто лет назад, очевидно, больше, чем для двадцать лет назад.

6.3.5. *Нормализация с уменьшением масштаба*

Речь идет о нормализации в масштабе, меньшем, чем масштаб смещения. Предполагаем, что в отсутствие маркера точности для такой нормализации предпочтительнее интервальный (неточный) метод.

Ядро класса применимости неточной нормализации с уменьшением масштаба составляют ККТА с единичной величиной смещения, в которых масштаб смещения больше минимального масштаба. В таких случаях нормализация в МС-точном режиме слишком груба и не всегда корректна — например, если опорное время оказывается близко к границам календарного интервала масштаба смещения. Так, *через неделю* ≠ «на следующей неделе» (а именно таков содержательно был бы результат МС-точной нормализации), поскольку, будучи сказано в понедельник, может относиться к воскресенью той же недели. Аналогично, *через месяц* ≠ «в следующем месяце», год назад ≠ «*в прошлом году*».

При неточной нормализации с уменьшением масштаба за отправной масштаб может быть принят любой масштаб, меньший, чем масштаб смещения, — вплоть до минимального масштаба ККТА. Последнее целесообразно только при небольшой разнице между масштабом смещения и минимальным масштабом, например *неделю назад* («неделя»-«день»), дейктическое *через месяц* («месяц»-«день»).

Результат нормализации в режиме с уменьшением масштаба — точный интервал выбранного масштаба, отстоящий от отправного на величину смещения, плюс-минус погрешность. Масштаб погрешности совпадает с отправным масштабом. Что касается числовой величины погрешности, то она может быть тем больше, чем больше разница между масштабом смещения и отправным.

Важно: длительность интервала, в который нормализуется ККТА в режиме с уменьшением масштаба, всегда строго меньше длительности единицы масштаба смещения.

Примеры. Дейктическое *через неделю* можно нормализовать в минимальном масштабе в интервал «день, отстоящий на неделю вправо от отправного дня, плюс-минус один день». Дейктическое *год назад*, очевидно, предпочтительнее нормализовать не в минимальном масштабе, а в масштабе «месяц» — в интервал «месяц, отстоящий от отправного месяца влево на год, плюс-минус один месяц».

Пограничные случаи.

При не единичной, но малой («2»-«3») числовой величине смещения и несовпадении масштаба смещения и минимального масштаба, очевидно, допустима нормализация и в МС-точном режиме. По крайней мере, в ходе небольшого эксперимента некоторые из опрошенных склонны были трактовать *2 года назад* как «в позапрошлом году». Любопытно, что в сфере действия показателя неточности такие ККТА ведут себя иначе, чем ККТА со средним и большим числом смещения. Ср., например, *примерно три года назад* и *примерно 5 лет назад*. В первом случае естественной представляется интерпретация «ровно три года назад плюс-минус небольшая погрешность», т.е. скорее в режиме с уменьшением масштаба. Второе же предполагает неточную МС-нормализацию («4–6 лет назад»).

6.3.6. *Нормализация с увеличением масштаба смещения*

Если величина смещения представляет собой сотни или тысячи лет, то нормализация, даже неточная, в масштабе смещения («год») окажется неоправданно «прицельной».

Попытаемся, к примеру, нормализовать пятьсот лет назад, сказанное в октябре 2016 года, в МС-неточном режиме. Получим интервал «год, отстоящий от опорного на 500 лет, плюс-минус погрешность масштаба «год»», т. е. «1516 год плюс-минус погрешность масштаба «год»». При небольшой величине погрешности оценка будет почти точечной (мы получим интервал длиной в несколько лет — и с высокой вероятностью ошибемся). Если же увеличивать размер погрешности (а по умолчанию она симметричная), то прежде, чем интервал покроет XVI век, он захватит большую часть XV века, что (интуитивно) тоже не слишком хорошо.

В таких случаях имеет смысл преобразовать величину смещения, увеличив ее масштаб. ККТА *пятьсот лет* назад преобразуется

в ККТА *пять веков назад* (с масштабом смещения «век»). К полученному ККТА вполне применим МС-точный режим нормализации. В результате получим интервал «век, отстоящий от опорного века на пять веков». Таким образом, если сейчас XXI век, то пятьсот лет назад — это в XVI веке, что, скорее всего, и имелось в виду.

Заключение

В статье предложено различать два класса темпоральных адвербиалов, использующихся для календарной привязки событий, — календарные и квазикалендарные. Задача нормализации квазикалендарных адвербиалов (понимаемая как определение абсолютного значения) оказывается значительно сложнее, чем нормализация контекстно-зависимых адвербиалов календарного типа. Наряду с общими подзадачами (установление типа отсылки, идентификация временных ориентиров — опорного времени, времени отсчета), она требует выбора подходящего режима нормализации и определения его параметров (отправного масштаба, если выбран режим не в масштабе смещения; размера окрестности для неточных режимов). На выбор режима нормализации конкретного ККТА влияют:

- соотношение минимального масштаба ККТА и масштаба единицы смещения;
- диапазон числовой величины смещения;
- дополнительные признаки числовой величины смещения.

Между областями применимости разных режимов нормализации ККТА не всегда можно провести четкие границы, но для каждого режима описаны случаи, составляющие ядро его класса. Пограничные случаи допускают нормализацию в разных режимах.

Список литературы

[1] L. Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. Technical report, MITRE, September. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf. ↑ 209,210

[2] 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, Mani, L., Pustejovsky, J., Gaizauskas, R., eds. Oxford University Press, 2005.. ↑ 209,210,211

[3] P. Mazur. Broad-Coverage Rule-Based Processing of Temporal Expressions. Ph.D. dissertation, Macquarie University (Australia) and Wrocław University of Technology (Poland), 2012.. ↑ 209,210,211

[4] B. Han and A. Lavie. 2004. A framework for resolution of time in natural language. ACM Transactions on Asian Language Information Processing (TALIP), 3(1):11- 32, March.. ↑ 209,210

[5] I. Mani and G. Wilson. 2000. Robust temporal processing of news. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL), Hong Kong, October. Association for Computational Linguistics, pages 69-76.. ↑ 210,211

[6] M. Negri and L. Marseglia. 2005. Recognition and normalization of time expressions: ITC-IRST at TERN 2004. Technical Report WP3.7, Information Society Technologies, February.. ↑ 210,211

[7] B. Han, D. Gates, and L. Levin. 2006. From language to time: A temporal expression anchorer. In Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME), pages 196-203, Budapest, Hungary, June. Institute of Electrical and Electronics Engineers.. ↑ 210,211

[8] F. Schilder and Ch. Habel. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, ACL-2001. Toulouse, 2001, 65- 72.. ↑ 210,211

[9] E. Saquete, R. Muñoz, and P. Martínez-Barco. 2003. Terseo: Temporal expression resolution system applied to event ordering. In Text, Speech and Dialogue, pages 220-228.. ↑ 210,211

[10] P. Mazur and R. Dale. 2008. What's the Date? High Accuracy Interpretation of Weekday Names. In the Proceedings of the 22nd International Conference on Computational Linguistics (Coling). 16-24 August, Manchester, UK, pages 553-560.. ↑ 211

[11] W. Sun, A. Rumshisky, Ö. Uzuner. «Normalization of Relative and Incomplete Temporal Expressions in Clinical Narratives», Journal of the American Medical Informatics Association, 04/2015; 22(5).. ↑ 211

[12] R. Dale and P. Mazur. 2007. The Semantic Representation of Temporal Expressions in Text. In the Proceedings of the 20th Australian Joint Conference On Artificial Intelligence. Gold Coast, Queensland, Australia, 2-6 December 2007. SpringerVerlag Lecture Notes in Artificial Intelligence (LNAI) series.. ↑ 211

[13] A. X. Chang and Ch. D. Manning. 2012. SUTIME: A library for recognizing and normalizing time expressions. In 8th International Conference on Language Resources and Evaluation (LREC 2012).. ↑ 211

[14] J. Strötgen, M. Gertz. «HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions», Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, Uppsala, Sweden, 15-16 July 2010. Pp. 321- 324.. ↑ 211,212

[15] C. Smith. (**??**). Temporal structures in discourse. In R. P. Meier, H. Aristar-Dry and E. Destruel (Eds.), Text, time, and context (pp. 285-302). Dordrecht: Springer Netherlands. (Studies in Linguistics and Philosophy, 87). http://link.springer. com/10.1007/978-90-481-2617-0_12. ↑

[16] Сулейманова Е. А. Семантический анализ контекстных дат // Программные системы: теория и приложения. — 2015. — № 4(27). — С. 367—399. ↑

[17] Падучева Е. В. К семантической классификации временных детерминантов предложения. Язык: система и функционирование: сб. науч. трудов (под ред. Ю. Н. Караулова). М.: Наука, 1988. С. 190-201.. ↑ 213

[18] Кржижкова Е. Темпорально-квантитативная детерминация глагола: опыт трансформационного анализа // Ceskoslovenska rusistika. 1966. XI. С. 86-93.. ↑ 213

[19] Падучева Е. В. Семантические исследования (Семантика времени и вида в русском языке; Семантика нарратива). — М.: Школа «Языки русской культуры», 1996.. ↑ 213

[20] H. Kamp, J. van Genabith, U. Reyle. Discourse Representation Theory. In Dov Gabbay and Franz Guenthner (Eds.) Handbook of Philosophical Logic. Kluwer. 2005.. ↑ 213

[21] Булыгина Т. В. Языковая концептуализация мира (на материале русской грамматики) / А. Д. Шмелёв, Т. В. Булыгина. — М.: Школа «Языки русской культуры», 1997.— 577 с.. ↑ 217

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

Пример ссылки на эту публикацию:

Е. А. Сулейманова. «О двух видах текстовых временных координат», *Программные системы: теория и приложения*, 2016, **7**:4(31), с. 209–228.

URL: http://psta.psiras.ru/read/psta2016_4_209-228.pdf

Об авторе:

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

Elena Suleymanova. *On two types of time-referring expressions.*

Abstract. The paper suggests a view on categorizing text expressions that are generally referred to by information extraction community as time-point expressions, or temporal coordinates,. Two types of expressions are identified which differ in the way they refer to time. The issues of normalization (i. e. identifying the absolute value) are addressed for both types of expressions. (*In Russian*).

Key words and phrases: natural language processing, temporal information extraction, normalization of context-dependent time expressions.

Sample citation of this publication:

Elena Suleymanova. "On two types of time-referring expressions", *Program systems: Theory and applications*, 2016, **7**:4(31), pp. 209–228. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2016_4_209-228.pdf