

«Год назад»: о нормализации одного типа относительных временных выражений*

Е. А. Сулейманова, И. В. Трофимов

Институт программных систем им. А. К. Айламазяна РАН, г. Переславль-Залесский, Россия

Аннотация. Статья посвящена проблеме вычисления абсолютного значения временных выражений типа *два месяца назад*, *через полтора года* и т. п., объединяемых авторами в группу с условным названием «координаты-смещения». Подход, практикуемый в настоящее время в рамках задачи аннотирования темпоральной информации, предписывает считать нормализованным значением таких выражений календарную дату. Было установлено, что представляет собой значение таких выражений с точки зрения человека и насколько нормализация их в виде даты того или иного масштаба согласуется с человеческими оценками, полученными в результате опроса. На основе ответов были построены эталонные значения выражений рассматриваемого типа в виде функций принадлежности.

Ключевые слова: обработка естественного языка, извлечение темпоральной информации, темпоральные выражения, нормализация временных выражений, абсолютное значение относительных указаний на время, дейктические и анафорические темпоральные отсылки, TimeML.

DOI 10.14357/20718594200103

Введение

Неотъемлемая составляющая задачи извлечения из текста фактической информации — локализация во времени соответствующего факту события. Информация, позволяющая локализовать событие на временной оси, передается посредством разнообразных по форме и содержанию текстовых выражений — темпоральных, или временных, выражений (time expressions, temporal expressions).

Для того чтобы содержащаяся в тексте темпоральная информация была доступна различным аналитическим приложениям, она должна быть нормализована, т. е. приведена к стандартизованному машиночитаемому формату.

Практический интерес представляют, прежде всего, те временные выражения, которые служат для привязки событий к календарю. В самом типичном случае эту функцию выполняют

даты — выражения, описывающие (имеющие своим референтом) конкретную календарную единицу того или иного масштаба. Поскольку нормализация полной, или абсолютной, даты представляет собой техническую задачу, проблемы нормализации затрагиваются обычно в связи с относительными или неполными датами (*в прошлом году*, *в прошедшем месяце*, *в начале недели*). Под нормализацией относительной или неполной даты понимается не столько запись ее в требуемом формате, сколько установление абсолютного значения — идентификация соответствующей календарной единицы.

Статья посвящена проблеме нормализации относительных выражений типа *год назад*, *через две недели*, *три месяца назад* и т. п. Как и относительные даты, такие выражения служат для привязки событий к календарю, а для установления их значения тоже необходим контекст. Но на этом сходство заканчивается.

* Работа выполнена при финансовой поддержке РФФИ, грант № 19-07-00991.

✉ Сулейманова Елена Анатольевна. E-mail: yes2helen@gmail.com

Сравним выражения *(в) прошлом году* и *год назад*. Каким бы ни был контекст, для даты *прошлый год* результат вычислений всегда будет представлять собой ‘конкретный календарный год, предшествующий текущему году’. Эта часть значения даты *прошлый год* инвариантна относительно контекста. Поэтому, будучи сказано в любой момент от 1 января 2019 года до 31 декабря 2019 года, *в прошлом году* означает ‘в 2018 году’, а разница в дате привязки всего в один день меняет значение этого выражения на целый год: для 31 декабря года 2018 *в прошлом году* означает ‘в 2017 году’.

Попытка сформулировать в таких же терминах контекстно-независимое значение выражения *год назад* терпит фиаско. ‘Того же числа того же месяца, что и сегодня, но в прошлом году’? ‘В одноименном месяце прошлого года’? ‘В прошлом году’? Любая трактовка может оказаться верной, и все три могут оказаться ошибочными.

От относительных дат выражения типа *n лет назад* и подобные им отличает одно обстоятельство, имеющее принципиальное значение для нормализации [1]. Такие выражения не содержат в себе прямого указания на то, *что* представляет собой их значение (их собственное время). Имеется лишь указание на *величину смещения* этого времени относительно некоторой точки привязки.

Введем для таких выражений термин *координата-смещение*. TimeML Specification Language¹, международный стандарт аннотирования темпоральной информации, причисляет такие выражения к датам, и нормализовать их предписывает соответственно — в виде даты (подробнее об этом в следующем разделе).

Очевидно, что такое решение продиктовано соображениями простоты и единообразия разметки темпоральной информации в тексте. Мы задались целью экспериментальным путем оценить, насколько велики «информационные погрешности» при таком упрощенном подходе к нормализации выражений типа «координата-смещение».

Прежде чем перейти к описанию эксперимента и анализу его результатов, рассмотрим положения стандарта TimeML, задающие уровень state-of-the-art в нормализации выражений интересующего нас типа.

1. Нормализация временных выражений типа «координата-смещение»: положение дел

По материалам разных документов [2-4] подход международного стандарта темпоральной разметки TimeML к нормализации координат-смещений и схожих с ними выражений можно сформулировать в виде следующих положений.

В отдельную группу стандарт выделяет выражения, в которых явно указана дата привязки (*two weeks from next Tuesday, three years ago today*). Их предписывается нормализовать в виде даты масштаба даты привязки. Такие выражения мы не считаем координатами-смещениями. Это скорее даты, сравните в русском языке конструкцию *(в) этот день три года назад*. Как бы то ни было, нормализация таких выражений в виде даты масштаба точки привязки представляется вполне корректной.

Случаи, в которых явное указание на дату привязки отсутствует (*a year ago, in two weeks* и т. п.), следует нормализовать, руководствуясь правилом «масштаба выражения» (“Expression Granularity” Rule [2]). Это правило гласит: в отсутствие в контексте явной даты привязки, масштаб значения следует определять по существительному-вершине² (имеется в виду существительное, называющее масштаб смещения).

В соответствии с этим подходом, любую координату-смещение с дейктической отсылкой следует нормализовать в виде даты масштаба смещения, поскольку ее точка привязки — «момент речи», который не может быть явно упомянут в самом тексте. Следовательно, *неделю назад* = ‘на прошлой неделе’, *два месяца назад* = ‘в позапрошлом месяце’ и т. п. Аналогично, значением координат-смещений *год назад, два года назад, 50 лет назад, триста лет назад, тысячу лет назад* должен быть признан точный календарный год. Смещение, выраженное в десятилетиях и веках, правило разрешает пересчитывать в годах. Это означает, что выражение *два века назад* в тексте, датированном 2019 годом, можно нормализовать либо как век (значением будет «XIX век»), либо как точный календарный год («1819»). Правило «масштаба выражения» не распространяется на

¹ <http://www.timeml.org>

² “Expression Granularity” Rule: When no explicit anchor exists in the document context, use only the head noun to determine the precision of the VAL (цитируется по источнику [2, с. 23]).

выражения с очень большими величинами смещения: они нормализуются как «геологические эры», с использованием специальной нотации (*210 million years ago* — «MA210»).

Что касается анафорических координат-смещений, привязанных к датированному событию, то стандарт 2005 года [2] разрешал вычислять их значение (атрибут *value*), но не давал специальных указаний относительно выбора масштаба для нормализации. Как бы то ни было, вариантов для выбора два: нормализовать либо в виде даты «масштаба выражения» (т. е. масштаба смещения), либо в виде даты, имеющей тот же масштаб, что и дата-антецедент. Таким образом, в примере *Первый полет 737 совершил в декабре 1966 года, а уже год спустя Lufthansa получила свой первый борт* значением выражения *год спустя* можно считать либо «1967 год», либо «декабрь 1967 года».

Стандарт 2006 года [3] в таких случаях ограничивается фиксацией в разметке темпорального отношения между двумя событиями: они связываются аннотацией TLink с атрибутом *magnitude*; значением этого атрибута в конечном итоге считается нормализованная запись смещения, которое помечается отдельной аннотацией как длительность (*Duration*). Абсолютное значение анафорического темпорального выражения в терминах календарного времени не вычисляется, т. е. в рассматриваемом нами смысле такие выражения не нормализуются. В более поздних версиях TimeML [4] эта тема не затрагивается.

Суть нашего экспериментального исследования состояла в следующем:

1. С помощью опроса группы людей получить эталонные значения координат-смещений для тестовых примеров.

2. Реализовать подход state-of-the-art к нормализации координат-смещений, взяв за основу рекомендации стандарта TimeML, и применить его к тестовым примерам. Это означает: а) применить нормализацию в масштабе смещения как к дейктическим, так и к анафорическим координатам-смещениям — для полноты картины; б) к выражениям со смещением в веках и десятилетиях дополнительно применить нормализацию с уменьшением масштаба до «года»; в) для анафорических выражений выполнить еще и нормализацию в масштабе даты привязки — если ее масштаб отличается от масштаба смещения.

3. Оценить, в какой мере полученные результаты нормализации соответствуют человеческим оценкам (эталонным значениям).

2. Эталонные значения выражений типа «координата-смещение»

2.1. Получение данных для построения эталонов

Семнадцать респондентам были предложены фрагменты новостных текстов, в которых были помечены 25 выражений типа «координата-смещение», локализующих произвольные события. Ни одно из выражений не содержало явных маркеров неточности (такие выражения, как *около месяца назад*, *более ста лет назад*, *года через три*, из рассмотрения исключались). Опрашиваемым предлагалось в каждом случае описать в свободной форме *минимально возможный* (т. е. максимально точный) отрезок календарной шкалы, в который, по их мнению, с большой вероятностью попадает время события.

Отрезок календарной шкалы, в который с большой вероятностью попадает время события, принимался нами за оценку респондентом значения координаты-смещения — на том основании, что именно так соотносятся показатели времени и время самого события в типичном случае темпоральной локализации. В таких случаях, когда временное выражение используется для локализации события во времени и является ответом на вопрос «когда произошло/произойдет событие?», имеет место следующее соотношение: собственное время темпорального выражения включает в себя (ограничивает) время локализуемого события³.

Примеров, в которых это было бы не так, среди предложенных для оценки не встречалось. Не было ни одного случая, когда время координаты-смещения оказывалось бы внутри неопределенно длящегося времени ситуации⁴. Кроме того, не было и случаев, в которых

³ В таком употреблении временное выражение представляет собой *показатель объемлющего времени* (Падучева Е. В. Семантические исследования. Семантика времени и вида в русском языке; Семантика нарратива. — М.: Школа «Языки русской культуры», 1996). Ср. с *показателем включенного времени* (там же).

⁴ Как в примере *Три года назад он уже был губернатором* (время губернаторства не ограничивается временем координаты-смещения, которая здесь представляет собой показатель включенного времени).

Табл. 1. Полученные от 17 респондентов оценки значения выражения *год назад* во фрагменте: «Новое здание на месте старой конструкции XIX столетия открылось лишь год назад». Дата публикации документа 10 марта 2019 года

1	Январь-май 2018	7	Первая половина 2018	13	8 февраля — 9 апреля 2018
2	Зима 2017-2018	8	Начало 2018 — весна 2018	14	Февраль-март 2018
3	Начало 2018	9	Февраль-апрель 2018	15	Февраль-апрель 2018
4	Март 2018	10	Первый квартал 2018	16	Февраль-апрель 2018
5	2018	11	Начало 2018	17	Февраль-март 2018
6	Начало 2018	12	Середина марта 2018		

Табл. 2. Индивидуальные оценки для выражения *два года назад* с датой привязки 18.03.2019

u1	2017	u7	2017	u13	2017-START
u2	2017-H1	u8	2017-H1	u14	2017-H1
u3	2017	u9	2017-START	u15	2017
u4	2017-03	u10	2017-START	u16	2017-H1
u5	2017	u11	2017	u17	2017-02:2017-03
u6	2017	u12	2017-03		

опрашиваемый в силу своих экстралингвистических знаний мог бы оценить время события интервалом, значительно более узким, чем оценил бы время самого темпорального выражения вне контекста⁵.

В трех случаях из 25 отмечено точное попадание всеми участниками опроса в одну и ту же дату. В остальных случаях респонденты в той или иной степени расходились в своих оценках (Табл. 1). Полученные в результате опроса данные были использованы для построения тестового множества *testset-25*.

По такой же методике были собраны данные для второго тестового множества (*testset-100*). В этом случае семи⁶ участникам опроса было предложено 100 смоделированных примеров, в которых для указания на время гипотетического события использовались выражения с предлогом *назад* — *год назад*, *два месяца назад*, *три недели назад*, *тридцать пять лет назад* и т. д. Все эти выражения — координаты-смещения дейктического типа (точка привязки — «момент речи»; «момент речи» в каждом случае задавался явно датой масштаба «день»). В примерах были задействованы разнообразные величины (числа) и единицы смещения. Примеры со смещением в «днях» во второе множество решено было не включать, поскольку

ку с точки зрения нормализации эти случаи особого интереса не представляют.

Ни в одном из 100 случаев не отмечено полного совпадения мнений всех опрошенных.

2.2. Представление эталонного значения координаты-смещения

Эталонные значения координат-смещений для тестовых данных представляются как нечеткие множества. Нечеткое множество задается функцией принадлежности, ставящей в соответствие каждому элементу нечеткого множества число от нуля до единицы — степень его принадлежности множеству. Элементами множества выступают календарные единицы.

На вход процедуре построения эталонного значения для некоторой координаты-смещения поступает множество оценок времени этой координаты, данных разными людьми в ходе опроса по описанной ранее методике.

Предварительно индивидуальные оценки приводятся к специальному формату записи (Табл. 2). За основу формата записи дат взят стандарт TimeML, который был переработан под нужды нашей календарной модели с ее расширенной трактовкой дат: мы относим к датам нечеткие фазы (начало, середина, конец) и условно точные части (половина, четверть) календарных единиц. Нечеткие фазы представлены в модели функциями принадлежности, построенными по результатам соответствующих опросов. Кроме того, мы ввели специальный формат для произвольных календарных интервалов, чтобы иметь возможность записывать интервальные оценки.

⁵ Как, например, *Три года назад он пошел в первый класс* (время события могло бы быть уточнено до «три года назад осенью», что, безусловно, не может считаться значением самой координаты-смещения).

⁶ Небольшое число участников объясняется трудоемкостью как участия в опросе, так и обработки его результатов.

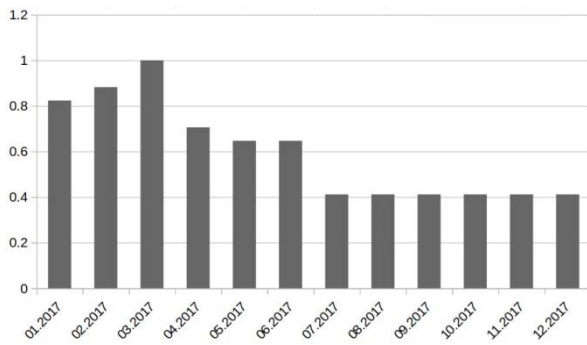


Рис. 1. Эталонное значение выражения два года назад с точкой привязки 18.03.2019

Затем каждая из индивидуальных оценок интерпретируется в календаре. В календаре используются 5 масштабов: «день», «месяц», «год», «век», «тысячелетие». Заметим, что нечеткие фазы календарных единиц для текущей задачи предварительно приводятся к четким интервалам: в качестве таковых используются альфа-сечения с альфа равным 0.5.

Таким образом, каждая индивидуальная оценка представляет собой отрезок календарной шкалы, состоящий из календарных единиц некоторого масштаба. За масштаб эталона принимается масштаб наименьшей из таких единиц.

Все индивидуальные отрезки приводятся к масштабу эталона. Отрезок, образованный объединением индивидуальных отрезков (а точнее, множество составляющих его календарных единиц масштаба эталона), принимается за носитель нечеткого множества — будущего эталона.

Затем строится функция принадлежности эталонному значению (Рис. 1). Значение функции для каждого элемента носителя вычисляется следующим образом: каждое вхождение этого элемента в один из индивидуальных отрезков засчитывается за единицу; полученная сумма делится на общее число оценок (т. е. число опрошенных). Таким образом, степень принадлежности календарной единицы эталонному значению отражает степень уверенности, с которой эта единица была отнесена к эталону группой респондентов.

3. Метод оценки соответствия результата нормализации эталонному значению

Мера соответствия результата нормализации эталонному значению должна учитывать как

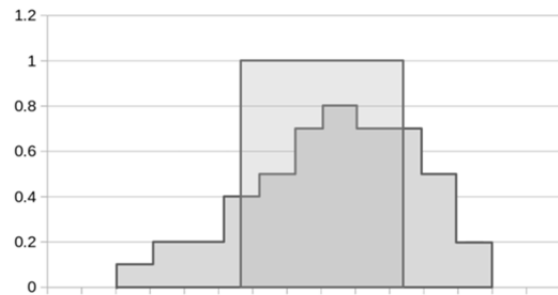


Рис. 2. Фигуры, соответствующие эталону и результату нормализации

точность попадания в эталон, так и полноту его покрытия. Эталон, описываемый функцией принадлежности, двумерен. Для оценки соответствия результата эталону важны оба измерения: по оси абсцисс располагаются календарные единицы, а ординаты определяют значимость совпадения или несовпадения результата с эталоном в каждой из этих единиц⁷.

Результат нормализации в виде даты представляет собой четкий отрезок календарной шкалы. Для сопоставления с нечетким эталоном результат нормализации также можно представить в виде функции принадлежности. Эта функция имеет вид $f(x) = 1$ для всех входящих в него календарных единиц x , независимо от их масштаба.

В качестве меры сравнения результата и эталона будем использовать степень совпадения фигур (криволинейной трапеции — для эталона и прямоугольника — для результата нормализации), образованных графиками функций, осью абсцисс и прямыми, проходящими через начало и конец графиков (Рис. 2). Для сопоставления эталон и результат приводятся к масштабу «день».

За точность принимается отношение площади пересечения к площади прямоугольника. Полнота — отношение площади пересечения к площади эталонной фигуры. F-мера определяется формулой:

$$F = 2 \times \frac{\text{точность} \times \text{полнота}}{\text{точность} + \text{полнота}}. \quad (1)$$

⁷ Ордината выражает степень принадлежности календарной единицы эталону. При сравнении результата с эталоном это имеет следующий смысл: чем больше степень принадлежности календарной единицы эталону, тем выше ценится включение ее в результат и тем выше цена ошибки в противном случае.

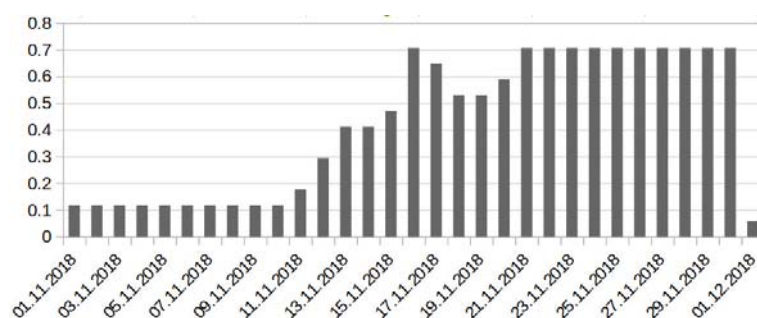


Рис. 3. Пример эталона

Чем больше F-мера, тем лучше результат нормализации соответствует эталону. Мы используем сбалансированную F-меру, которая признает равную значимость точности и полноты (ее принято обозначать $F1$).

Заметим, что в принципе возможен и другой путь сравнения четкого результата с нечетким эталоном — через преобразование последнего в четкий отрезок. Но, каков бы ни был способ такого преобразования, суть его — огрубление, а, следовательно, потеря части данных эталона. В наших модельных данных графики эталонных функций имеют ступенчатую форму, далекую от колоколообразной и даже не всегда выпуклую (Рис. 3). Какой способ преобразования выбрать, а также насколько дефазифицированный,

таким образом, эталон будет пригоден для оценки качества нормализации — все эти вопросы требуют отдельного исследования. Выбранная нами мера соответствия более проста и вполне отвечает задаче.

4. Модели индивидуальной оценки

Для обсуждения результатов эксперимента нам потребуется понятие *модели (индивидуальной) оценки*.

Анализ материалов, полученных в результате опросов, позволил выделить несколько моделей, которым следовали респонденты при оценке значения координат-смещений (Табл. 3).

Табл. 3. Модели индивидуальной оценки значения координаты-смещения

Модель оценки (условное название)	Значение координаты-смещения, вычисленное в соответствии с моделью	Пример(ы) значения выражения <i>пятьдесят лет назад</i> , вычисленного в соответствии с моделью (дата привязки «1 ноября 2019 года»)
Точечная оценка в масштабе смещения	Календарная единица масштаба смещения	«1969 год»
Интервальная оценка в масштабе смещения	Произвольный отрезок, состоящий из двух и более календарных единиц масштаба сме- щения	«1967–1971 годы»
Точечная прицельная оценка	Календарная единица, масштаб которой меньше масштаба смещения. Выбор масшта- ба ограничен «снизу» масштабом даты при- вязки. При точечной прицельной оценке зна- чение всегда располагается внутри календарной единицы масштаба смещения	«1 ноября 1969 года» «ноябрь 1969 года» «вторая половина 1969 года»
Интервальная прицельная оценка	Отрезок, состоящий из двух и более календар- ных единиц, имеющих масштаб меньше масштаба смещения. В отличие от точечной, интервальная прицельная оценка может пере- секать границу между двумя смежными кален- дарными единицами масштаба смещения	«октябрь–ноябрь 2014 года» «вторая половина 1969 года — первая половина 1970 года»

5. Реализация подхода state-of-the-art к нормализации выражений типа «координата-смещение»

Для того чтобы воспроизвести в экспериментах рекомендуемый TimeML подход к нормализации координат-смещений, мы реализовали четыре способа (режима) точной нормализации. Технические подробности реализации мы опускаем, поскольку они потребовали бы детального описания календарной модели.

1. *Точная нормализация в масштабе смещения* — нормализация в виде даты масштаба смещения, независимо от масштаба даты привязки.

2. *Точная нормализация для нецелых величин смещения* — смещение пересчитывается в единицах меньшего масштаба (подробнее об этом — в Разделе 6).

3. *Точная нормализация с уменьшением масштаба смещения до года* — альтернативный вариант нормализации для смещений в десятилетиях и веках, допускаемый TimeML.

4. *Точная нормализация в масштабе даты привязки* (отличается от первого режима только в тех случаях, когда масштаб явно упомянутой даты привязки *меньше* масштаба смещения). Результат нормализации имеет тот же масштаб, что и дата привязки.

Заметим, что четыре точных режима нормализации покрывают лишь точечные модели оценки значений. Эксперимент состоял из следующих этапов (первые три режима применялись к примерам из обоих тестовых множеств):

- в исходном масштабе смещения;
- для нецелых величин смещения;
- с уменьшением масштаба смещения до года;
- в масштабе даты привязки к анафорическим случаям из тестового множества testset-25.

Для первых трех режимов детально анализировать показатели сопоставления результатов нормализации с эталонами удобнее на материале тестового множества testset-100, так как в него включены примеры с разнообразной величиной и всеми возможными масштабами смещения, за исключением смещения в «днях». Это множество состоит из бесконтекстных примеров выражений типа «координата-смещение» с отсылкой к «моменту речи», который задавался явным образом.

Результаты, полученные на тестовом множестве testset-25, рассмотрим отдельно, акцентируя внимание на различиях между дейктическими и анафорическими случаями. Безусловно, число

опрошенных недостаточно велико, чтобы гарантировать достоверность и сбалансированность эталонных значений. Поэтому, оценивая результаты эксперимента, можно говорить лишь о тенденциях. Но тенденции прослеживаются явно даже на таком скромном материале.

6. Обсуждение результатов эксперимента

6.1. Точная нормализация в исходном масштабе смещения

Результаты применения режима точной нормализации в масштабе смещения (далее для краткости *МС-нормализации*) к тестовому множеству testset-100 свидетельствуют о следующем.

1. *Лучшие показатели соответствия эталонам МС-нормализации продемонстрировала на выражениях со средней величиной смещения в годах и месяцах.* На таких выражениях F1-мера превысила значение 0.8. Заметим, что порог 0.8 F-мера преодолела всего на восьми примерах из 100 (Табл. 4, раздел 1). Эталоны для этих выражений почти совпадают с календарной единицей масштаба смещения (Рис. 4, а), за исключением случаев пограничной точки привязки. Об этом далее.

2. *С приближением числовой величины смещения к единице снижается точность совпадения результата МС-нормализации с эталоном* (Табл. 4, раздел 2). С уменьшением величины смещения растет число прицельных оценок среди ответов, а в примерах с единичным смещением в годах и месяцах оценки в масштабе смещения отсутствуют вообще (Рис. 5, а).

3. *Близость точки привязки к границам календарной единицы масштаба смещения снижает как точность, так и полноту совпадения результата МС-нормализации с эталоном.* Пограничное расположение точки привязки склоняет респондентов к интервальной прицельной оценке, захватывающей фрагменты смежных календарных единиц масштаба смещения. Это заметнее всего на единичных и малых величинах смещения в годах и месяцах (Рис. 5, б), однако эффект от пограничного расположения точки привязки можно наблюдать и на выражениях со средними величинами смещения в годах и месяцах (Рис. 4, б). «Трансграничные» прицельные оценки значительно снижают как полноту, так и точность совпадения результатов МС-нормализации с эталоном (Табл. 4, раздел 3, а и б).

Табл. 4. Сводная таблица примеров для точной нормализации в масштабе смещения

Выражение	Дата привязки («момент речи»)	Результат МС-нормализации			
		значение	точность	полнота	F1
1. Лучшие результаты, полученные применением МС-нормализации на тестовом множестве testset-100					
Пять месяцев назад	2019-05-12	2018-12	0.88	1.00	0.94
Десять лет назад	2019-05-12	2009	0.87	1.00	0.93
Пять лет назад	2019-05-12	2014	0.87	1.00	0.93
Четыре года назад	2019-05-12	2015	0.87	1.00	0.93
Шесть лет назад	2019-05-12	2013	0.87	1.00	0.93
Четыре месяца назад	2019-05-12	2019-01	0.80	1.00	0.89
Две недели назад	2019-01-02	2018-W51	0.76	0.97	0.85
Шесть месяцев назад	2019-05-12	2018-11	0.84	0.85	0.85
2. Низкая точность совпадения результатов МС-нормализации с эталоном при малой величине смещения					
Три года назад	2019-05-12	2016	0.51	1.00	0.68
Три месяца назад	2019-05-12	2019-02	0.49	1.00	0.66
Два месяца назад	2019-05-12	2019-03	0.48	1.00	0.65
Два года назад	2019-05-12	2017	0.45	1.00	0.62
Месяц назад	2019-05-12	2019-04	0.37	1.00	0.54
Год назад	2019-05-12	2018	0.16	1.00	0.27
3а. Влияние положения точки привязки на точность и полноту (малая величина смещения)					
Три месяца назад	2019-05-12	2019-02	0.49	1.00	0.66
Три месяца назад	2019-01-02	2018-10	0.30	0.61	0.40
Два месяца назад	2019-05-12	2019-03	0.48	1.00	0.65
Два месяца назад	2019-01-02	2018-11	0.27	0.51	0.35
Месяц назад	2019-05-12	2019-04	0.37	1.00	0.54
Месяц назад	2019-01-02	2018-12	0.18	0.57	0.27
Три года назад	2019-05-12	2016	0.51	1.00	0.68
Три года назад	2019-01-02	2016	0.27	0.55	0.36
Два года назад	2019-05-12	2017	0.45	1.00	0.62
Два года назад	2019-01-02	2017	0.21	0.63	0.32
Год назад	2019-05-12	2018	0.16	1.00	0.27
Год назад	2019-01-02	2018	0.09	0.53	0.15
3б. Влияние положения точки привязки на точность и полноту (средняя величина смещения)					
Пять месяцев назад	2019-05-12	2018-12	0.88	1.00	0.94
Пять месяцев назад	2019-01-02	2018-08	0.41	0.52	0.45
Четыре месяца назад	2019-05-12	2019-01	0.80	1.00	0.89
Четыре месяца назад	2019-01-02	2018-09	0.40	0.61	0.48
Десять лет назад	2019-05-12	2009	0.87	1.00	0.93
Десять лет назад	2019-01-02	2009	0.73	0.71	0.72
Шесть лет назад	2019-05-12	2013	0.87	1.00	0.93
Шесть лет назад	2019-01-02	2013	0.67	0.70	0.68
4. Низкая точность результатов МС-нормализации на примерах с большим масштабом смещения					
Пять веков назад	2019-05-12	15	0.46	0.93	0.62
Три века назад	2019-05-12	17	0.46	0.93	0.62
Два века назад	2019-05-12	18	0.33	0.91	0.48
Столетие назад	2019-05-12	19	0.15	0.83	0.25
тысячелетие назад	2019-05-12	1	0.08	0.72	0.14
5. Низкая полнота результатов МС-нормализации на примерах с круглой величиной смещения					
Тридцать лет назад	2019-05-12	1989	0.93	0.42	0.58
Тридцать пять лет назад	2019-05-12	1984	1.00	0.54	0.70
Пятьдесят лет назад	2019-01-02	1969	0.93	0.16	0.28
125 лет назад	2019-03-14	1894	1.00	0.53	0.69

Выражение	Дата привязки («момент речи»)	Результат МС-нормализации			
		значение	точность	полнота	F1
Сто лет назад	2019-05-12	1919	1.00	0.14	0.24
Двести пятьдесят лет назад	2019-05-12	1769	1.00	0.04	0.08
Тысячу лет назад	2019-05-12	1019	1.00	0.01	0.02
6. Показатели для примеров со смещением в «неделях»					
Неделю назад	2019-01-02 (среда)	2018-W52	0.37	1.00	0.54
Неделю назад	2019-05-12 (воскр.)	2019-W18	0.31	0.88	0.45
Две недели назад	2019-01-02 (среда)	2018-W51	0.76	0.97	0.85
Две недели назад	2019-05-12 (воскр.)	2019-W17	0.59	0.71	0.64
Три недели назад	2019-01-02 (среда)	2018-W50	0.67	0.78	0.72
Три недели назад	2019-05-12 (воскр.)	2019-W16	0.63	0.67	0.65
Четыре недели назад	2019-01-02 (среда)	2018-W49	0.71	0.70	0.71
Четыре недели назад	2019-05-12 (воскр.)	2019-W15	0.78	0.76	0.78

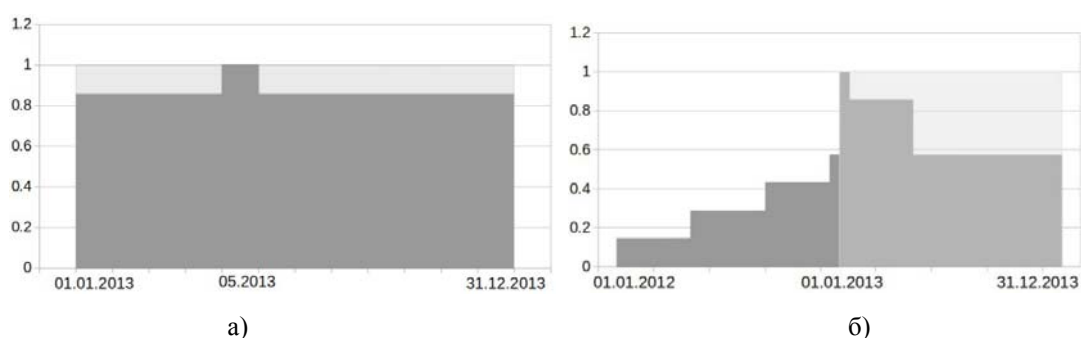


Рис. 4. Выражение *шесть лет назад*: эталон (темно-серая фигура) и результат его МС-нормализации «2013 год» (прямоугольник)

а) – дата привязки 2019-05-12, б) – дата привязки 2019-01-02

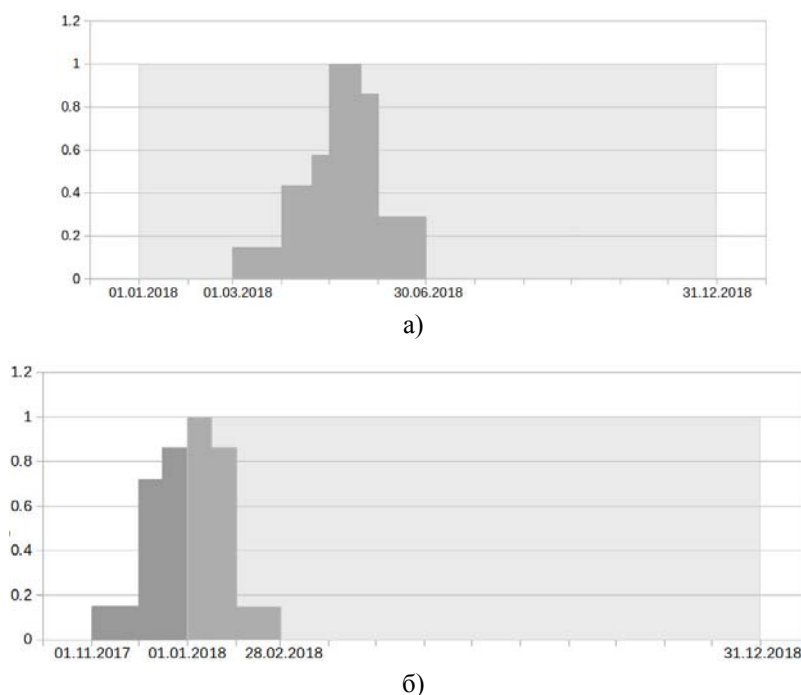


Рис. 5. Взаимное расположение эталона и результата МС-нормализации «2018 год» для выражения *год назад*

а)– дата привязки 2019-05-12, б) – дата привязки 2019-01-02

4. На выражениях с большим масштабом смещения результаты МС-нормализации имеют очень низкую точность совпадения с эталоном (Табл. 4, раздел 4). Среди ответов отсутствуют оценки в масштабе смещения, все оценки прицельны.

5. На выражениях с большой круглой величиной смещения в годах отмечается снижение полноты совпадения результата МС-нормализации с эталоном. В индивидуальных оценках для выражений с большой круглой величиной смещения в годах (число лет кратно «10» или «5») преобладают интервальные оценки в масштабе смещения или даже в большем масштабе (для сотен и тысяч лет), в связи с чем результаты МС-нормализации таких выражений имеют низкие показатели полноты при максимальной или высокой точности (Табл. 4, раздел 5). Недобор точности совпадения для некоторых примеров и здесь объясняется присутствием в ответах прицельных оценок.

6. Выражения со смещением в неделях имеют некоторую специфику. Для выражения *неделю назад* показатели аналогичны примерам *месяц назад* и *год назад* с серединным и пограничным положением точки привязки (Табл. 4, раздел 6). Все респонденты дали прицельные интервальные оценки (отрезки длиной в 2–3 дня), что объясняет низкую точность результата МС-нормализации в виде недели.

Для выражения *две недели назад* тип оценок тот же, но длина интервала увеличивается до шести–семи дней. Отсюда почти полное совпадение эталона по длине с календарной неделей для примера с серединной точкой привязки («среда»). И в остальных случаях среди ответов прицельных интервальных оценок в днях значительно больше, чем точечных оценок в масштабе недели.

6.2. Точная нормализация для нецелых величин смещения

Поскольку в документации по TimeML указаний относительно нормализации таких случаев найти не удалось, было решено использовать следующую схему изменения масштаба смещения: «тысячелетие», «век», «десятилетие» >> «год»; «год» >> «месяц»; «месяц», «неделя» >> «день» (месяц принимается равным 30 дням).

Числовая величина смещения пересчитывается соответственно. Затем выполняется нормализация в полученном масштабе. Результаты такой нормализации имеют низкую полноту (Табл. 5): во всех случаях в эталонах преобладают интервальные оценки, которые охватывают отрезки существенно большей протяженности, чем результат нормализации.

6.3. Точная нормализация с уменьшением масштаба смещения до года

Этот режим был применен к выражениям со смещением в веках (столетиях) и десятилетиях. Результаты нормализации в виде года имеют крайне низкую полноту (Табл. 6): во всех случаях эталоны охватывают отрезки существенно большей протяженности, чем один календарный год.

6.4. Сопоставление результатов точной нормализации с эталонными значениями на тестовом множестве testset-25

К тестовому множеству testset-25 были применены все 4 режима нормализации. Для действительных случаев сравнение с эталоном результатов нормализации в трех первых режимах подтверждает закономерности, которые наблюдались на данных testset-100⁸ (Табл. 7).

Табл. 5. Результаты точной нормализации для примеров с нецелой величиной смещения

Выражение	Дата привязки («момент речи»)	Результат нормализации с пересчетом смещения			
		значение	точность	полнота	F1
Полторы недели назад	2019-05-12	2019-05-01	1.00	0.33	0.50
Четверть века назад	2019-05-12	1994	0.93	0.22	0.36
Полвека назад	2019-05-12	1969	1.00	0.11	0.21
Полтора месяца назад	2019-05-12	2019-03-28	1.00	0.06	0.11
Полтора столетия назад	2019-05-12	1869	0.86	0.04	0.08
Два с половиной века назад	2019-05-12	1769	0.71	0.02	0.04

⁸ К одному из 25 примеров (*спустя 50 лет*, дата привязки в анафорическом контексте — «начало 1960-х») не применился ни один из штатных режимов нормализации.

Табл. 6. Результаты нормализации с уменьшением масштаба смещения до года

Выражение	Дата привязки («момент речи»)	Результат нормализации с уменьшением масштаба смещения до года			
		значение	точность	полнота	F1
Три десятилетия назад	2019-05-12	1989	0.93	0.14	0.25
Столетие назад	2019-05-12	1919	0.86	0.05	0.09
Два века назад	2019-05-12	1819	0.88	0.02	0.05
Три века назад	2019-05-12	1719	1.00	0.02	0.04
Пять веков назад	2019-05-12	1519	1.00	0.02	0.04
Три столетия назад	2019-05-12	1719	1.00	0.02	0.04

Табл. 7. Показатели для результатов трех режимов нормализации на тестовом множестве testset-25

Выражение	Дата привязки		Результат точной нормализации в масштабе смещения			
	«момент речи»	дата события	значение	точн.	полнота	F1
Четыре дня назад	2019-03-07		2019-03-03	1.00	1.00	1.00
За пять лет до выступления VW		2018	2013	1.00	1.00	1.00
Годом позже		1984	1985	1.00	1.00	1.00
11 лет назад	2015-09-01		2004	0.96	1.00	0.98
25 лет назад	2019-03-06		1994	0.94	1.00	0.97
Через пять лет	2019-03-06		2024	0.84	1.00	0.91
Спустя двое суток	2019-03-06		2019-03-08	1.00	0.81	0.89
Шесть лет назад	2019-03-18		2013	0.82	0.95	0.88
Через два года	2019-03-07		2021	0.79	1.00	0.88
За двадцать дней до полета		1961-04-12	1961-03-23	0.94	0.73	0.82
Полгода назад	2019-03-14		2018-09	0.94	0.65	0.77
Два года назад	2019-03-18		2017	0.60	1.00	0.75
125 лет назад	2019-03-14		1894	1.00	0.53	0.69
За три недели до выборов		2019-03-31	2019-W10	0.65	0.68	0.66
Через месяц после приземления		1961-04-12	1961-05	0.39	1.00	0.57
120 лет назад	2019-03-12		1899	0.94	0.37	0.53
Четверть века назад	2019-03-06		1994	0.88	0.31	0.46
Год назад	2019-03-10		2018	0.29	0.98	0.45
Двадцать лет спустя		1964	1984	1.00	0.22	0.36
Год спустя	1997-12	1997-12	1998	0.21	0.84	0.33
Год спустя	1966-12	1966-12	1967	0.17	0.76	0.27
Ровно через 13 лет после ареста		2004-12-19	2017	0.06	1.00	0.11
За 2,5 месяца до этого		2019-02-06	2018-11-23	0.71	0.05	0.10
Полтора столетия назад	2019-03-12		1869	0.76	0.04	0.07

Специфика эталонных значений для *анафорических координат-смещений* состоит в следующем: при наличии в анафорическом контексте явно упомянутой даты привязки выбор масштаба результата оказывается, ограничен «снизу» масштабом этой даты⁹.

Этим обстоятельством объясняется полное совпадение с эталоном результата нормализа-

ции в виде года для выражений *годом позже* и *за пять лет до выступления VW* (Табл. 8). Оба выражения привязаны к событиям, которые датированы календарным годом, что исключило для респондентов возможность прицельной оценки в меньшем масштабе. Сравнение: для выражения *год назад* с привязкой к «моменту речи» результаты нормализации в виде года имели полноту менее 0.2.

⁹ Это следует из соображений здравого смысла.

Табл. 8. Результаты применения нормализации в масштабе даты привязки

Выражение	Дата привязки (дата события)	Результат точной нормализации в масштабе смещения			
		значение	точность	полнота	F1
Год спустя	1966-12	1967-12	1.00	0.38	0.55
Год спустя	1997-12	1998-12	1.00	0.35	0.51
За три недели до выборов	2019-03-31	2019-03-10	0.76	0.12	0.20
Через месяц после приземления	1961-04-12	1961-05-12	0.82	0.07	0.12
За 2,5 месяца до этого	2019-02-06	2018-11-23	0.71	0.05	0.10
Ровно через 13 лет после ареста	2004-12-19	2017-12-19	1.00	0.05	0.09

Для выражений со смещением в днях действительные и анафорические употребления с точки зрения нормализации не различаются, поскольку дата привязки всегда представляет собой календарный день¹⁰. Можно предположить, что координаты-смещения с некруглой величиной в днях вполне успешно нормализуются в масштабе смещения (в точный день). Для смещения с числовой величиной «1» и «2» эталон может не совпадать с точным днем в связи с неоднозначной интерпретацией таких выражений¹¹. При величине смещения, кратной «10», результат нормализации в виде дня может недобирать полноты по сравнению с эталоном из-за наличия в ответах интервальных оценок — как в примере *за двадцать дней до полета*¹².

6.5. Точная нормализация в масштабе даты привязки

Этот режим нормализации мы оценивали на выражениях с анафорической отсылкой ко времени события, явно датированного в тексте, если масштаб его даты отличается от (а именно меньше) масштаба смещения. Дата события в таких случаях и принималась за дату отсчета. Точная нормализация в масштабе даты привязки была применена к шести примерам из множества *testset-25* (Табл. 8). Во всех случаях *результаты точной нормализации в масштабе даты привязки имеют низкую полноту относительно эталонных значений*.

¹⁰ Речь идет о календарной модели, в которой минимальной единицей является «день/сутки».

¹¹ В устном опросе респонденты испытывали затруднения при установлении значения выражений *через день*, *через два дня*, *день назад*, *два дня назад*.

¹² То, что в данном конкретном примере результат нормализации «23 марта 1961 года» имеет точность меньше единицы, объясняется, по всей видимости, вычислительной ошибкой в одном из ответов.

Заключение

Эксперименты на небольшом материале показали, что результаты *точной* нормализации временных выражений типа «координата-смещение» как в масштабе смещения, так и с уменьшением масштаба, в большинстве случаев значительно расходятся с оценкой значения этих выражений людьми.

Очевидно, нормализация выражений типа «координата-смещение» требует более дифференцированного подхода. При выборе способа нормализации необходимо учитывать разные аспекты: масштаб смещения, соотношение масштабов точки привязки и смещения, числовую величину смещения и ее дополнительные признаки, расположение точки привязки на календарной шкале. Кроме того, тот факт, что эталонное значение большинства координат-смещений в тестовых данных представляет собой нечеткое множество, говорит о том, что игнорировать фактор неопределенности при нормализации таких выражений было бы, по меньшей мере, странно. В связи с этим отметим недавнюю публикацию [4], в которой представлен опыт нечеткого моделирования «длительностей», маркированных явными показателями неточности или неопределенности: *approximately 10 days* «примерно 10 дней», *less than a month* «менее месяца».

Как минимум, в качестве альтернативы *точной нормализации координаты-смещения в виде даты того или иного масштаба* (в соответствии с предписаниями стандарта TimeML) можно рассмотреть методы, допускающие нормализацию в виде произвольного отрезка календарной шкалы, т. е. реализующие не только точечную, но и интервальную оценку.

Литература

1. Сулейманова Е. А. О двух видах текстовых временных координат // Программные системы: теория и приложения. 2016. № 4(31). С. 209–229.
2. Ferro L., Gerber L., Mani I., Sundheim B., Wilson G. TIDES 2005 standard for the annotation of temporal expressions. Technical report, MITRE. September.2005. <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-timex2-guidelines-v0.1.pdf>
3. Saur'ı Roser, Littman Jessica, Knippen Bob, Gaizauskas Robert, Setzer Andrea, Pustejovsky James. TimeML Annotation Guidelines. Version 1.2.1. January 31. 2006. https://catalog ldc.upenn.edu/docs/LDC2006T08/timeml_annotationguide_1.2.1.pdf
4. Guidelines for Temporal Expression Annotation for English for TempEval 2010. TimeML Working Group August 14. 2009. <http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf>
5. Tissot Hegler, Didonet Del Fabro Marcos, Derczynski Leon, Roberts Angus. Normalisation of imprecise temporal expressions extracted, Knowledge and Information Systems. 2019. Volume 61. Issue 3. Pp. 1361–1394. <https://doi.org/10.1007/s10115-019-01338-1>.

“A Year Ago”: on Normalisation of one Type of Temporal Expressions

E. A. Suleymanova, I. V. Trofimov

Program Systems Institute of RAS, Pereslavl-Zalessky, Russia

Abstract. The paper addresses a particular issue within the framework of temporal information normalisation. It focuses on time expressions of the kind ‘three years ago’, ‘two months later’, ‘a week ago’ etc. According to TimeML, the international standard for time markup and annotation, such expressions are to be normalised as dates. We conducted an experiment in order to find out how people interpret time expressions of this kind, and to what extent normalising such expressions as dates is consistent with human judgments. We used two questionnaires to collect human judgments, and then used the answers to create ‘model values’ in the form of membership functions. We implemented four modes of normalisation to reproduce the state-of-the-art method. Evaluation of these normalisation modes on our test data showed, in most cases, great discrepancies between the normalisation results and the ‘model values’. This might be an indication that alternative methods of normalisation would be appropriate. Experiment analysis revealed some normalisation patterns that could serve as a basis for an alternative solution.

Keywords: natural language processing, temporal information extraction, temporal expressions, timex normalisation, absolute value of relative temporal expressions, deictic and anaphoric temporal references, TimeML.

DOI 10.14357/20718594200103

References

1. Suleymanova E. A. 2016. O dvukh vidakh tekstovykh vremennykh koordinat [On two types of time-referring expressions]. Programmye sistemy: teoriya i prilozheniya [Program systems: theory and applications]. 4(31): 209–229. Available at: http://psta.psiras.ru/read/psta2016_4_209-229.pdf
2. 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>
3. Roser Saur'ı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. TimeML Annotation Guidelines. Version 1.2.1. January 31, 2006 https://catalog ldc.upenn.edu/docs/LDC2006T08/timeml_annotationguide_1.2.1.pdf
4. Guidelines for Temporal Expression Annotation for English for TempEval 2010. TimeML Working Group August 14, 2009 <http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf>
5. Tissot, H., Del Fabro, M.D., Derczynski, L. et al. Knowl Inf Syst (2019) 61: 1361. <https://doi.org/10.1007/s10115-019-01338-1>.