ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(31), 2016, с. 249–265

УДК 004.89:004.912

Н. С. Ландо

TimeML для разметки русскоязычных текстов. Оценка перспектив

Аннотация. Статья посвящена анализу возможности применения языка TimeML для разметки временных выражений и их связей с упоминаниями событий в русскоязычных текстах. Выявлен ряд специфических для русского языка конструкций, требующих внесения корректив в инструкцию для аннотаторов, предложены варианты изменений отдельных пунктов инструкции.

В заключении делается вывод о целесообразности практического приложения доработанной версии языка TimeML к русскоязычным текстам как в качестве языка разметки, так и в качестве формата представления извлекаемой автоматически темпоральной информации.

Ключевые слова и фразы: автоматическая обработка текста, извлечение информации, разметка корпусов текстов, темпоральные выражения.

Введение

Все большую актуальность в современном динамичном мире приобретают задачи автоматизации тех или иных процессов. В том числе речь идет и об автоматическом извлечении информации из текстов на естественном языке, что позволяет значительно ускорить обработку и систематизацию больших объемов данных. Извлечение информации из текста (information extraction) — это представление того, о чем в нем говорится, в некотором структурированном виде. Одним из компонентов такой информации является локализация тех или иных событий (см. ниже) во времени.

Чтобы оценить, насколько эффективна разрабатываемая система автоматического извлечения информации, как правило, про водится сравнение результатов ее работы на некоторой базе текстов с подготовленной заранее «идеальной» разметкой тех же текстов. Для того,

Работа выполнена в рамках НИР «Моделирование модально-временного аспекта описания ситуаций в задаче извлечения информации из текстов», номер гос. регистрации 01201455353 (0077-2014-0005).

чтобы подобное сопоставление оказалось возможно, результаты, выданные системой, и «идеальная» разметка должны быть представлены в одном формате.

В настоящее время большое распространение и практическое применение в мире получил язык TimeML, разработанный группой ученых во главе с Джеймсом Пустейовски (James Pustejovsky) из Университета Брандейса. Первая версия языка TimeML была представлена в 2004 году. В 2009 году язык TimeML был признан Международной организацией по стандартизации (International Organization for Standardization, ISO) как международный стандарт для разметки времени и событий в тексте — ISO-TimeML [1]. Уже более десятилетия крупнейшие международные соревнования разработчиков систем извлечения информации не обходятся без этого языка в той части, что касается темпоральной информации. Подробнее о подобных соревнованиях и о современных методах автоматического анализа темпоральных выражений мы уже говорили в [2].

Разрабатывая собственную систему извлечения информации ИСИДА-Т [3], наша команда также столкнулась с задачей извлечения информации о времени. Решение данной задачи предполагает, в первую очередь, построение модели времени и описание на ее основе способов выражения временной информации в тексте. В поисках подобной модели мы обнаружили, что для русского языка достаточно полного и пригодного для практического использования инструмента до сих пор нет. Встречаются либо описания вне связи с приложением к системам АОТ [4], либо модели, создаваемые под конкретную и достаточно узкую задачу [5]. Вполне логично, что в подобной ситуации мы увидели два выхода: либо создавать собственную оригинальную модель времени, либо искать подходящий инструмент среди разработок зарубежных коллег.

В данной работе мы взяли язык TimeML, уже достаточно широко применяемый к самым разным языкам мира и не только европейским [6, 7]. В качестве итога проведенного нами анализа мы предполагаем ответить на вопрос, настолько ли язык TimeML универсален, что с его помощью, — возможно, с некоторыми поправками и дополнениями, — хорошо описывается и русскоязычный материал, или от него все-таки стоит отказаться и разрабатывать собственный инструмент.

Особо отметим также, что изначально TimeML — это язык разметки, осуществляемой вручную аннотатором. Создание больших

корпусов размеченных текстов (типа TimeBank) заставило разработчиков задуматься и об автоматизации процесса разметки на начальном ее этапе, то есть фактически двигаться в сторону извлечения информации. Поскольку наши интересы заключаются и в создании работающей системы, и в возможности ее протестировать, будем оценивать TimeML с обеих этих позиций.

1. Как устроен язык TimeML

Язык TimeML позволяет размечать в тексте упоминания событий, указания на время, а также указывать на последовательность событий, связывать упоминания событий и указания на время между собой.

Под «событиями» при этом понимаются любые ситуации, которые случаются или происходят. События могут быть точечными или длящимися определенный период времени. Также к событиям относятся те предикаты (predicates), которые описывают состояния или положения дел (states or circumstances), в которых что-то становится или является верным[1].

Всего язык TimeML оперирует четырьмя тегами для отрезков текста — EVENT, TIMEX3, SIGNAL и MAKEINSTANCE, — а также тремя тегами для установления отношений — TLINK, SLINK и ALINK.

1.1. Разметка событий

Для разметки упоминаний событий используются теги EVENT (событие) и MAKEINSTANCE (экземпляр события). Так, в случаях типа *Иван работал в понедельник и во вторник* глагол будет помечен как EVENT, за которым стоит два MAKEINSTANCE.

Обязательными атрибутами тега EVENT являются уникальный ID-номер *Event ID number*, генерируемый автоматически инструментом аннотирования каждый раз, когда какой-то подстроке присваивается тег EVENT, а также Класс (Class). Каждое событие принадлежит к одному из следующих классов: ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE. PERCEPTION, REPORTING, STATE,

Атрибуты тега MAKEINSTANCE это *Event Instance ID number* — уникальный ID-номер аналогично ID-номеру события, а также ID-номер события, к которому привязан данный «экземпляр».

[1]См. отрывок из инструкции по аннотированию текста по стандарту TimeML в переводе М. Кормалевой: http://ai-center.botik.ru/Airec/index.php/ru/other/mpubtranslations/38--timeml-

Остальные атрибуты — грамматические: Tense, Aspect, Cardinality, Polarity, Modality, Part of Speech, pos — достаточно спорные и заслуживают того, чтобы стать темой отдельного исследования (ср. опыт применения к французскому материалу в [8]).

Два типа отношений — SLINK и ALINK — устанавливаются исключительно между двумя событиями. SLINK — тег для установления субординационных связей между событиями (модальных, фактивных, условных и т.п.). Связь ALINK устанавливается между событием и аспектуальным глаголом (выражением), которому подчинен соответствующий предикат. Значениями атрибута relType данного тега будут INITIATES, CULMINATES, TERMINATES, CONTINUES, REINITIATES [9].

Очевидно, что отношения обоих типов имеют прямое отношение к темпоральности в тексте и должны быть протестированы на возможность применения к материалу на русском языке. Однако в настоящей работе мы хотели бы сосредоточиться вокруг самого явного способа выражения времени в тексте, то есть временных выражений и их тегирования. Разметку событий будем считать априори заданной. Далее по тексту упоминания событий в примерах будут заключены в квадратные скобки.

1.2. Разметка временных выражений

Временные выражения (temporal expressions, TE) в современной версии языка TimeML помечаются тегом TIMEX3. Как следует из названия тега, это уже не первая его версия. Сами разработчики в качестве принципиального отличия от предыдущих версий (TIMEX и TIMEX2) указывают сужение границ размечаемого временного выражения — минимум зависимых, никаких приложений, придаточных предложений времени и т.п [10].

Так, например, в предложении

За два дня до ***выборов*** *на Украине* ***определилась*** *тройка лидеров*

выборов — это EVENT (и MAKEINSTANCE), так же, как и *определилась*. В качестве TIMEX3 будет размечено словосочетание *два дня*. Предлоги же *за* и *до* получат каждый помету SIGNAL, которая служит для разметки элементов текста, указывающих на отношение, существующее между двумя сущностями — либо TE и событием, либо между двумя событиями, либо между двумя TE. «Сигналами» могут быть предлоги, например, *за, до* (см. пример выше), *на (на два дня), в течение (в течение месяца)* и др., союзы *после того как, до того как, когда* и др., а также некоторые знаки препинания, например, тире в выражении *5–10 августа*.

TimeML

```
<SIGNAL sid="s1">
За
</SIGNAL>
<TIMEX3 tid="t1" type="DURATION" value="P2D" temporalFunction="false">
два дня
</TIMEX3>
<SIGNAL sid="s2">
до
</SIGNAL>
<EVENT eid="e1" class="OCCURRENCE" tense="NONE" aspect="NONE">
выборов
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"/>
на Украине
<EVENT eid="e2" class="OCCURRENCE" tense="PAST" aspect="PERFECTIVE">
определилась
</EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2"/>
тройка лидеров.
<TLINK eventInstanceID="ei1" signalID="s2" relatedToEvent="ei2"
relType="AFTER" magnitude="t1"/>
```

Рис. 1. Разметка предложения *За два дня до выборов на Украине определилась тройка лидеров*

Обязательными атрибутами TIMEX3 являются Timex ID number по аналогии с событиями, Type со значениями DATE — календарное время, TIME — время дня, DURATION — длительность и SET — последовательность, Value — нормализованное значение временного выражения (в соответствии со стандартом ISO-TimeML), а также temporalFunction для указания на то, является ли значение выражения полностью определенным или «недоопределенным» (см. п. 2.3 ниже).

Остальные атрибуты — anchorTimeID, beginPoint, endpoint, freq, functionInDocument, mod, quant — указываются только для выражений определенных типов. Подробнее об этих атрибутах см. в [9].

1.3. TLINK

Наконец, самое главное: между событиями *выборы* и *определилась* в примере выше устанавливается отношение TLINK. Разметка предложения в целом будет выглядеть так, как это показано на рис. 1. Отношения TLINK могут быть установлены не только между двумя

событиями, но и между TIMEX3 и EVENT: для предложения

Бывший президент Индонезии Абдуррахман Вахид ***прибыл*** *в пятницу на* ***лечение*** *в США*

аннотация завершается фрагментом

TimeML

```
<TLINK eventInstanceID="ei1" signalID="s1" relatedToTime="t1"
       relType="INCLUDES"/>
```

в котором ei1 соответствует событию *прибыл*, s1 — предлогу *в*, а t1 — указанию на время *пятницу*.

Из примеров выше очевидно, что формат задания TLINK предполагает указание ID тех событий (экземпляров событий) или те, между которыми устанавливается связь, указание signalID, если таковой присутствует в тексте, а также типа отношения — BEFORE, AFTER, IS_INCLUDED, INCLUDES, IDENTITY, DURING, DURING_INV, SIMULTANEOUS, IAFTER, IBEFORE, BEGUN_BY, ENDED_BY, BEGINS, ENDS.

Набор отношений взят из интервальной логики Аллена [11]. К оригинальным алленовским 13 типам авторы TimeML добавили тип IDENTITY для случая, когда речь идет об одном и том же событии (*В прошлом месяце он* ***ездил*** *в Париж, и вчера рассказывал нам о* ***своей поездке***), то есть когда фактически имеет место кореферентность. Спорный момент. В [8] приводится соображение, с которым трудно не согласиться: введение отношения типа IDENTITY приводит к смешению понятий темпоральности и кореферентности.

Вероятно, проблема в том, что авторы TimeML допустили установление отношений типа алленовских не исключительно между временными интервалами, но и между событиями как наполнениями неких интервалов. Более того, участниками TLINK в принципе может быть и «неоднородная» пара EVENT-TIMEX3 (см. пример выше). Если все-таки развести события и временные интервалы, можно было бы говорить о кореферентности событий, с одной стороны, и об идентичности соответствующих им временных интервалов, с другой. Вместе с тем, разметка самих событий как участников TLINK упрощает ее структуру.

Набор типов отношений, очевидно, может быть доработан и в сторону увеличения: авторы той же французской работы, например, предлагают ввести тип IS_EXCLUDED, для *воскресенья* в выражении *магазин открыт ежедневно, кроме воскресенья*.

2. TimeML в применении к русскому языку

В целом процедура разметки с помощью инструментария языка TimeML выглядит вполне логичной и продуманной. Тем не менее, наш опыт применения данной процедуры к материалу русскоязычных новостных текстов выявил ряд сложностей, которые не позволяют применить инструкцию по аннотированию текстов по стандарту TimeML в том виде, как она есть. Ниже будут перечислены основные моменты, в которых мы столкнулись с затруднениями. Мы постараемся также подумать, каким образом можно было бы изменить инструкцию, чтобы адаптировать ее под русскоязычный материал.

2.1. Проблема №1 — что является временным выражением

Информация о времени разбросана по тексту в настолько разных формах (семантика слов, относящихся к разным частям речи, видовременные характеристики предикатных слов, выбор синтаксической конструкции), что вычленить ее — крайне непростая задача. На сегодняшний день можно смело утверждать, что модели, полностью описывавшей бы темпоральность в тексте на естественном языке, нет. Авторы TimeML, так же, как и прочие исследователи, вынуждены просто игнорировать те явления, которые они пока не в силах описать своей моделью. Задав в качестве обязательных атрибутов Type и Value тега TIMEX3, о которых сказано выше, язык TimeML постановил, что размечены могут быть только те фрагменты, которые

(*а*) отвечают либо на вопрос «когда?» (типы DATE и TIME), либо на вопрос «как долго?» (тип DURATION), либо на вопрос «как часто?» (тип SET);
(*б*) могут быть нормализованы в соответствии со стандартом ISO-TimeML.

«Отлавливание» TIMEX3 в тексте происходит с помощью заданного перечня так называемых слов-триггеров. Для того чтобы то или иное выражение было размечено как TIMEX3, оно должно содержать одно из слов из данного перечня. Это тем более важно в том случае, если «отлавливание» как первый этап разметки производится автоматически.

Примерами слов-триггеров являются английские аналоги слов *день, неделя, сегодня, сейчас, три (он пришел в три)* и т.д. Часть слов-триггеров иногда употребляется и не во временном значении. В этом случае триггерами они не считаются: ср. *следующий* в *на следующий*

день и *следующая дверь*. Аналогичным образом не размечаются выражения, содержащие слова-триггеры, но сами по себе не являющиеся временными выражениями, например, названия произведений таких как *«1984» Джорджа Оруэлла*.

В то же время целый ряд слов, которые несомненно имеют отношение к выражению времени в тексте, но не могут быть описаны в указанном выше формате, приходится исключать из списка «триггерных». Соответственно, TIMEX3 в этом месте не размечается. «Нетриггерными» (согласно существующей инструкции) оказываются аналоги русских слов *мгновение, момент* (единицами времени дня признаны часы, минуты и секунды, но никак не моменты), *немедленно, вскоре, ранее, часто, однажды* и др.

Далее все зависит от постановки задачи. Так, в пространстве новостных сообщений, медицинской и пр. документации, из которых, как правило, и состоят тестовые корпуса, слова, подобные перечисленным выше, встречаются относительно редко, и объем «упущенной» информации не будет критически велик. ИСИДА-Т также имеет дело, прежде всего, с новостными текстами и документами, поэтому мы считаем, что можем здесь последовать ограничениям, предложенным авторами TimeML. Тем не менее, очевидно, что в будущем модель времени в TimeML в целом должна быть усовершенствована с тем, чтобы охватить весь спектр темпоральных выражений. Данная задача достойна отдельного масштабного исследования, в качестве подготовки к которому было бы крайне полезно вести учет всем временным выражениям, не подлежащим разметке в TimeML с целью их дальнейшей систематизации и описания.

2.2. Проблема №2 — границы временного выражения

Как уже было сказано выше, современная версия языка TimeML ратует за максимально короткие TIMEX3. В состав темпорального выражения включаются определения, стоящие слева от триггерного слова (перед ним), но не включается все, что стоит справа (после него). Если справа приложение, которое само по себе имеет темпоральное значение, то оно выделяется в самостоятельный TIMEX3. Зависимые от верхушки выражения предложно-падежные группы и придаточные предложения, стоящие справа и обозначающие связанное с ним событие, не входят в группу TIMEX3. Между ним и данным событием будет установлено соответствующее отношение TLINK (см. пример ниже).

TimeML

```
<TIMEX2 VAL="2016-12-25">
неделю до
<TIMEX2 VAL="2017-01-01">
Нового года
</TIMEX2>
</TIMEX2>
```

Рис. 2. Разметка временного выражения *(за) неделю до Нового года* согласно версии TIMEX2

TimeML

```
<TIMEX3 tid="t1" type="DURATION" value="P1W" beginPoint ="t2"
endPoint="t3">
неделю
</TIMEX3>
до
<TIMEX3 tid="t2" type="DATE" value="2017-01-01" temporalFunction="true"
anchorTimeID="t0">
Нового года
</TIMEX3>
<TIMEX3 tid="t3" type="DATE" value="2016-12-25" temporalFunction="true"
anchorTimeID="t1"/>
```

Рис. 3. Разметка временного выражения [*за*] *неделю до Нового года* согласно версии TIMEX3

Кроме того, согласно оригинальной версии инструкции TimeML (для английского языка), выделяемые как TIMEX3 выражения могут быть именами существительными, прилагательными, наречиями, именными группами, но! не предложными группами. Предлоги *on, in, of, from* и др. никогда не входят в состав TIMEX3. Соответственно, в русском языке такие выражения как *в пятницу*, *на два дня* мы должны рассматривать как сочетание SIGNALов *в* и *на* и TIMEX3 *пятницу, два дня*.

Упомянем также тот факт, что, в отличие от версии TIMEX2, последний вариант инструкции не предполагает никаких вложений одного TIMEX в другое. Так, если ранее временное выражение в *Зимний сезон* ***начнется*** *в горах Сочи за неделю до Нового года* выглядело бы так, как это показано на рис. 2, то теперь это целых три отдельных TIMEX3 (см. рис. 3), причем последний из них «пустой» — ему не соответствует никакой отрезок текста. Остается, правда, вопрос,

почему взаимоотношение между *Новым годом* и событием *начнется* не описать соответствующим TLINK без привлечения «пустого» t, аналогично тому, как это происходит в примере ниже:

В ноябре — через несколько недель после того, как Игорь Слюняев ***сменил*** *Олега Говоруна в кресле главы ведомства — Иванова была* ***назначена*** *директором департамента правового обеспечения министерства регионального развития.*

Учитывая все выше сказанное, в этом предложении следует разметить два TIMEX3 — *ноябре* (t1) и *несколько недель* (t2), три SIGNAL — *в* (s1), *через* (s2) и *после того, как* (s3), а также два EVENT (и MAKEINSTANCE) — *сменил* (ei1) и *назначена* (ei2).

Взаимосвязь между всеми этими элементами будет описана следующими отношениями:

TimeML

```
<TLINK eventInstanceID="ei2" signalID="s1" relatedToTime="t1"
relType="INCLUDES"/>

<TLINK eventInstanceID="ei1" signalID="s3" relatedToEvent="ei2"
relType="BEFORE" magnitude="t2"/>
```

Тем не менее, попытавшись строго следовать перечисленным выше принципам определения границ временных выражений при разметке текстов на русском языке, мы столкнулись со случаями, когда подобная разметка выглядит абсолютно нелогичной. Так, например, в русском языке есть выражения типа *без 15 четыре*, которые совершенно «разваливаются» без предлога (если только не вводить новый тип временных выражений «часть времени дня» и усматривать здесь два TIMEX3, связанных через предлог *без*). В английском языке аналогичное выражение не выносит предлог вперед — *quarter to 4* — и все в целом размечается как один TIMEX3.

Мы предлагаем отметить данный тип задания времени дня при составлении русской версии инструкции TimeML как особый случай и размечать как единый TIMEX3, включая предлог *без*.

Обзор работ исследователей, пытающихся применять TimeML к другим языкам, отличным от английского, показал, что и там лексические, грамматические и синтаксические особенности языка заставляют вносить в этом пункте коррективы. Так, например, в [12] указывается, что при разметке текста на итальянском языке следует включать в TIMEX3 предлоги и модификаторы, которые влияют

TimeML

```
На
<TIMEX3 tid="t1" type="DATE" value="2016-W42" temporalFunction="true"
anchorTimeID="t0">
прошлой неделе
</TIMEX3>
```

Рис. 4. Разметка временного выражения *На прошлой неделе*

на определение значения временного выражения (*verso le 21.30* «около 21.30», *l'anno in esame* «рассматриваемый год»). Как один TIMEX3 размечаются также любые показания часов — *11 e 30* «11.30», *le 12 e mezza* «12 с половиной» (ср. с русской конструкцией с *без* выше).

2.3. Проблема № 3 — разметка недоопределенных выражений. Эллипсис и анафора

Отдельную проблему для аннотаторов и тем более для разработчиков систем автоматического извлечения информации представляют различного вида анафора и эллипсис. В рамках темы данной статьи речь идет о тех временных выражениях, значение которых не может быть вычислено без привлечения контекста либо вообще внетекстовой информации, например, времени создания текста (Data Creation Time, DCT).

TimeML как язык аннотирования вполне справляется с подобными случаями, для описания которых в перечень атрибутов TIMEX3 вводятся temporalFunction и anchorTimeID. TemporalFunction — это собственно указание на то, достаточно ли информации в самом выражении или нет. Считается, что по умолчанию ее недостаточно и атрибут имеет значение TRUE, которое меняется на FALSE, если выражение все-таки содержит всю информацию, необходимую для определения его значения. Примеры полностью определенных выражений: *twelve o'clock January 3, 1984; summer of 1964; Friday, October 1, 1999; 9 a.m. Friday, October 1, 1999; the morning of January 31, 1999.* Их русские аналоги также не вызывают сложностей с определением значения Value.

Пример «недоопределенного» выражения:

На прошлой неделе акции «Мечела» ***подорожали*** *на впечатляющие 30%.*

Аннотатор подобного текста автоматически отсчитывает значение выражения *прошлая неделя* от времени создания текста (24.10.2016),

которому при разметке присваивается отдельный ID (здесь t0). В разметке же самого недоопределенного выражения появится anchorTimeID — ID того временного выражения, с помощью которого высчитывается значение размечаемого TIMEX3 (рис. 4).

Но это в том случае, если разметка производится человеком. Как только мы ставим перед собой задачу автоматизации процесса разметки, и в том числе недоопределенных временных выражений, мы фактически переходим к поискам путей автоматического извлечения темпоральной информации из текста. В этом, пожалуй, и состоит главный вызов для TimeML как для формата представления извлекаемой автоматически темпоральной информации.

С одной стороны, работа в этом направлении давно идет, ведь, как уже было сказано в самом начале данной работы, язык TimeML создавался как единая платформа для тестирования разнообразных систем извлечения информации. Авторы систем, принимающих участие в подобных соревнованиях, неизбежно приводят полученные на выходе данные к формату TimeML. Иначе выявить победителя было бы невозможно. В качестве довольно успешного опыта можно привести, например, проект Stanford JavaNLP, в рамках которого реализован анализатор временных выражений SUTime [**13**].

С другой стороны, практически все участники соревнований, включая команду SUTime, отмечают следующие проблемы сектора извлечения темпоральной информации: затруднена нормализация недоопределенных выражений (как правило, все они привязываются к времени создания текста, что далеко не всегда оправдано), невозможность учесть весь спектр существующих в языке конструкций (варианты задания промежутков и пр. в случае SUTime), необходимость достаточно сильной коррекции при попытке применения программы к текстам на ином языке, чем тот, под который она создавалась.

Возвращаясь к вопросу о формальных моделях именно для русского языка, отметим также работу [**14**], автор которой задается целью разработки лингвистической модели календарного времени. Нам представляется наиболее целесообразным воспользоваться языком TimeML и уже готовыми наработками его авторов и их последователей в виде разнообразных корпусов аннотированных текстов и работ, описывающих опыт применения TimeML к языкам с очень разной структурой, обогатить и усовершенствовать его, в том числе, обратившись к работам российских исследователей темпоральности.

Заключение

Итак, нами описан опыт приложения инструкции по аннотированию текстов согласно стандарту TimeML к текстам на русском языке. Выявлен ряд позиций инструкции, которые необходимо изменить либо дополнить при адаптации к русскоязычному материалу. Приходится признать, что в том виде, в каком инструкция есть на данный момент, ее никак нельзя назвать полной и исчерпывающе описывающей указания на время в тексте.

Вместе с тем, проведенное нами исследование позволяет сделать заключение о возможности применения языка TimeML для разметки временных выражений и событий в текстах определенной тематики (например, новостных сообщениях или разнообразных документов, не изобилующих сложными художественными оборотами и предпочитающими максимально конкретное выражение смыслов) на русском языке при условии тщательной доработки и адаптации инструкции для аннотаторов с учетом специфики грамматики и синтаксиса русского языка. Главное преимущество TimeML, на наш взгляд, заключается в том, что на сегодняшний момент это наиболее широко применяемый в мире инструмент, созданный под задачу представления темпоральной информации в тексте на естественном языке. Расширение спектра описываемой им информации, поиск средств для разметки тех фрагментов текста, которые явно имеют отношение к темпоральности, но современной версией языка никак не отлавливаются, а также охват все большего количества языков мира должны способствовать не только совершенствованию самого TimeML как языка аннотирования, но и развитию систем автоматического извлечения информации.

Благодарности. *Автор статьи благодарит весь коллектив ИЦИИ ИПС им. А. К. Айламазяна РАН и особенно Е. А. Сулейманову и И. В. Трофимова за продолжительные и плодотворные дискуссии на темы, связанные с данной работой.*

Список литературы

[1] J. Pustejovsky, K. Lee, H. Bunt, L. Romary. “ISO-TimeML: An International Standard for Semantic Annotation”, *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC’10 (Valletta, Malta, 17–23 May, 2010), pp. 394–397. ↑ 250

[2] Н. С. Ландо. «Современные методы автоматического анализа темпоральных выражений в текстах на естественном языке», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 419–439, URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf ↑ 250

[3] Д. А. Кормалев, Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. «Извлечение информации из текста в системе ИСИДА-Т», *Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции*, Труды XI Всероссийской научной конференции RCDL'2009 (Петрозаводск, Россия, 17–21 сентября, 2009), КарНЦ РАН, Петрозаводск, 2009, с. 247–253. ↑ 250

[4] Е. Я. Титаренко. «Выражение временной локализованности ситуаций в русском языке», *Вісник Дніпропетровського університету: Серія «Мовознавство»*, **15**/**3**:11 (2009), с. 133–140, URL: http://www.nbuv.gov.ua/old_jrn/Natural/Vdpu/Movozn/2009_15_3/article/22.pdf ↑ 250

[5] В. В. Цибульский, А. С. Ежов, Г. А. Поляков, Г. Г. Феклистов. *Анализ и классификация характеристик времени и сроков в российских нормативных правовых актах*, Диалог 2012 (Бекасово, Россия, 30 мая–3 июня, 2012), URL: http://www.dialog-21.ru/media/1389/119.pdf ↑ 250

[6] A. Bittar et al. "French TimeBank: An ISO-TimeML Annotated Reference Corpus", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (Portland, Oregon, USA, June 19–24, 2011), pp. 130–134, URL: https://www.aclweb.org/anthology/P11-2023 ↑ 250

[7] J.-S. Jeong et al. "Korean TimeML and Korean TimeBank", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (Portorož, Slovenia, May 23–28, 2016), pp. 356–359, URL: http://www.lrecconf.org/proceedings/lrec2016/pdf/175_Paper.pdf ↑ 250

[8] A. Lefeuvre-Halftermeyer et al. "Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (Portorož, Slovenia, May 23–28, 2016), pp. 3802–3806. ↑ 252, 254

[9] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky. *TimeML Annotation Guidelines*, Version 1.2.1, January 31, 2006, URL: http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf ↑ 252, 253

[10] *Guidelines for Temporal Expression Annotation for English for TempEval 2010*, TimeML Working Group, August 14, 2009, 14 p., URL: http://www.timeml.org/tempeval2/tempeval2trial/guidelines/timex3guidelines-072009.pdf ↑ 252

[11] J. F. Allen. "An interval-based representation of temporal knowledge", *Proceedings of the 7th International Joint Conference on Artificial Intelligence*. V. 1, IJCAI'81 (University of British Columbia, Vancouver, Canada, 24–28 August, 1981), San Francisco, CA, USA, 1981, pp. 221–226. ↑ 254

[12] T. Caselli et al. “Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank”, *Proceedings of the Fifth Linguistic Annotation Workshop*, LAW-V (Portland, Oregon, USA, 23–24 June, 2011), ACL, 2011, pp. 143–151, URL: http://www.aclweb.org/anthology/W11-0418 ↑ 258

[13] A. X. Chang, C. D. Manning. “SUTIME: A Library for Recognizing and Normalizing Time Expressions”, *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012 (Istanbul, Turkey, May 21–27, 2012), URL: http://nlp.stanford.edu/pubs/lrec2012-sutime.pdf ↑ 260

[14] Е. А. Сулейманова. «Семантический анализ контекстных дат», *Программные системы: теория и приложения*, **6**:4(27) (2015), с. 367–399, URL: http://psta.psiras.ru/read/psta2015_4_367-399.pdf ↑ 260

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

Пример ссылки на эту публикацию:

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

Об авторе:

Наталья Сергеевна Ландо

Инженер Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН. Научные интересы: извлечение информации из текстов на естественных языках, автоматическая обработка текстов, формальные лингвистические модели

e-mail: nlando1@gmail.com

Natal'ya Lando. *TimeML markup language for Russian. Future outlook.*

Abstract. The article discusses the possibility of applying the TimeML markup language for annotating temporal and event expressions in Russian. The author reveals some cases specific to Russian that do not quite fit in the TimeML guidelines, and suggests possible updates to get around the problem. The conclusion is that an updated version of TimeML for Russian can serve both as a markup language and as a storage format for automatically extracted temporal information. (*In Russian*).

Key words and phrases: natural language processing, information retrieval, annotation language, time expressions.

References

[1] J. Pustejovsky, K. Lee, H. Bunt, L. Romary. "ISO-TimeML: An International Standard for Semantic Annotation", *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC'10 (Valletta, Malta, 17–23 May, 2010), pp. 394–397.

[2] N. S. Lando. "Up-to-Date Methods of Automatic Time Expression Resolution in Natural Language Texts", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:4(27) (2015), pp. 419–439 (in Russian), URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf

[3] D. A. Kormalev, Ye. P. Kurshev, Ye. A. Suleymanova, I. V. Trofimov. "Information extraction in ISIDA-T system", *Digital Libraries: Advanced Methods and Technologies*, Proceedings of the RCDL 2009 (Petrozavodsk, Rossiya, 17–21 sentyabrya, 2009), KarNTs RAN, Petrozavodsk, 2009, pp. 247–253 (in Russian).

[4] Ye. Ya. Titarenko. "Expression of time localization of the situation in Russian language", *Visnik Dnipropetrovs'kogo universitetu: Seriya "Movoznavstvo"*, **15/3**:11 (2009), pp. 133–140 (in Russian), URL: http://www.nbuv.gov.ua/old_jrn/Natural/Vdpu/Movozn/2009_15_3/article/22.pdf

[5] V. V. Tsibul'skiy, A. S. Yezhov, G. A. Polyakov, G. G. Feklistov. *Analysis and Classification of Time and Terms Characteristics in Russian Legal Acts*, Dialog 2012 (Bekasovo, Rossiya, 30 maya–3 iyunya, 2012) (in Russian), URL: http://www.dialog-21.ru/media/1389/119.pdf

[6] A. Bittar et al. "French TimeBank: An ISO-TimeML Annotated Reference Corpus", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (Portland, Oregon, USA, June 19–24, 2011), pp. 130–134, URL: https://www.aclweb.org/anthology/P11-2023

[7] J.-S. Jeong et al. "Korean TimeML and Korean TimeBank", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (Portorož, Slovenia, May 23–28, 2016), pp. 356–359, URL: http://www.lrecconf.org/proceedings/lrec2016/pdf/175_Paper.pdf

[8] A. Lefeuvre-Halftermeyer et al. "Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility", *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (Portorož, Slovenia, May 23–28, 2016), pp. 3802–3806.

[9] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, J. Pustejovsky. *TimeML Annotation Guidelines*, Version 1.2.1, January 31, 2006, URL: http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf

[10] *Guidelines for Temporal Expression Annotation for English for TempEval 2010*, TimeML Working Group, August 14, 2009, 14 p., URL: http://www.timeml.org/tempeval2/tempeval2trial/guidelines/timex3guidelines-072009.pdf

[11] J. F. Allen. "An interval-based representation of temporal knowledge", *Proceedings of the 7th International Joint Conference on Artificial Intelligence*. V. 1, IJCAI'81 (University of British Columbia, Vancouver, Canada, 24–28 August, 1981), San Francisco, CA, USA, 1981, pp. 221–226.

[12] T. Caselli et al. "Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank", *Proceedings of the Fifth Linguistic Annotation Workshop*, LAW-V (Portland, Oregon, USA, 23–24 June, 2011), ACL, 2011, pp. 143–151, URL: http://www.aclweb.org/anthology/W11-0418

[13] A. X. Chang, C. D. Manning. "SUTIME: A Library for Recognizing and Normalizing Time Expressions", *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012 (Istanbul, Turkey, May 21–27, 2012), URL: http://nlp.stanford.edu/pubs/lrec2012-sutime.pdf

[14] Ye. A. Suleymanova. "Semantic Analysis of Contextual Dates", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:4(27) (2015), pp. 367–399 (in Russian), URL: http://psta.psiras.ru/read/psta2015_4_367-399.pdf

Sample citation of this publication:

Natal'ya Lando. "TimeML markup language for Russian. Future outlook", *Program systems: Theory and applications*, 2016, **7**:4(31), pp. 249–265. (*In Russian*). URL: http://psta.psiras.ru/read/psta2016_4_249-265.pdf