ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(31), 2016, с. 177–195

УДК 004.89:004.912

Н. А. Власова, А. В. Подобряев

Извлечение сложных временных выражений из текстов в рамках задачи автоматического выявления ситуаций

Аннотация. В рамках проблемы автоматического выявления ситуаций в публицистических текстах на русском языке рассматривается задача поиска сложных временных выражений. Выделение именных групп, содержащих временные выражения, понимается как подзадача частичного синтаксического анализа (shallow parsing). Предлагается алгоритм, состоящий из предварительной сегментации и последующего поиска границ именных групп в выделенном сегменте с помощью машинного обучения (CRF-модели). Приводятся результаты экспериментов.

Ключевые слова и фразы: автоматическое извлечение информации, выделение именованных сущностей, извлечение ситуаций, выделение именных групп, временные выражения, синтаксический анализ, CRF.

Введение

Задача автоматического выявления временных выражений периодически становится объектом исследований специалистов по обработке текстов [1–5], а также обзор [6]. От качества решения этой задачи зависят многие другие — извлечение информации об отношениях и ситуациях, построение вопросно-ответных систем, автоматический анализ текстовых данных в медицине и многое другое. Поэтому рассматривать выявление временных маркеров как самодостаточную задачу нецелесообразно. Важно понимать, для каких целей понадобится информация о времени и как она должна быть представлена для дальнейшей работы.

Практически любой текст на естественном языке погружен во временной контекст. Временная информация в тексте выражается на разных уровнях, особенно в таких морфологически и лексически

Работа выполнена при финансовой поддержке РФФИ, проект 15-07-05277 А.

© Н. А. Власова, А. В. Подобряев, 2016
© Институт программных систем имени А. К. Айламазяна РАН, 2016
© Программные системы: теория и приложения, 2016

богатых языках, как русский. Это морфологический уровень (категория времени и вида у глагола, как личных форм, так и безличных — причастий, деепричастий, инфинитива), синтаксический (придаточные предложения с временными союзами и союзными словами, предложные группы), лексический уровни (слова разных частей речи, выражающие временное значение). Каждое значение отражается в элементах семантического представления текста. См. примеры ниже, в которых выделены слова, содержащие временную информацию.

Встреча президентов России и Франции состоялась ***3 июля 2014 года****.*

Правительством будут предприняты попытки решить проблему ***прошлогодней*** *задолженности за газ.*

Гости уехали, ***когда уже совсем стемнело****.*

Встречи проходят ***каждые два дня / раз в неделю / по вторникам / каждую среду четного месяца / в первые выходные зимы / в этот день / накануне Нового года****.*

До этого *он* ***семь месяцев*** *возглавлял правительство Московской области.*

В договоре ***от 2013 года*** *ничего не сказано о заморозке уровня добычи нефти.*

Пресс-служба Лукашенко сообщила ***на прошлой неделе****, что руководивший КГБ* ***четыре года*** *Зайцев отстранен* ***на время расследования самоубийства офицера, информация о котором просочилась в прессу****.*

Разработанных механизмов отражения в семантику всего многообразия временной информации пока нет. Однако задача выявления и интерпретации временной информации перед исследователями все же стоит. Эта задача частично решается в некоторых практических приложениях. Выявление в тексте временных выражений (до интерпретации) подразумевает ответы на следующие вопросы:

(*1*) четкое определение понятия временного выражения для обеспечения возможности разметки;
(*2*) определение конечной цели извлечения временных выражений;
(*3*) разработка алгоритма извлечения;
(*4*) разметка корпуса текстов для обучения и проверки алгоритма извлечения;
(*5*) оценка результатов работы алгоритма.

Исследования, в которых затрагивается тема автоматического извлечения временных выражений, обычно имеют именно такую структуру. В начале 2000-ых годов был разработан стандарт разметки TimeML, который предусматривает разметку временной информации всех типов совместно с информацией о событиях [7]. Разметка TimeML опирается на семантическое представление предложения. Однако наряду с работами в рамках TimeML есть и такие исследования, в которых обосновывается другая позиция — временное выражение нужно выделять в тексте как некоторое синтаксическое единство, поскольку именно такой подход позволит в дальнейшем автоматически интерпретировать полученные при извлечении результаты [8]. В исследованиях, опирающихся на разметку TimeML, авторы при решении практических вопросов сталкиваются с трудностями, поскольку работа с текстом происходит не на семантическом уровне, а на графическом, морфологическом, поверхностно-синтаксическом, а также с использованием частичной семантической информации. Поэтому приходится приспосабливать данные, которые получены при разметке по стандарту, к информации, доступной на этих уровнях, что приводит исследователей к поверхностно-синтаксическим структурам, включающим в себя временные маркеры, размеченные по стандарту TimeML [1], [8].

Таким образом, целесообразно решать задачу выделения в тексте временных выражений, исходя из интересов решения задачи более высокого уровня — например, автоматического извлечения информации о ситуациях. В данном аспекте интересно рассмотреть статью [9] об автоматическом извлечении временных указателей из текстов на французском языке. Здесь речь идет как раз о сложных временных выражениях, характеризующих ситуации. Приводится классификация временных выражений — они делятся на одиночные и допускающие контекст слева и (или) справа. Выделение в тексте таких выражений проводится рекурсивными правилами с опорой на маркеры временных выражений, которые задаются списком. В зависимости от частеречной принадлежности слова из правого и левого контекстов определяется, входит это слово во временное выражение или нет. К сожалению, в данной работе не приводятся результаты тестирования системы, основанной на разработанных правилах. Также не рассматриваются вопросы снятия омонимии, нет определения понятия временного выражения. В других работах, посвященных автоматическому извлечению информации о времени, исследователи также решают вопросы удобства представления извлеченных данных [2, 3, 10, 11].

Для русского языка подобных исследований, насколько нам известно, не проводилось. Есть несколько работ, в которых предпринята попытка выявления временных выражений. Однако в этих исследованиях нет четкого определения, что считается временным выражением, не рассматриваются вопросы дальнейшего использования полученной в результате извлечения информации [4], [12]. Так, например, в работе [4] говорится: «*Временным выражением (temporal expression, timex) называется выражение естественного языка, несущее временную окраску и обозначающее точку во времени, промежуток времени или периодичность некоторого события. Понятие временного выражения довольно расплывчато*». Далее приводится ссылка на формат разметки TimeML, после чего описывается эксперимент по автоматическому выявлению временных выражений. Так как в данной работе не сказано, какие именно выражения будут считаться временными в русском языке, трудно понять, что именно извлекалось из текстов.

Понятно, что задача маркирования в тексте некоторых последовательностей словоформ, имеющих временное значение, является чисто технической и не очень сложной, поскольку простые временные выражения (даты, указания на календарное время и пр.) обычно имеют жестко определенную структуру, которая легко поддается описанию. В таком случае результаты выявления временных выражений заведомо будут высокими. Для выявления оказываются применимыми любые традиционные подходы — основанные на правилах или на машинном обучении с простыми признаками, что и демонстрируется в упомянутых выше работах. Однако ценность такого маркирования невелика, поскольку практически не позволяет приблизиться к решению задачи интерпретации полученной информации в привязке к другим извлеченным из текста данным. Просто информация о времени, как мы уже указывали выше, не представляет интереса.

1. Проблемы автоматического извлечения временной информации из текстов на русском языке

При анализе текстов на русском языке исследователи сталкиваются со следующими проблемами выделения временной информации.

1.1. Проблема уровня, на котором выражается временная информация

Рассмотрим примеры:

статья, опубликованная ***ранее***;

прошлогодняя *статья*;

июльская *конференция*;

статья ***прошлого года***;

статья, написанная ***в прошлом году***;

в прошлом году *автор опубликовал эту статью.*

Видно, что информация о времени может выражаться в словах разных частей речи, разными синтаксическими конструкциями. Синтаксическая группа, выражающая временную информацию, может зависеть как от предикатного слова, так и не от предикатного слова. При этом, конечно, в каждом случае имеется в виду некоторая ситуация, погруженная во временной контекст. Например, в словосочетании «*прошлогодняя статья*» имеется в виду ситуация написания статьи и то, что эта ситуация имела место в прошлом году. Однако информация о событии в таких примерах «свернута», извлекать ее автоматически на существующем уровне развития технологий анализа текстов на русском языке не представляется возможным. При решении задачи автоматического извлечения временной информации речь обычно идет о тех временных указателях, которые характеризуют предикаты, в которых явным образом выражается информация о ситуациях.

1.2. Определения границ выражения, обозначающего время

Рассмотрим примеры:

Таким образом, поправки в устав, принятые ***накануне вступления в должность прежнего губернатора Сергея Шойгу****, отменены.*

Принятые ***накануне*** *поправки в устав были отменены.*

Как видно из приведенных примеров, при анализе текстов на русском языке важно понимать, где проходит граница временного выражения, так как от этого зависит его интерпретация. Так, в первом примере синтаксическая группа «**накануне вступления в должность прежнего губернатора Сергея Шойгу**» может быть интерпретирована через анализ ситуации «вступление в должность», а во втором примере слово «*накануне*» должно быть проинтерпретировано через анализ информации из предыдущих предложений или анализ метаинформации о тексте (дата создания документа).

1.3. Проблема омонимии при определении границ временного выражения

Приведенные ниже примеры иллюстрируют трудности, с которыми придется столкнуться при определении границ временных выражений из-за явления морфологической омонимии.

2 недели ожидания *ни к чему не привели*;

за 2 недели *разговора так и не состоялось*;

за месяц *работы были закончены*;

за месяц работы *сотрудники успели только оформить документы.*

1.4. Проблема определения предикатного слова, к которому относится синтаксическая группа, обозначающая время

Рассмотрим примеры.

Спустя несколько недель после свержения президента Виктора Януковича *он был назначен новым правительством на пост губернатора Днепропетровской области.*

Возглавлявший Московскую область ***последние полгода*** *Сергей Шойгу* ***в начале ноября*** *был назначен министром обороны вместо Анатолия Сердюкова.*

Сергей Шойгу, возглавлявший Московскую область ***последние полгода, в начале ноября*** *был назначен министром обороны вместо Анатолия Сердюкова.*

До этого *он* ***в течение трех лет*** *возглавлял Администрацию президента.*

Легко видеть, что в предложении может быть несколько синтаксических групп, выражающих временную информацию. Эти группы могут зависеть как от одного и того же слова, так и от разных. Перед разработчиками алгоритма по автоматическому извлечению временной информации стоит задача определить границы каждого временного выражения и те слова, к которым относятся выявленные временные указатели.

Становится понятным, что для решения практических задач по выявлению временной информации необходимо четко описать те временные выражения, которые подвергаются извлечению. В настоящей работе определение временного указателя подчинено задаче извлечения информации о ситуациях и исходит из возможностей системы

автоматического анализа текста ИСИДА-Т [13]. Это графематический, морфологический анализ, частичный синтаксис и информация о некоторых видах именованных сущностей.

2. Определение временного указателя

Как уже сказано выше, временные выражения, которые интересуют специалистов при решении задач по автоматическому извлечению информации, характеризуют ситуацию, которая описывается в предложении. Ситуации в тексте чаще всего задаются предикатами. Самые простые для анализа ситуации — это ситуации, информация о которых выражена глагольными формами: личной формой глагола, причастием или деепричастием. Временные характеристики предикатов такого вида выражаются с помощью обстоятельства времени.

Поэтому нецелесообразно считать временными выражениями любые последовательности словоформ, обозначающие время, например, такие группы слов, как:

(1) *статья **2015 года*** (здесь указание на год не характеризует ситуацию. Вернее, упоминания ситуации есть, но оно не задано явно: имеется в виду, что статья была написана в 2015 году);
(2) *в возрасте **менее 14 лет*** (маркер времени не относится к ситуации);
(3) ***20 лет** прошло после окончания войны* (синтаксическая группа, описывающая время, является участником ситуации. Здесь, конечно, идет речь о времени, но группа «**20 лет**» не характеризует ситуацию, описываемую в тексте).

С другой стороны, при описании и автоматическом выявлении участников ситуаций необходимо определить границы синтаксической группы, представляющей участника ситуации. Такая группа, описывающая временной аспект ситуации, соответствует некоторому поддереву синтаксического дерева предложения. Вершина этого поддерева зависит от предикатного слова, обозначающего ситуацию.

Это якобы было известно администрации Обамы ⟨за несколько дней до описываемых событий⟩.

Необходимо предупредить участников поездки ⟨за несколько дней⟩ из-за необходимости взять нужные справки.

В настоящей работе было принято решение ограничиться только ситуациями, которые заданы в тексте глаголом в личной форме, причастием и деепричастием. Предполагается, что такое определение

временных указателей позволит впоследствии подойти к автоматическому извлечению информации о временных отношениях, связывающих различные ситуации, упоминаемые в тексте.

Итак, договоримся считать *временным указателем* синтаксическую группу с временным значением, вершина которой непосредственно зависит от предиката, описывающего ситуацию. Временной указатель в предложении играет роль сирконстанта.

3. Классификация временных указателей

Временные указатели могут иметь самую разнообразную структуру. Мы разделили все множество временных указателей на следующие группы.

(*1*) Даты с точным указанием числа, месяца, года, столетия и т.п, в том числе с предлогом:

23 февраля 2016 года, в XX-ом веке, в 1992 году.

(*2*) Синтаксические группы с вершиной, выраженной наречием или именем существительным с временным значением с зависимыми словами без предлога:

долго, рано, поздно, утром, днем, ночью, рано утром, поздно вечером, ранним утром, этой ночью, той же ночью и т.п.

(*3*) Предложные именные группы с зависимым существительным — безусловным указателем на время (и с зависимыми от этого существительного словами):

в этот понедельник, в апреле, (назначено) на вторник, до следующего сентября, за три часа, в три часа утра, в прошлом году, в начале двадцатого века, в конце текущего месяца, в начале года, в середине столетия, осенью этого года.

(*4*) Именные группы с предлогом — безусловным указателем на время:

в течение, во время, в продолжение, в ходе, в период, по истечении, на протяжении, накануне, спустя.

(*5*) Именные группы, обозначающие указание на время, с предлогом, не являющимся однозначным указателем на время, и с существительным, которое тоже не является однозначным временным указателем:

после обеда, до праздника, за прогулку и т.п.

Эти предлоги могут выражать и не временные отношения, ср. *дойти до угла дома, спрятаться за шторой.*

В некоторых случаях такие группы можно однозначно интерпретировать как временные указатели, поскольку они являются подгруппой внутри составного временного указателя, например:

⟨*Через три недели после отставки с поста министра*⟩ *Зайцев заявил, что...*

В данном случае группа ⟨*после отставки с поста министра*⟩ входит в состав временного указателя ⟨*через три недели после отставки с поста министра*⟩. Ср.

⟨*После отставки с поста министра*⟩ *Зайцев заявил, что...*

С временными указателями этой группы автоматическая обработка будет возможна только в том случае, если в наличии будет онтология, в которой для каждого существительного будет указано, может ли оно обозначать событие. В настоящей работе такие группы рассматриваться не будут.

Отметим, что временные указатели разных типов могут в тексте следовать друг за другом, образуя сложный временной указатель. Например:

⟨*рано утром во вторник*⟩;

⟨*вечером 21 ноября*⟩;

⟨*летом 2013 года*⟩;

⟨*с момента его основания в 2009 году*⟩;

⟨*в период выборов в Государственную Думу в 2011 году*⟩;

⟨*в единый день голосования 8 сентября*⟩;

⟨*25 октября поздно вечером*⟩.

4. Проблемы автоматического выявления временных указателей разных типов

Временные указатели первых трех типов легко исчислить и извлечь из текста — они компактны и перечислимы. Заметим при этом, что если не рассматривать вопрос о том, к чему относится временное выражение, то полнота и точность получаются хорошие [**4**, **5**, **12**].

Что касается групп 4 и 5, то это необозримое множество. Именно поэтому правилами и машинным обучением на простых признаках выявить такие временные выражения с хорошим результатом по полноте и точности не получится. Здесь нужен либо полный синтаксический анализ, либо другие решения.

В случае группы 5 необходимо знать, обозначает ли слово, зависящее от предлога, событие. Если да, то это временной указатель. Поэтому при наличии достаточно полной онтологии можно рассматривать и такие указатели. Другой вариант — выявление временных показателей с ситуациями определенного типа, которые задаются списком. В настоящей работе принято решение такие временные указатели (пятый тип) не рассматривать.

Вопрос об автоматическом извлечении временных указателей третьего типа из текстов на русском языке, насколько нам известно, не ставился. В существующих исследованиях речь идет исключительно о регулярных выражениях. Однако по причинам, указанным в разделе 2, для качественного извлечения информации о событиях необходимо полностью выделять синтаксическую группу указателя на время.

5. Постановка задачи и описание алгоритма

Практическая задача, которая решается в настоящей работе — автоматическое извлечение временных указателей без интерпретации и без построения отношения между временным указателем и ситуацией. Также ставится задача оценить количество временных указателей разных типов в текстах новостного содержания.

Работа проводится в рамках решения более глобальной задачи — выделения границ именных групп в русскоязычных текстах новостного содержания [14]. Метод выделения именных групп основан на машинном обучении с использованием нетривиальных признаков, основанных в частности на информации из базы знаний. Кроме того, обучение проводится не на предложениях, а на специально выделяемых фрагментах, содержащих именные группы. С помощью алгоритма внутри фрагментов проводятся границы между именными группами.

Что касается выявления временных указателей, то предлагается использовать комбинированный подход. Временные указатели первых трех типов, представляющие из себя регулярные выражения, выявляются в текстах с помощью правил на специальном языке в системе автоматического анализа текстов ИСИДА-Т [13].

Временные указатели третьего типа обязательно содержат опорное слово — предлог или предложную конструкцию («*в течение*», «*в продолжение*», «*спустя*», «*накануне*» и т.д.) Пометив опорные слова, мы можем выбрать фрагменты, которые заведомо содержат временные указатели.

В итоге на вход алгоритму, основанному на машинном обучении поступает размеченная коллекция фрагментов. Наша задача — определить границы синтаксической группы, составляющей временное выражение. В процессе анализа должны отсеяться те временные выражения, которые были выявлены правилами, но при этом не являются временными указателями в том смысле, в котором мы их определили в настоящей работе.

Алгоритм, основанный на машинном обучении, использует следующие признаки:

(*1*) простые графематические: информация о типе токена (число или слово, латиница или кириллица, с большой или с маленькой буквы, вхождение в выражение в кавычках);
(*2*) простые морфологические:
- частеречный тег для каждой словоформы внутри фрагмента;
- падежная характеристика каждой словоформы (без снятия падежной омонимии);
- значение словоизменительной категории «число» для каждой словоформы во фрагменте (без снятия омонимии);
- принадлежность слова к определенному выделенному классу словоформ (помимо частеречных тегов). Это классы личных, относительных, указательных и определительных местоимений;

(*3*) информация из базы знаний после обработки текста:
- вхождение в состав словосочетания, выделенного как именованная сущность определенного типа («имя лица», «название организации», «геополитическая единица», «указание на время»);
- информация о слове из базы знаний (связь слова с концептом в базе знаний системы ИСИДА-Т);

(*4*) комбинированные признаки:
- длина фрагмента (число словоформ, входящих во фрагмент, при этом уже выделенные именованные сущности считаются за одну словоформу);
- последовательность частеречных тегов внутри фрагмента;
- типы границ фрагмента (начало предложения, конец предложения, предикатное слово, знак препинания);
- последовательность выявленных именованных сущностей и концептов внутри фрагмента с учетом количества слов между ними, не связанных с концептами в базе знаний;
- порядковый номер словосочетания (словоформы), представляющего именованную сущность определенного вида.

6. Некоторые статистические наблюдения

В результате разметки можно было провести некоторые статистические наблюдения над содержащимися в новостных текстах на русском языке конструкциями, описывающими время. Прежде всего необходимо отметить, что доля синтаксических групп, содержащих опорные слова, которые указывают на время, но при этом не являются временными указателями, составляет примерно 27% от общего количества синтаксических групп. Это, например, такие выражения, как: «*тюремный срок*», «*в возрасте 25 лет*», «*5 лет лишения свободы*», «*итоги 2008 года*», «*охрана окружающей среды*», «*учебный год*», «*рабочее время*», «*благодарность за годы работы*», «*повестка дня*», «*по состоянию на сегодняшний день*», «*аванс за январь*» и другие конструкции подобного рода. А также сложные конструкции типа:

Новый глава Рособрнадзора на несколько месяцев старше своего коллеги из Минкомсвязи Николая Никифорова.

Стратегия развития «Почты России» до 2023 года сейчас находится на рассмотрении в Министерстве связи.

Речь идет о нецелевом использовании средств в период выборов в Государственную Думу в 2011 году.

Отметим, что именно такие конструкции ошибочно ловятся системами выявления временных выражений, основанными на правилах, а также на машинном обучении с простыми графематическими и морфологическими признаками. Ср. обсуждение подобных примеров в [4]. Из оставшихся примерно 12 тысяч временных указателей около 80% составляют указатели первых трех типов (см. классификацию в разделе 3). То есть указатели 4-го типа — сложные именные группы, содержащие указание на время, составляют примерно одну пятую часть всех временных указателей. Итак, алгоритм, выявляющий границы временных указателей, должен правильно определить, где проходят границы именных групп внутри фрагментов: с одной стороны, отделить те слова и конструкции, которые могут быть временными указателями, но в данном конкретном случае не характеризуют ситуацию, описываемую глаголом, причастием или деепричастием, а с другой стороны, правильно выявить границу временного указателя.

7. Результаты и обсуждение

7.1. Метод обучения

Рассмотрим последовательность слов выделенного фрагмента, содержащего временное выражение. Перед нами стоит задача про-

ведения границ внутри данного фрагмента. Рассмотрим эту задачу, как задачу сегментации (разметки) последовательности с зависимостями между наблюдаемыми переменными (словами). В то же время достаточно далекие друг от друга слова можно считать независимыми.

Такое явление описывает линейная дискриминативная графовая модель CRF [15]. Напомним некоторые определения. Пусть имеется граф, в вершинах $v_1, \ldots, v_n$ которого определены случайные величины $\xi = (\xi_1, \ldots, \xi_n)$, принимающие значения в множествах $X_1, \ldots, X_n$. Пусть граф и эти случайные величины обладают марковским свойством, т. е.

$$P(\xi_k = x_k \mid \xi_i = x_i, i \neq k) = P(\xi_k = x_k \mid \xi_i = x_i, v_i \in O(v_k)),$$

где $O(v_k)$ — множество вершин графа, соседних с v_k. По теореме Хаммерсли–Клиффорда [16] распределение случайной величины ξ является распределением Гиббса, т. е.

$$P(\xi = x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c), \qquad Z = \sum_{x \in X} \prod_{c \in C} \psi_c(x_c),$$

где $x = (x_1, \ldots, x_n) \in X = X_1 \times \cdots \times X_n$, C — множество клик графа (т. е. полных подграфов), далее если $c = \{v_{i_1}, \ldots, v_{i_{|c|}}\}$ — клика, то $x_c = (x_{i_1}, \ldots, x_{i_{|c|}})$, а функция $\psi_c : X_{i_1} \times \cdots \times X_{i_{|c|}} \to \mathbb{R}_+$.

Если множество вершин графа разбито на две части $V = X \sqcup Y$, наблюдаемую X и скрытую Y, то задача состоит в том, чтобы по наблюдаемому набору значений x восстановить скрытые значения y, т. е. найти условное распределение

$$p(y \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(x_c, y_c), \qquad Z(x) = \sum_{y'} \prod_{c \in C} \psi_c(x_c, y'_c).$$

Вид функций ψ_c фиксирован — это экспоненты от линейных комбинаций признаков. Коэффициенты λ этих линейных комбинаций подбираются в процессе обучения, чтобы на обучающем множестве $\{(x^i, y^i) \mid i = 1, \ldots, N\}$ максимизировать логарифм функции правдоподобия:

$$\sum_{i=1}^{N} \log p(y^i \mid x^i) - \frac{1}{2\sigma^2} ||\lambda||^2,$$

где вычитаемое — регуляризационный член (нужный для того, чтобы избежать переобучения), а параметр σ определятся экспериментально.

Мы используем линейную модель CRF, с максимальными кликами третьего порядка, то есть каждая метка y соединена с соседними и с двумя наблюдаемыми переменными (словами) справа и слева.

Таблица 1. Результаты тестирования

Точность	Полнота	F-мера
88.97%	94.11%	91.46%

7.2. Результаты

Для обучения и тестирования алгоритма были размечена коллекция из 2000 текстов новостного содержания. С помощью системы правил на языке PSL в системе обработки текстов ИСИДА-Т в текстах были автоматически выделены опорные слова и выражения, содержащие временную информацию. Кроме того, все тексты были разбиты на фрагменты, заведомо содержащие синтаксические группы временных указателей (о методе разбиения текста на фрагменты см. [**14**]). В итоге было получено 15472 фрагмента, содержащих временную информацию. Обучающее множество состояло из 1500 текстов (10703 фрагмента, 9971 граница фрагментов), а тестовое из 500 текстов (4769 фрагментов, 4143 границы фрагментов). Результаты, достигнутые на тестовом множестве, представлены в таблице 1.

Таким образом, можно констатировать, что разработанная система признаков и алгоритм на базе графовой вероятностной модели позволяют получить хороший результат при выявлении временных указателей, характеризующих ситуации, выраженные глагольными формами. Комбинация признаков, использующих информацию разных языковых уровней, включая сведения из базы знаний, т. е. доступную для автоматического анализа частичную семантическую информацию, позволяет выявить границы временных указателей с хорошими показателями точности и полноты. Основные ошибки при выявлении временных указателей возникают при анализе таких предложений, как

СМИ сообщили об отставке Пихоя с поста гендиректора ***в ноябре прошлого года****.*

В данном примере выражение «*в ноябре прошлого года*» в принципе может относиться и к глаголу «*сообщили*», и к существительному «*отставке*». Так что здесь есть двусмысленность в исходном предложении. Разработанный алгоритм выделяет такую конструкцию как временной указатель.

Заключение

В настоящей работе было определено понятие временного указателя с целью выявить автоматически в тексте конструкции, которые выражают временную информацию и характеризуют ситуации, выраженные глагольными формами. Приведенный алгоритм показывает достаточно хороший результат. Таким образом, можно использовать результаты выявления для дальнейшей работы — автоматической привязке временного указателя к предикатному слову, что также является нетривиальной задачей в рамках систем, использующих частичный синтаксический анализ. Однако решение этой задачи позволит подойти к интерпретации извлекаемой из текстов информации. С другой стороны, выделение границ сложных временных выражений с использованием системы разноуровневых признаков показывает эффективность такого подхода для решения задач сегментации текстов, этот поход может быть применен для выделения синтаксических групп разных типов.

Список литературы

[1] V. Moriceau, X. Tannier. “French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime”, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC’14 (26–31 May, 2014 , Reykjavik, Iceland), ELRA, 2014, URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/45_Paper.pdf ↑177,179

[2] T. A. Miller, S. Bethard, D. Dligach, Ch. Lin, G. K. Savova. “Extracting Time Expressions from Clinical Text”, *Proceedings of the Workshop on Biomedical Natural Language Processing*, BioNLP’15 (July 30, 2015, Beijing, China), 2015, pp. 81—91, URL: http://www.aclweb.org/anthology/W15-3809 ↑ 177,179

[3] P. Jindal, D. Roth. “Extraction of Events and Temporal Expressions from Clinical Narratives”, *Journal of Biomedical Informatics*, **46**, suppl. (December 2013), pp. S13–S19, URL: http://sharps.org/wp-content/uploads/JINDAL-JBI.pdf ↑ 177,179

[4] А. А. Романенко. *Применение условных случайных полей в задачах обработки текстов на естественном языке*, Выпускная квалификационная работа магистра, МФТИ, М., 2014, URL: http://www.machinelearning.ru/wiki/images/f/fc/Romanenko2014Application.pdf ↑ 177,180,185,188

[5] Н. А. Власова. «Об одной проблеме автоматического извлечения временной информации из русскоязычных текстов», *Программные системы: теория и приложения*, **5**:4(22) (2014), с. 231–242, URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf ↑ 177,185

[6] 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 ↑ 177

[7] J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, I. Mani. "The Specification Language TimeML", *The Language of Time: A Reader Mani*, eds. J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, pp. 545–557. ↑ 179

[8] B. Boguraev, R. K. Ando. *TimeML-Compliant Text Analysis for Temporal Reasoning*, IBM, 2005, URL: http://riejohnson.com/rie/timeml-ijcai05.pdf ↑ 179

[9] N. Vazov. "A System for Extraction of Temporal Expressions from French Texts", TALN'01 (July 2–5, 2001, Tours, France), pp. 315–324, URL: http://tln.li.univ-tours.fr/Tln_Colloques/TALN2001-RECITAL2001/Actes/tome1_PDF/partie2_p30_322/art29_p313_322.pdf ↑ 179

[10] F. Schilder, Ch. Habel. "From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages", *Proceedings of the ACL Workshop on Temporal and Spatial Information Processing*, ACL'01 (July 9–11, 2001, Toulose, France), pp. 65–72, URL: http://modul.mercubuana.ac.id/files/openjournal/OpenJournalOfTechnology/text/W01-1309.pdf ↑ 179

[11] A. X. Chang, Ch. D. Manning. "SUTIME: A Library for Recognizing and Normalizing Time Expressions", LREC'12 (May 21–27, 2012, Istanbul, Turkey), 2012, pp. 3735–3740, URL: http://nlp.stanford.edu/pubs/lrec2012-sutime.pdf ↑ 179

[12] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. «Conditional random field in segmentation and noun phrase inclination tasks for Russian» (Бекасово, 4 — 8 июня 2014 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **13 (20)**, По материалам ежегодной Международной конференции «Диалог», Изд-во РГГУ, М., 2014, с. 297, 10 с., URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/KudinovMS.pdf ↑ 180,185

[13] Д. А. Александровский, Д. А. Кормалев, М. С. Кормалева, Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. «Развитие средств аналитической обработки текста в системе ИСИДА-Т», *Труды Десятой национальной конференции по искусственному интеллекту с международным участием КИИ'2006*. Т. 2 (25–28 сентября, Обнинск), Физматлит, М., 2006, с. 555–563, URL: http://www.raai.org/resurs/papers/kii-2006/doklad/Alexandrovsky.doc ↑ 183,186

[14] Н. А. Власова, А. В. Подобряев. «К вопросу об определении границ именных групп при решении задач автоматического извлечения информации из текстов на русском языке», *Программные системы: теория и приложения*, **7**:1(28) (2016), с. 153–170, URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf ↑ 186,190

[15] Ch. Sutton, A. McCallum. “An Introduction to Conditional Random Fields”, *Foundations and Trends in Machine Learning*, **4**:4 (2011), pp. 267–373. ↑ 189

[16] J. M. Hammersley, P. Clifford.. *Markov fields on finite graphs and lattices*, 1971, URL: http://www.recognition.mccme.ru/pub/papers/CRF/hammersley71markov.pdf ↑ 189

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

Пример ссылки на эту публикацию:

Н. А. Власова, А. В. Подобряев. «Извлечение сложных временных выражений из текстов в рамках задачи автоматического выявления ситуаций», *Программные системы: теория и приложения*, 2016, **7**:4(31), с. 177–195.
URL: http://psta.psiras.ru/read/psta2016_4_177-195.pdf

Об авторах:

Наталья Александровна Власова

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

e-mail: nathalie.vlassova@gmail.com

Алексей Владимирович Подобряев

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

e-mail: alex@alex.botik.com

Natalia Vlasova, Alexey Podobryaev. *Complex time expressions recognition problem in application to automatic information extraction from Russian texts.*

Abstract. We consider the problem of complex time expressions recognition in Russian news texts with application to automatic information extraction. We describe an algorithm for finding noun phrases that contain time expressions. This algorithm has two parts: the pre-segmentation and the selection of noun phrase borders inside the segments via machine learning (CRF-model). We receive results of experiments. (*In Russian*).

Key words and phrases: information extraction, named entities recognition, noun phrase chunking, time expressions, CRF.

References

[1] V. Moriceau, X. Tannier. "French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC'14 (26–31 May, 2014 , Reykjavik, Iceland), ELRA, 2014, URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/45_Paper.pdf

[2] T. A. Miller, S. Bethard, D. Dligach, Ch. Lin, G. K. Savova. "Extracting Time Expressions from Clinical Text", *Proceedings of the Workshop on Biomedical Natural Language Processing*, BioNLP'15 (July 30, 2015, Beijing, China), 2015, pp. 81—91, URL: http://www.aclweb.org/anthology/W15-3809

[3] P. Jindal, D. Roth. "Extraction of Events and Temporal Expressions from Clinical Narratives", *Journal of Biomedical Informatics*, **46**, suppl. (December 2013), pp. S13–S19, URL: http://sharps.org/wp-content/uploads/JINDAL-JBI.pdf

[4] A. A. Romanenko. *Application of CRF to natural language processing*, Master's thesis, MIPT, M., 2014 (in Russian), URL: http://www.machinelearning.ru/wiki/images/f/fc/Romanenko2014Application.pdf

[5] N. A. Vlasova. "On one problem of automatic information extraction from Russian texts", *Program systems: theory and applications*, **5**:4(22) (2014), pp. 231–242 (in Russian), URL: http://psta.psiras.ru/read/psta2014_4_231-242.pdf

[6] N. S. Lando. "Up-to-date methods of automatic time expression resolution in natural language texts", *Program systems: theory and applications*, **6**:4(27) (2015), pp. 419–439 (in Russian), URL: http://psta.psiras.ru/read/psta2015_4_419-439.pdf

[7] J. Pustejovsky, B. Ingria, R. Sauri, J. Castano, J. Littman, R. Gaizauskas, A. Setzer, G. Katz, I. Mani. "The Specification Language TimeML", *The Language of Time: A Reader Mani*, eds. J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, pp. 545–557.

[8] B. Boguraev, R. K. Ando. *TimeML-Compliant Text Analysis for Temporal Reasoning*, IBM, 2005, URL: http://riejohnson.com/rie/timeml-ijcai05.pdf

[9] N. Vazov. "A System for Extraction of Temporal Expressions from French Texts", TALN'01 (July 2–5, 2001, Tours, France), pp. 315–324, URL: http://tln.li.univ-tours.fr/Tln_Colloques/TALN2001-RECITAL2001/Actes/tome1_PDF/partie2_p30_322/art29_p313_322.pdf

© N. A. Vlasova, A. V. Podobryaev, 2016
© Ailamazyan Program Systems Institute of RAS, 2016
© Program systems: Theory and Applications, 2016

[10] F. Schilder, Ch. Habel. “From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages”, *Proceedings of the ACL Workshop on Temporal and Spatial Information Processing*, ACL’01 (July 9–11, 2001, Toulose, France), pp. 65–72, URL: http://modul.mercubuana.ac.id/files/openjournal/OpenJournalOfTechnology/text/W01-1309.pdf

[11] A. X. Chang, Ch. D. Manning. “SUTIME: A Library for Recognizing and Normalizing Time Expressions”, LREC’12 (May 21–27, 2012, Istanbul, Turkey), 2012, pp. 3735–3740, URL: http://nlp.stanford.edu/pubs/lrec2012-sutime.pdf

[12] M. S. Kudinov, A. A. Romanenko, I. I. Piontkovskaja. “Conditional random field in segmentation and noun phrase inclination tasks for Russian” (June 4–8, 2014, Bekasovo, Russia), Computational Linguistics and Intellectual Technologies, vol. **13 (20)**, Papers from the Annual conference “Dialogue”, RGGU, M., 2014, pp. 297, 10 p., URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/KudinovMS.pdf

[13] D. A. Aleksandrovkiy, D. A. Kormalev, M. S. Kormaleva, E. P. Kurshev, E. A. Suleymanova, I. V. Trofimov. “Development of the ISIDA-T system’s tools for analytic text processing”, *Proceedings of the X national conference of artificial intelligence KII’2006*. V. 2 (September 25–28, 2006, Obninsk, Russia), Fizmatlit, M., 2006, pp. 555–563 (in Russian).

[14] N. A. Vlasova, A. V. Podobryaev. “To the noun phrase recognition problem in application to automatic information extraction from Russian texts”, *Program systems: theory and applications*, **7**:1(28) (2016), pp. 153–170 (in Russian), URL: http://psta.psiras.ru/read/psta2016_1_153-170.pdf

[15] Ch. Sutton, A. McCallum. “An Introduction to Conditional Random Fields”, *Foundations and Trends in Machine Learning*, **4**:4 (2011), pp. 267–373.

[16] J. M. Hammersley, P. Clifford.. *Markov fields on finite graphs and lattices*, 1971, URL: http://www.recognition.mccme.ru/pub/papers/CRF/hammersley71markov.pdf

Sample citation of this publication:

Natalia Vlasova, Alexey Podobryaev. “Complex time expressions recognition problem in application to automatic information extraction from Russian texts”, *Program systems: Theory and applications*, 2016, **7**:4(31), pp. 177–195. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2016_4_177-195.pdf