ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(27), 2015, с. 419–439

УДК 004.912:519.767.6

Н. С. Ландо

Современные методы автоматического анализа темпоральных выражений в текстах на естественном языке

Аннотация. Статья посвящена вопросу извлечения из текста на естественном языке информации о времени путем анализа указаний на время, которые в нем содержатся.

В работе представлен обзор современных достижений в данной области, прежде всего в применении к текстам на английском языке. Рассмотрены методы и подходы, используемые авторами западных систем автоматического анализа текстов, а также предложение по использованию аналогичного подхода к русскоязычному материалу. Выявлены наиболее проблемные с точки зрения анализа участки текста, а именно — анафорические, дейктические и эллиптированные элементы темпоральных выражений

Ключевые слова и фразы: автоматическая обработка текста, извлечение информации, временна́я референция, темпоральные выражения.

Введение

Одним из самых серьезных вызовов, которые поставили перед собой современные исследователи в области искусственного интеллекта и математической лингвистики, является задача обработки естественного языка, или Natural Language Processing (NLP). В частности, к сфере NLP относится проблема извлечения информации из текста. В данной статье речь пойдет об извлечении информации о времени, которая используется далее при привязке упомянутых в тексте событий ко временно́й оси, то есть временно́й референции событий. Под «событием» в данном случае понимается любое действие, процесс, ситуация, состояние и т. п., упомянутые в тексте (событие «в метаязыковом смысле» ([1]).

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

© Н. С. Ландо, 2015
© Институт программных систем имени А. К. Айламазяна РАН, 2015
© Программные системы: теория и приложения, 2015

Чтобы привязать события ко времени, необходимо для начала извлечь из текста события и указания на время и затем связать одни с другими. На время в естественном языке указывают либо видовременные характеристики глагольных форм, либо специальные временны́е, или темпоральные, выражения — Time Expressions (TE), например, *вчера*, *в прошлом году*, *20-го апреля*, *потом* и т. д., а также придаточные предложения времени. В нашем случае речь пойдет именно о темпоральных выражениях.

1. Сложности при интерпретации TE

Предположим, что все TE в тексте уже найдены. Теперь встает вопрос, как их проинтерпретировать. Для этого необходимо построить модель времени, которое традиционно представляется в виде оси времени (реальной, воображаемой и т. п.). Задача интерпретации сводится тогда определению точки (интервала, отрезка, группы точек или интервалов) на этой оси, которые соответствуют данному указанию на время.

Задача привязки TE к временно́й оси усложняется следующим обстоятельством. В одних случаях указание может быть вполне конкретно, например, *5-го июня 2002-го года*. Построив ось реального времени, мы легко можем отметить на ней отрезок, соответствующий этой дате. Однако, в ряде других случаев подобная интерпретация может быть затруднена невозможностью извлечения достаточной информации непосредственно из TE в связи с тем, что оно содержит дейктические или анафорические элементы, или же имеет место эллипсис элементов.

Под *дейксисом* принято понимать использование языковых выражений и других знаков, которые могут быть проинтерпретированы лишь при помощи обращения к физическим координатам коммуникативного акта — его участникам, его месту и времени (см. [**2**]).

Для времени это, например, интерпретация таких выражений как *сейчас, сегодня, вчера, на прошлой неделе* и т. п.

Анафора — использование языковых выражений, которые могут быть проинтерпретированы лишь с учетом другого, как правило, предшествующего, фрагмента текста (см. [**3**]). Анафорические элементы есть в выражениях *в том же году, за день до того* и т. п.

К проблеме разрешения дейксиса и анафоры вплотную примыкает и проблема восстановления эллиптированных компонентов темпорального выражения. Так, в примере «*В апреле 2000-го года Елена вышла*

замуж, а в сентябре у нее уже родился ребенок» для выражения *в сентябре* носитель русского языка безошибочно восстановит год — тот же 2000-й. Восполнение подобных пробелов автоматически — это уже сложная задача в рамках NLP.

2. Поиск решений

2.1. Проведение международных конференций-соревнований

В 80-х годах прошлого столетия появилась идея объединить усилия исследователей в области извлечения информации из текста на естественном языке в поисках более эффективных решений имеющихся проблем. С этого времени регулярно начали проводиться соревнования между заинтересованными группами исследователей, желающих протестировать свои разработки. Результаты каждого подобного соревнования затем публикуются и обсуждаются на соответствующей конференции (воркшопе).

Первыми соревновательными площадками из этой серии стали Message Understanding Conferences (MUC) по инициативе и при финансировании DARPA (Defense Advanced Research Projects Agency) в США (семь конференций с 1987 по 1997 год). Еще один проект, спонсированный DARPA, — это соревнования Automatic Content Extraction (ACE), проводившиеся с 1999 по 2008 год. Далее эстафету переняли TempEval в рамках SemEval (2007(1), 2010(2), 2013(3), 2015 (QA-TempEval и Clinical TempEval)).

По мере проведения подобных мероприятий меняется и развивается их концепция, но суть ее остается неизменной. На входе перед участниками ставится определенная задача (задачи, tasks), — например, извлечь из корпуса текстов темпоральные выражения, — а также уже размеченный в соответствии с поставленной задачей корпус текстов, который можно использовать для предварительного обучения тестируемых систем. На этапе собственно соревнования команды получают другой, неразмеченный, корпус и испытывают на нем свои системы. Результаты затем сравниваются с эталонной разметкой того же корпуса, сделанной предварительно вручную носителями языка. Начиная со второй конференции MUC (1989), формат разметки был приведен к единому, унифицированному виду, что дало платформу для объективной оценки и сравнения результатов.

2.2. Критерии оценки

Основными параметрами для оценки эффективности тестируемых систем были признаны *полнота* (*recall*) — отношение числа правильно размеченных единиц, к общему числу единиц в корпусе, которые нужно было разметить, и *точность* (*precision*) — отношение числа правильно размеченных единиц к общему числу размеченных алгоритмом единиц, а также *f-мера* (*f-measure*) — характеристика, которая позволяет дать оценку одновременно по точности и по полноте:

$$f\text{-measure} = \frac{2PR}{P+R},$$

где P — precision, а R — recall.

2.3. Задача извлечения и нормализации TE

Темпоральные выражения впервые фигурируют в инструкциях к MUC-6 (1995) в рамках задачи по распознаванию именованных сущностей (Named Entity Recognition (NER). Наряду с TE, участники должны были извлечь упоминания лиц, организаций, географических названий и числовых выражений.

Согласно принятой на тот момент инструкции, пометы TIMEX на языке SGML[1] получали исключительно «абсолютные» TE. Под «абсолютными» понимались выражения, указывающие на конкретный отрезок времени. Таким выражениям приписывался единственный атрибут TYPE со значением TIME для времени дня и DATE для дней и годов. Таким образом, разметке подлежало выражение *20 minutes after 10* и не подлежали *a few minutes after the hour*, *20 minutes after the hour*. По той же причине неразмеченными оказываются выражения *the current year*, *the last 10 years* и т. п. В выражении *July last year* как TIMEX размечался *July*, а *last year* отбрасывалось. Иными словами, дейктические и анафорические элементы полностью исключались из рассмотрения как требующие дополнительной информации для конкретизации значения (подробнее см. [4]).

MUC-7 продвинулся несколько дальше и включил в разметку именованных сущностей «относительные» TE. В качестве относительных TE (Relative Temporal Expression, RTE) размечались либо указания на дату, исчисляемые по дате написания документа (*вчера*, *сегодня*

[1]SGML (англ. Standard Generalized Markup Language — стандартный обобщённый язык разметки) — метаязык, на котором можно определять язык разметки для документов.

и т. п.), либо указания на часть конкретно заданного временнóго интервала (например, *утро* как начало уже заданного дня). Таким образом, в поле зрения участников MUC-7 (по крайней мере, тех из них, кто участвовал в части, посвященной TE), начинают попадать указания, содержащие дейктические элементы, типа *в прошлом месяце*, *в следующем году*, *последние два месяца*, *10 лет назад*, *через четыре месяца* и т. п. (подробнее см. в [**5**]).

История TIMEX получила свое продолжение в 2000 году, когда под эгидой программы DARPA Translingual Information Detection, Extraction and Summarization (TIDES) был разработан новый стандарт — TIMEX2. Принципиальное отличие и прорыв новой программы по сравнению с предыдущей стала постановка в качестве задачи не только распознавания, но и нормализации, то есть интерпретации, приведения дат и указаний на время дня к единому формату, предусмотренному стандартом ISO 8601 (в виде YYYY- MM-DD и т. п.).

Начиная с 2004 года, задача по распознаванию TE была выделена из общей задачи по распознаванию именованных сущностей в отдельную задачу — Time Expression Recognition and Normalization (TERN), то есть задачу распознавания и нормализации (интерпретации) временны́х выражений. К этому времени для разметки TE применялся уже специальный язык TimeML, первая версия которого была окончательно оформлена в 2002 году на семинарах TERQAS (Time and Event Recognition for Question Answering Systems), организованных профессором Джеймсом Пустейовски из Университета Брандейса. Результатом работы семинаров стало появление TimeML version 1.0, а также корпуса TimeBank как иллюстрации применения этого языка. В настоящее время, и в частности на TempEval, используется еще более новый стандарт TIMEX — TIMEX3 (TimeML Specification version 1.2.1).

В 2009 году в соответствии с замыслом его создателей TimeML в его обновленной версии ISO-TimeML стал международно признанным инструментом для разметки времени и событий в текстах на естественных языках (утвержден Международной организацией по стандартизации ISO). Описание языка ISO-TimeML см., в частности, в [**6**].

2.4. Модель времени

Как уже было сказано выше, для интерпретации временны́х выражений и установления связей между ними и событиями в тексте необходимо построение модели времени. Основой для такой модели

Таблица 1. Базовые отношения между двумя интервалами по Аллену

https://en.wikipedia.org/wiki/Allen's_interval_algebra

Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
X **m** Y Y **mi** X	X Y	X meets Y (*i* stands for ***i****nverse*)
X **o** Y Y **oi** X	X Y	X overlaps with Y
X **s** Y Y **si** X	X Y	X starts Y
X **d** Y Y **di** X	X Y	X during Y
X **f** Y Y **fi** X	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

в TIMEX, — а, следовательно, и во всех перечисленных выше соревнованиях, — стала так называемая интервальная логика Аллена [**7**].

Все события, включая те, которые, как нам кажется, происходят одномоментно (например, *нашел*), по Аллену, имеют интервал на временной оси (пусть и нулевой), в течение которого они происходят. В тексте на эти интервалы указывают, прежде всего, временны́е выражения. Так, например, *now* (*сейчас*) в концепции Аллена также является интервалом, который может подвергаться корректировке, если, допустим, время в тексте сдвинулось на день вперед. «Вчерашнее» *now* становится в этом случае *yesterday* (*вчера*), и появляется новое *now*.

В современной версии теории интервалы, соответствующие разным событиям, соотносятся друг с другом тринадцатью разными способами, представленными в таблице 1.

2.5. Основные задачи, связанные с TE

Несколько варьируясь от одного раунда к другому, базовые задачи, ставящиеся перед участниками вышеупомянутых соревнований, сводятся к следующему:

(1) извлечение и нормализация TE (определить границы всех темпоральных выражений, сделав пометы TIMEX (позднее TIMEX2, TIMEX3), и приписать каждому TE характеристики TYPE (значения TIME, DATE, DURATION и SET) и VAL (унифицированный вид в соответствии со стандартами TIMEX2 и TIMEX3));
(2) извлечение и классификация событий (определить границы всех указаний на события, сделав пометы EVENT с соответствующими характеристиками);
(3) установление темпоральных отношений между событиями и TE, между событиями и временем создания документа (Document Creation Time, DCT) и между событиями попарно. Для TempEval-1 и TempEval-2 это только отношения типа BEFORE, AFTER, OVERLAP, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER, а также VAGUE для случаев полной неопределенности. На TempEval-3 это уже весь набор отношений языка TimeML (см. интервалы Аллена выше).

3. Эффективные системы и методы

Рассмотрим методы, которые использовали наиболее успешные из участников соревнований TempEval (последние площадки) для решения поставленных задач. В данном случае нас интересует, прежде всего, первая из перечисленных выше задач, то есть извлечение и нормализация TE. Особое внимание обратим на решение проблемы анализа дейктических, анафорических и эллиптированных элементов в TE.

Нужно сказать, что применяющиеся системами машинного анализа методы в целом можно разделить на три группы: (1) методы, основанные на машинном обучении, (2) методы, основанные на разработанных специалистами-лингвистами правилах и (3) комбинированные (гибридные) методы. Организаторы TempEval-3 [8], анализируя его результаты, отмечали, что если ранее задачу извлечения TE лучше выполняли системы с правилами в основе, то теперь уже машинное обучение достигло такого уровня, что позволяет показывать даже

превосходящие результаты. При этом нормализация по-прежнему обеспечивается лучше правилами. Кроме того, они указывают на большое значение постобработки (postprocessing), то есть повторной обработки уже полученных машиной результатов, а также на важность использования как можно большего объема лингвистической информации при условии, что она используется корректно.

3.1. XRCE-T

Несмотря на то, что в перечень задач для участников первого из соревнований серии TempEval (в рамках SemEval-2007) ни распознавание TE, ни их нормализация как раз не входили, обратим внимание на одну из команд, а именно XEROX Research Centre Europe (XRCE-T), которая описала свои результаты в [**9**].

Дело в том, что, имея перед собой цель установить темпоральные отношения, о которых было сказано выше, данная группа исследователей решила не брать за основу те TE и события, которые в тот раз уже были размечены в тестовом корпусе как TIMEX3 и EVENT, а фактически применила свой собственный глубинносинтаксический анализатор XIP (Xerox Incremental Parser), точнее его модификацию XTM (темпоральный анализатор), а затем адаптировала результаты обработки под стандарты TempEval.

Система XRCE-T начинала с того, что по своим правилам извлекала темпоральные выражения и события, связывала одни с другими и упорядочивала события во времени. Базой для работы анализатора являются правила. Система проводит токенизацию, морфологический анализ, частеречную разметку текста, затем поверхностно-синтаксический анализ — членение на фрагменты, распознавание именованных сущностей (NER) — и, наконец, глубинно-синтаксический анализ.

Распознавание и интерпретация TE — это так называемый анализ локального уровня. На рис. 1 ниже изображена схема такого анализа для выражения *4 years ago*, приведенная в [**9**, стр. 493].

С помощью правил, способных учитывать контекст слева и/или справа, выражение вычленяется из текста. Затем для выражения строится адвербиальный узел ADV. `Merge_anchor_and_dur` — это функция в языке Python с тремя узлами в качестве аргументов (`#0` — это все выражение слева в целом).

Из схемы видно, что система использует понятие референтного времени, то есть *времени отсчета*. В данном случае это момент речи

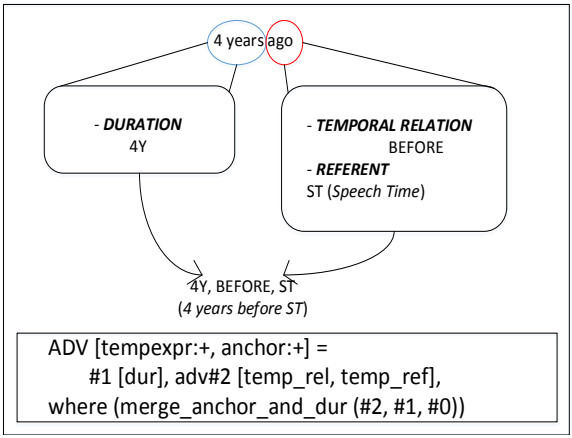

Рис. 1. Схема анализа локального уровня выражения *4 years ago*

(Speech Time). В рамках соревнования TempEval-1 время отсчета автоматически приравнивалось ко времени создания документа.

Установление отношений типа TEMP между TE и предикатами (событиями) осуществлялось уже на уровне предложений и было не более чем частным случаем установления отношений между предикатом и зависимыми от него предложными группами (PP).

Несмотря на то, что принципы, заложенные в основу XRCE-T, довольно сильно отличались от стандартов TempEval, начиная от модели времени и принципов выделения TE и событий и кончая набором устанавливаемых отношений, исследователи сочли, что, вмешательство в работу своего анализатора на этапе извлечения и установления отношений, сильно ухудшило бы результаты его работы в целом.

В качестве примера принципиальных расхождений в [**9**] приводится разное деление на токены составных TE. Так, выражение *ten days ago yesterday* анализатор XTM разметит как одно TE, а выражение *during 10 days in December* разделит на два TE — *during 10 days* и *in December*. Темпоральное отношение, устанавливаемое в TimeML, между упоминанием события и TE, обозначающим его длительность, XTM описывает как атрибут события. За счет подобных расхождений и того, что команда отдала предпочтение собственному подходу к выделению TE и событий и установлению отношений между ними, как считают авторы, система и показала самые низкие уровни полноты. Точность же была показана одна из лучших, что позволило им

оценить выбранную ими тактику как оправдавшую себя.

3.2. HeidelTime

На TempEval-2 в 2010 году распознавание TE и их нормализация уже были выделены как отдельная задача (Task A). Самыми успешными из проектов, участвовавших в TempEval-2, оказались TRIPS и TRIOS (Рочестерский университет, США), а также система HeidelTime (Гейдельбергский университет, Германия).

В той части соревнования, которая касалась извлечения и интерпретации темпоральных выражений (Task A), лучшие результаты показала именно разработанная в стенах Гейдельбергского университета (Германия) система HeidelTime — f-мера=86% при извлечении ([**10**]).

Каждое темпоральное выражение, согласно разработчикам, может быть представлено в виде тройки $tei = \langle e_i, t_i, v_i \rangle$, где e_i — выражение в том виде, в каком оно встретилось в тексте, t_i — тип (`Date`, `Time`, `Duration` или `Set`), v_i — значение выражения в стандартном виде (на языке TimeML). Соответственно, перед HeidelTime ставится задача извлечь из текста все возможные выражения ei и приписать им корректные характеристики t_i и v_i.

Алгоритм работы HeidelTime использует разработанные вручную правила. Правила, применяемые для извлечения темпоральных выражений, опираются, прежде всего, на набор стандартных паттернов подобных выражений. Но не только. Так, в частности, могут быть задействованы ограничения на часть речи для предыдущего или последующего токена и т. п.

После того, как все предложения текста последовательно обработаны блоком правил, извлекающих темпоральные выражения, и все выражения найдены, необходимо распределить их по типам и привести к стандартному виду. Здесь уже начинают работать правила нормализации, которые также опираются на базы стандартных паттернов выражений.

Проблема, как и следовало ожидать, возникает тогда, когда для интерпретации выражения необходима точка отсчета извне, то есть, в случае относительных выражений. В этом случае основной алгоритм HeidelTime признает значение такого выражения недостаточно определенным. Так, значение выражения *last June* будет записано как `UNDEF-last-June`, *June* в *in June* — как `UNDEF-June`, *year-earlier* в *the year-earlier results* — как `UNDEF-REF-last year`. При этом в качестве

времени отсчета для первых двух выражений автоматически устанавливается время создания документа, а последнего — ближайшая упомянутая выше дата.

Окончательно задача разрешения подобной неоднозначности, равно как и отбрасывание ошибочно извлеченных выражений, переносится системой на этап постобработки. Все значения, помеченные `UNDEF-REF`, отсчитываются от ближайшей упомянутой выше даты. В противном случае отсчет производится от времени создания документа. Соответственно, `UNDEF-last-June` — это июнь, предшествующий времени создания документа. С `UNDEF-June` все несколько сложнее и для правильной интерпретации выражения *«in June»* требуется дополнительная лингвистическая информация. Так, например, в предложении *«In June, new results will be published»* глагол стоит в форме будущего времени. Соответственно, программа заменит значение `UNDEF-June` на `UNDEF-next-June` и только затем рассчитает дату по времени создания документа.

Примечательно, что проинтерпретировать и привести выражение к стандартному виду в данном случае фактически означает восстановить эллиптированный фрагмент даты (указание на год). Поскольку форматы для дат заранее заданы и дата, в частности, никак не может указывать на месяц без привязки к конкретному году, система просто обязана эту привязку осуществить.

В тех случаях, когда временем отсчета признается ближайшая упомянутая выше дата, берутся все уже извлеченные TIMEX3, для которых был установлен тип DATE, и в качестве v_{ref} (времени отсчета в стандартном виде) берется значение последнего из этих выражений. Так, для *year-earlier* программа отсчитает год от соответствующего v_{ref}.

3.3. TRIPS и TRIOS

Что касается TRIPS и TRIOS, то возможности данных систем позволили их разработчикам принять участие во всех без исключения соревновательных секторах, то есть и в задачах на извлечение информации о событиях и времени, и в задачах на установление отношений между таковыми [**11**].

Свой подход авторы определяют как сочетание глубинносемантического анализа, с одной стороны, и применения логикомарковских

сетей (Markov Logic Networks) и классификаторов, основанных на модели условных случайных полей (Conditional Random Fields), — с другой. Таким образом, в данном случае мы имеем дело с комбинацией правил и машинного самообучения.

В работе системы участвуют два модуля, которые запускаются параллельно: с одной стороны, проводится глубинно-семантический анализ (анализатор TRIPS), порождающий логические формы, из которых затем с помощью определенных правил извлекаются упоминания событий, их свойства и темпоральные выражения. Одновременно с этим классификаторы, работающие на основе машинного обучения, фильтруют события, извлекают их свойства, темпоральные выражения и классифицируют темпоральные отношения. TRIOS — это, собственно, TRIPS, обогащенная вышеперечисленными надстройками, призванными улучшить результаты работы системы в рамках задач TempEval-2 (прежде всего, в задачах на установление отношений).

Грамматика анализатора TRIPS — это контекстно-свободная грамматика, в основу которой были положены принципы грамматики непосредственных составляющих, а также грамматики GPSG (Generalized Phrase Structure Grammar) и грамматики управляемых вершинами фразовых категорий — Head-driven Phrase Structure Grammar (HPSG).

На вход в TRIPS текст поступает после предварительной его обработки целым рядом уже известных и широко применяемых анализаторов, осуществляющих разметку частей речи, расстановку границ составляющих, интерпретацию незнакомых слов с помощью WordNet, распознавание именованных сущностей. Далее методом поиска по первому наилучшему совпадению строится восходящее дерево составляющих. Специальная структура packed-forest chart позволяет хранить деревья более компактно и убыстряет процедуру анализа.

Авторы неоднократно подчеркивают, что они стремились сделать свою систему максимально универсальной, независимой от предметной области текста и применимой к текстам на любых естественных языках. Набор семантических типов, встречающихся в логических формах, включает около 2500 типов, семантических ролей — 30. По сути это несколько обобщенный вариант FrameNet[2].

Что касается распознавания темпоральных выражений, то его исследователи попытались улучшить, применив классификатор, основанный на модели условных случайных полей (CRF++). С помощью

[2] http://polit.ru/article/2014/02/21/ps_fillmore/

данного классификатора каждому токену темпорального выражения, уже размеченного тегами Beginnig-Inside-Outside (B-I-O), приписываются лексические и синтаксические свойства. Так, лексическими свойствами являются *word, shape, is year, is date of week, is month, is number, is time, is day, is quarter, is punctuation, if belong to word-list like initlist* (для слов типа *this, mid, first, almost, last, next* и т. п.), follow-list (для слов типа *century, centuries, day, days, hour, hours* и т. п.) и др.

Далее в ход идут шаблоны CRF++. Так, шаблон, согласно которому два следующих друг за другом слова, первое из которых входит в *init-list*, а второе — в *follow-list*, должны быть объединены в одно TE, позволит получить такие выражения как *this weekend, earlier morning, several years* и т. п.

Наконец, выражения проходят фильтрацию, во время которой система пытается построить для них характеристики TYPE и VAL (о характеристиках см. выше). Здесь уже работают правила. Выражение принимается как достоверное только в том случае, если и тип, и стандартный вид выражения удается определить.

Для интерпретации относительных выражений типа *last month, Sunday, today* в качестве точки отсчета автоматически выбирается время создания документа.

3.4. Структурная грамматика темпоральных выражений

Свой подход к проблеме интерпретации временны́х выражений протестировала на TempEval-2 и группа исследователей из Стэнфордского университета (США) [**12**].

Команда поставила перед собой задачу создать универсальную грамматику временны́х выражений, с помощью которой, имея на входе базу таких выражений и их стандартное представление, можно было бы обучить машину автоматической интерпретации TE.

В своем подходе авторы основываются на убеждении в том, что указания на время в естественном языке композиционны по своей природе. Иными словами, каждое слово, входящее во временно́е выражение, влияет на значение всего выражения в целом определенным образом. Так, например, в выражении «*the week before last week сначала*» к *week* применяется модификатор *last*, в результате чего получается 'прошлая неделя', а затем применяется *before* к *week* и *last week* и получается 'неделя, предшествующая прошлой'.

С формальной точки зрения, в данном случае строится вероятностная контекстно-свободная грамматика (Probabilistic Context- Free Grammar, PCFG). Любое темпоральное выражение, согласно данной грамматике, представляется как одно из трех:

- Range, промежуток (в том числе момент, Instant) — это период между двумя датами или двумя временами дня. Примеры: *Today*, *1987*, *Now*. Так же, как и у Аллена, момент — это промежуток, длина которого равна 0.
- Sequence — последовательность промежутков, не обязательно отстоящих на равных временны́х промежутках друг от друга. Примеры: *Friday*, *November 27th*, *last Friday*. Для последовательности могут быть построены уже несколько возможных интерпретаций, но одним из них приписывается больший вес, другим — меньший. Распределение вероятностей определяется по закону Гаусса.
- Duration — срок, период времени. Примеры: *Week*, *Month*, *7 days*.

Помимо этого, в состав TE могут входить вспомогательные элементы двух типов. Это Number для числительного, которое само по себе не имеет временно́го значения (например, *2* в *2 weeks*), и Nil, например, для артикля в *a week ago*, который ничего не добавляет к значению выражения в целом, но является дополнительным указателем на то, что это именно TE. «Сборка» сложного выражения осуществляется посредством функций (Function) с переменными перечисленных выше типов. Функциями являются, например,

shiftLeft — shift a Range or Sequence left by a Duration,
shrinkBegin — take the first Duration of a Range/Sequence,
catRight — take Duration units before the start of a Range,
n^{th} x of y — take the nth Sequence in y (Day of Week, etc).

Для каждого сложного (составного) TE система сначала приписывает значение каждому отдельному токену, а затем последовательно соединяет их с помощью функций, выстраивая дерево анализа. Результатом работы алгоритма становятся наиболее вероятные варианты интерпретации для всего выражения в целом.

В отличие от традиционного метода, предполагающего разметку регулярных выражений, а затем обработку вручную разработанными правилами, в данном случае мы имеем дело с рекурсивной системой, учитывающей вероятностный фактор, что позволяет, в частности, разрешать неоднозначность в TE. Так, например, здесь будет учтена разница в значении слова *last* в выражениях *last week of November*

и просто *last week*. Или синтаксическая неоднозначность выражения «*last Friday the 13th*», которое может быть проинтерпретировано и как [*last Friday*][*the 13th*], и как *last* [*Friday the 13th*].

На TempEval-2 программа была протестирована только в той части, что касалась интерпретации TE. В качестве объекта в данном случае был взят корпус с уже размеченными TE (золотой стандарт). В результате показатели эффективности программы оказались практически на уровне лидеров типа HeidelTime.

В качестве несомненных преимуществ своего подхода авторы называют возможность учесть синтаксическою и прагматическую неоднозначность, а также возможность предложить интерпретацию практически для всех TE. Разработанный алгоритм наиболее эффективен как раз при анализе составных выражений, даже более сложных, чем те, что встречались в корпусе, предложенном для тестирования.

Кроме того, подобная система может быть достаточно легко адаптирована как для анализа текстов на других языках, так и под задачи, отличные от интерпретации TE.

Ошибки, допущенные системой, авторы соотносят, прежде всего, с отсутствием необходимых лексических примитивов в ее арсенале. Так, например, *Nov.* (в *Nov. 13*) не было распознано как *November*, а попало в тип `Nil` по аналогии с артиклем в *the 13th*. Не было также проинтерпретировано выражение *1990s*, и *half* в *minute and a half*. Отсутствие функции *or* не позволило правильно проинтерпретировать выражение *weeks or months*.

Помимо этого, в частности, путь Витерби мог оказаться не верным (ошибка обучения) и взята, например, пятница прошлой недели вместо пятницы текущей недели, или имело место несовпадение настоящего времени отсчета со временем создания текста, от которого по умолчанию производился отсчет.

3.5. TempEval-3

Результаты TempEval-3, проходившего в рамках SemEval-2013, можно посмотреть в [8]. Извлечению и нормализации TE была посвящена Задача A (Task A). Несмотря на то, что по многим параметрам большая часть систем-участников показали близкие результаты, авторы обзора выделяют версию HeidelTime-t (tuned) как получившую самую высокую f-меру при определении характеристики VALUE для английского языка (77,61).

Сами разработчики системы свой опыт участия в этом соревновании описали в работе [13] и считают основным своим достижением высокие результаты, показанные системой в испанском секторе соревнований, что подтверждает эффективность HeidelTime как мультилингвального разметчика темпоральных выражений.

Что касается английской части, то авторы по-прежнему отмечают сложность правильной интерпретации неоднозначных TE и TE, привязанных ко времени отсчета. Чтобы снизить количество ошибок при нормализации выражений, они вынуждены отбрасывать подобные выражения, что приводит к снижению уровня полноты.

3.6. Опыт применения к русскоязычным текстам

Методы машинного обучения, столь успешно применяемые уже для извлечения информации о событиях и времени в текстах на английском и ряде других языков, о чем шла речь выше, не могут быть в той же мере использованы для анализа текстов на русском языке. Проблема заключается в отсутствии достаточно большого корпуса качественно аннотированных текстов на русском языке (типа TimeBank).

Тем не менее, идея применения методов, оправдавших себя при анализе англоязычного материала, к русскому языку, существует. Так, группа исследователей из Волгоградского государственного технического университета предлагает схему для «гибридного (то есть, комбинированного — прим. авт.) подхода к обработке временно́й информации в тексте на русском языке» [14].

Задача — установление в тексте на русском языке тех же временны́х отношений языка TimeML. Обработку предполагается проводить в 3 этапа: «На первой стадии используются правила для увеличения размера обучаемого набора. Здесь без особых усилий создается больший аннотированный корпус текстов. На втором этапе, решается задача классификации временны́х отношений с помощью метода машинного обучения. При этом используется корпус текстов для генерации нечетких правил нечеткой нейронной сети. Для оптимизации параметров нечетких правил мы используем метод опорных векторов» ([14, с. 4]).

Из приведенной выше цитаты очевидно, что реализация данной схемы начинается с того же создания корпуса размеченных текстов. Мы полагаем, что совместная работа по составлению подобного корпуса всех заинтересованных групп исследователей, как это было сделано

для английского языка, дала бы огромный толчок для отечественных разработок области автоматической обработки текстов.

Заключение

Результаты тестирования систем на международных площадках позволяют сделать заключение о том, что к настоящему времени уже разработаны достаточно эффективные методы извлечения из текста и нормализации выражений, указывающих на время. Машинное обучение, не требующее такого объема времени и труда специалистов, как написание правил, используется все чаще и показывает все лучшие результаты. Вместе с тем, отсутствие достаточно большого корпуса размеченных текстов на русском языке до сих пор препятствует созданию русскоязычных систем подобного уровня.

Нельзя, однако, не отметить, что ряд сложностей недостаточно учитывается существующими системами анализа и, как следствие, приводит к снижению точности и/или полноты результатов анализа. Так, в частности, на данный момент компьютерные технологии не обладают всеми необходимыми инструментами для качественного разрешения анафорических и дейктических темпоральных ссылок и эллипсиса. Их авторы находятся в поисках подходов к решению этой задачи и вынуждены пока либо абстрагироваться от нее, либо значительно ее упрощать.

Благодарности. *Автор статьи искренне благодарна Елене Анатольевне Сулеймановой и Игорю Владимировичу Трофимову за ценные замечания и помощь в подготовке данного материала.*

Список литературы

[1] Н. Д. Арутюнова, *Типы языковых значений. Оценка, событие, факт*, Наука, М., 1988, с. 169. ↑419

[2] А. Кибрик, «Дейксис», *Энциклопедия Кругосвет*, Универсальная научно-популярная онлайн-энциклопедия, ред. А. В. Добровольский и др., URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/DEKSIS.html?page=0,0 ↑420

[3] А. Кибрик, «Анафора», *Энциклопедия Кругосвет*, Универсальная научно-популярная онлайн-энциклопедия, ред. А. В. Добровольский и др., URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/ANAFORA.html ↑420

[4] Appendix B. Timex: Specific Guidelines, URL: http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_11.html#HEADING38 ↑422

[5] N. Chinchor, P. Robinson, “MUC-7 Named Entity Task Definition”, *Proceedings of the 7th Conference on Message Understanding*, 1997, pp. 29, URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html ↑423

[6] J. Pustejovsky, K. Lee, H. Bunt, L. Romary, “ISO-TimeML: An International Standard for Semantic Annotation”, *Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC’10* (17–23 May 2010, Valletta, Malta), pp. 394–397. ↑423

[7] J. F. Allen, “An interval-based representation of temporal knowledge”, *IJCAI’81 Proceedings of the 7th international joint conference on Artificial intelligence*. V. 1 (24–28 August 1981, Vancouver, Canada), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, pp. 221–226. ↑424

[8] N. UzZaman, H. Llorence, L. Derczynski, M. Verhagen, J. Allen, J. Pustejovsky, “SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations”, *Seventh International Workshop on Semantic Evaluation SemEval-2013*, Co-located with *Sem and NAACL. V. 2 (14–15 June 2013, Atlanta, Georgia, USA), ACL, 2013, pp. 1–9. ↑425,433

[9] C. Hagège, X. Tannier, “XRCE-T: XIP Temporal Module for TempEval Campaign”, *Proceedings of the 4th International Workshop on Semantic Evaluations SemEval-2007* (23–24 June 2007, Prague), ACL, 2007, pp. 492–495. ↑426,427

[10] J. Strötgen, M. Gertz, “HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions”, *Proceedings of the 5th International Workshop on Semantic Evaluation SemEval-2010* (11–16 July 2010, Uppsala, Sweden), ACL, 2010, pp. 321–324. ↑428

[11] N. UzZaman, J. F. Allen, “TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text”, *Proceedings of the 5th International Workshop on Semantic Evaluation SemEval-2010* (11–16 July 2010, Uppsala, Sweden), ACL, 2010, pp. 276–283. ↑429

[12] G. Angeli, C. D. Manning, D. Jurafsky, “Parsing Time: Learning to Interpret Time Expressions”, *NAACL HLT’12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (3–8 June 2012, Montreal, Canada), Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 446–455. ↑431

[13] J. Strötgen, J. Zell, M. Gertz, “HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3”, *Seventh International Workshop on Semantic Evaluation SemEval-2013*, Co-located with *Sem and NAACL. V. 2 (14–15 June 2013, Atlanta, Georgia, USA), ACL, 2013, pp. 15–19. ↑434

[14] А. В. Заболеева-Зотова, Д. К. Фамхынг, С. С. Захаров, «Гибридный подход к обработке временной информации в тексте на русском языке», *Труды одиннадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2008* (28 сентября–3 октября 2008 г., г. Дубна, Россия), URSS, М., 2008, с. 228–235. ↑434

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

Об авторе:

Наталья Сергеевна Ландо

Инженер Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН

e-mail: nlando1@gmail.com

Пример ссылки на эту публикацию:

Н. С. Ландо. «Современные методы автоматического анализа темпоральных выражений в текстах на естественном языке», *Программные системы: теория и приложения*, 2015, **6**:4(27), с. 419–439.

URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf

Natal'ya Lando. *Up-to-date methods of automatic time expression resolution in natural language texts.*

Abstract. The article is concerned with temporal information extraction from natural language text by means of time expression resolution. The paper studies the up-to-date achievements in the area mostly for English. Methods and techniques applied in the western automatic text processing systems are discussed, as well as the problem of using similar approaches for Russian.

Challenging issues are revealed, related to anaphora, deixis and ellipsis in temporal expressions. (*In Russian*).

Key Words and Phrases: natural language processing, information retrieval, time reference, time expressions.

References

[1] N. D. Arutyunova, *Tipy yazykovykh znacheniy. Otsenka, sobytiye, fakt*, Nauka, M., 1988, pp. 169.

[2] A. Kibrik, "Deyksis", *Entsiklopediya Krugosvet*, Universal'naya nauchno-populyarnaya onlayn-entsiklopediya, eds. A. V. Dobrovol'skiy i dr., URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/DEKSIS.html?page=0,0

[3] A. Kibrik, "Anafora", *Entsiklopediya Krugosvet*, Universal'naya nauchno-populyarnaya onlayn-entsiklopediya, eds. A. V. Dobrovol'skiy i dr., URL: http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/ANAFORA.html

[4] Appendix B. Timex: Specific Guidelines, URL: http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_11.html#HEADING38

[5] N. Chinchor, P. Robinson, "MUC-7 Named Entity Task Definition", *Proceedings of the 7th Conference on Message Understanding*, 1997, pp. 29, URL: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html

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

[7] J. F. Allen, "An interval-based representation of temporal knowledge", *IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence.* V. 1 (24–28 August 1981, Vancouver, Canada), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, pp. 221–226.

[8] N. UzZaman, H. Llorence, L. Derczynski, M. Verhagen, J. Allen, J. Pustejovsky, "SemEval-2013 Task 1: TEMPEVAL-3: Evaluating Time Expressions, Events, and Temporal Relations", *Seventh International Workshop on Semantic Evaluation SemEval-2013*, Co-located with *Sem and NAACL. V. 2 (14–15 June 2013, Atlanta, Georgia, USA), ACL, 2013, pp. 1–9.

[9] C. Hagège, X. Tannier, "XRCE-T: XIP Temporal Module for TempEval Campaign", *Proceedings of the 4th International Workshop on Semantic Evaluations SemEval-2007* (23–24 June 2007, Prague), ACL, 2007, pp. 492–495.

© N. S. Lando, 2015
© Ailamazyan Program System Institute of RAS, 2015
© Program systems: Theory and Applications, 2015

[10] J. Strötgen, M. Gertz, “HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions”, *Proceedings of the 5th International Workshop on Semantic Evaluation SemEval-2010* (11–16 July 2010, Uppsala, Sweden), ACL, 2010, pp. 321–324.

[11] N. UzZaman, J. F. Allen, “TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text”, *Proceedings of the 5th International Workshop on Semantic Evaluation SemEval-2010* (11–16 July 2010, Uppsala, Sweden), ACL, 2010, pp. 276–283.

[12] G. Angeli, C. D. Manning, D. Jurafsky, “Parsing Time: Learning to Interpret Time Expressions”, *NAACL HLT’12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (3–8 June 2012, Montreal, Canada), Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 446–455.

[13] J. Strötgen, J. Zell, M. Gertz, “HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3”, *Seventh International Workshop on Semantic Evaluation SemEval-2013*, Co-located with *Sem and NAACL. V. 2 (14–15 June 2013, Atlanta, Georgia, USA), ACL, 2013, pp. 15–19.“Gibridnyy podkhod k obrabotke vremennoy informatsii v tekste na russkom yazyke”, *Trudy odinnadtsatoy natsional’noy konferentsii po iskusstvennomu intellektu s mezhdunarodnym uchastiyem KII-2008* (28 sentyabrya–3 oktyabrya 2008 g., g. Dubna, Rossiya), URSS, M., 2008, pp. 228–235.

Sample citation of this publication:

Natal’ya Lando. “Up-to-date methods of automatic time expression resolution in natural language texts”, *Program systems: theory and applications*, 2015, **6**:4(27), pp. 419–439. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf