ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 4(22), 2014, с. 231–242

УДК 004.89:004.912

Н. А. Власова

Об одной проблеме автоматического извлечения временной информации из русскоязычных текстов

Аннотация. В настоящей работе рассматривается задача сопоставления временной информации событиям назначения и отставки лиц. Предлагается система правил для автоматического установления такого соответствия. Выполнено тестирование на размеченной коллекции новостных текстов на русском языке.

Ключевые слова и фразы: автоматическое извлечение информации, темпоральные выражения, события, временной аспект, текстовые коллекции.

Введение

В последнее время продолжают совершенствоваться алгоритмы автоматического извлечения информации из текстов. В процессе решения этой задачи среди исследователей и разработчиков выработался подход — поэтапное извлечение информации [1]. Это метод извлечения информации от простого к сложному — от выявления именованных сущностей к извлечению отношений между ними, ситуаций, связывающих сущности между собой, и, наконец, до интерпретации выявленной и извлеченной информации. В дальнейшем осуществляется переход к отождествлению извлеченных объектов, вывод информации с помощью онтологий, построение смыслового образа документа. Наиболее проработаны первые этапы — выявление сущностей, отношений, ситуаций. Эти модули анализа текстов имеют наиболее высокие показатели выявления и извлечения. Работы ведутся и над более сложными последующими этапами. Одной из подзадач автоматического извлечения информации из текстов является извлечение временной информации. Выявление и интерпретация

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

© Н. А. Власова, 2014
© Институт программных систем имени А. К. Айламазяна РАН, 2014
© Программные системы: теория и приложения, 2014

временных указателей — это этап, который вплотную подводит исследователей к интерпретации извлеченной информации.

1. О задаче автоматического извлечения временной информации

Извлечение временной информации из текста, в отличие от простого выявления временных сущностей, является подзадачей автоматического извлечения информации, причем довольно высокого уровня сложности. Ведь само по себе выявление временной сущности, в отличие, например, от упоминаний персон, организаций, геополитических единиц, не несет в себе никакой смысловой нагрузки и не может рассматриваться как самостоятельная цель извлечения информации. Временной указатель должен быть обязательно каким-то образом проинтерпретирован и отнесен к некоторой ситуации (ситуациям), а для этого необходимо, в свою очередь, чтобы на предварительных этапах анализа текста были выявлены и некоторым образом проинтерпретированы ситуации.

Время и временные указатели давно стали объектом внимания исследователей. Для английского языка были разработаны стандарты разметки TimeML [2], которые позволяют фиксировать в тексте временную информацию и снабжать выявленные указатели атрибутами, то есть выполнять некоторую интерпретацию временных выражений. Естественно, такая интерпретация невозможна без привязки временного указателя к ситуации. Поэтому в стандарте TimeML предусматривается разметка ситуаций и отношений между ситуациями и временными указателями. Для некоторых европейских языков были предприняты практические попытки извлекать информацию о времени [3].

Для русского языка подход к извлечению информации о времени разрабатывался в проекте OntosMiner [4]. Однако сейчас активные работы в рамках этого проекта не ведутся. Существуют некоторые работы, в которых рассматриваются проблемы извлечения временных выражений, однако практические решения с количественной оценкой результатов не приводятся [5].

Система автоматического извлечения информации из текстов, претендующая на выявление и интерпретацию простейших указаний на время в рамках одного предложения (самостоятельные слова и отдельные словосочетания, обозначающие время), предполагает качественное решение следующих подзадач:

(1) выявление границ временных указателей;
(2) интерпретация временных указателей хотя бы на простейшем уровне (предполагается наличие модели времени);
(3) выявление и извлечение информации о целевых фактах, для которых предполагается интерпретация времени;
(4) алгоритм связывания выявленных временных указателей и целевых фактов, окончательная интерпретация временных указателей в контексте.

Это необходимые этапы анализа, без реализации которых интерпретация временных указателей невозможна.

Допустим даже, что ситуации как-то помечены. Самое главное — связать ситуацию с временным указателем, особенно в системах с частичным синтаксическим анализом, какими является большинство современных систем автоматического извлечения информации, кроме систем ЭТАП [6] и Abbyy Compreno [7], ориентированных на перевод, а не на извлечение информации.

Для русского языка есть только попытка решить задачу выявления и некоторой классификации временных указателей [8]. Кроме того, в диссертации того же автора речь идет о разработке общего алгоритма автоматического временного анализа извлекаемых из дискурса ситуаций [9]. Однако в данной диссертации акцент сделан на выявлении временных отношений между ситуациями. Проблема сопоставления упоминаний о ситуациях и временных указателей не затрагивается.

Однако, как уже показано выше, как раз проблема соотнесения указания на время и упоминания ситуации, является одним из важных промежуточных этапов анализа текста. В настоящей работе рассматривается один из способов решения этой проблемы.

2. Постановка задачи

В системах автоматического извлечения информации из текстов, основанных на частичном синтаксическом анализе (таких систем в настоящее время большинство), проблема правильного связывания временного указателя с предикатным словом стоит очень остро. Установить автоматически время совершения события — непростая задача, даже если ограничиться рамками одного предложения и не принимать во внимание, что время может быть как-то вычислено

по более широкому контексту. В границах одного предложения тоже не все определяется однозначно. Обычно временной указатель относится к предикатному слову (чаще всего это форма глагола, например,«директор *уволился 25 ноября*», «Иван Родионов, *назначенный* на пост директора *неделю назад*»), однако встречаются и указания на время, относящиеся к существительному (указ от *15 октября*, революция *1917 года*, чемпионат мира по футболу *2018 года* и т.п.) Кроме того, предложения, в которых предстоит соотнести временной указатель и упоминание события, также могут быть устроены сложно. Приведем соответствующие примеры.

(1) В предложении два или несколько предикатных слов и один временной указатель.

> Новый премьер и состав кабинета министров будут *назначены* не раньше *14 декабря*, *заявил* спикер Верховной рады Владимир Литвин.

(2) В них два или несколько предикатных слов и несколько временных указателей.

> На должность руководителя департамента соцзащиты населения с *1 ноября 2013 года* *назначен* Александр Алисиевич, который *до настоящего времени* *возглавлял* комитет труда и занятости населения Новгородской области.

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

(3) Временной указатель синтаксически и по смыслу относится не к предикатному слову.

> Для подготовки к проведению Чемпионата мира по футболу *2018 года* на должность директора строительной корпорации был *назначен* Евгений Маслов.

В настоящей статье мы ограничиваемся рассмотрением только временных указателей, которые относятся к упоминаниям событий отставки или назначения в одном и том же предложении. Все другие способы указания на время ситуации не рассматриваются. Маркером ситуаций отставки или назначения могут быть только словоформы лексем «*назначить*» или «*уволить*».

Таким образом, будут рассматриваться конструкции типа «*предикатное слово* + *временной указатель*». Между выявленной временной сущностью и предикатным словом, описывающим ситуацию, необходимо установить отношение «`$время_события`».

Исследование проводилось в рамках системы автоматического извлечения информации из текстов ИСИДА-Т [10].

3. Об извлечении информации в системе ИСИДА-Т

В системе автоматического извлечения информации ИСИДА-Т есть независимо работающие модули выявления временных сущностей и извлечения информации о ситуациях отставки и назначения.

Методы извлечения информации основаны на правилах, которые пишутся экспертом на диалекте языка CPSL. Система предполагает поэтапный подход к анализу текста — от разбиения на предложения и слова к анализу морфологии, построению частичного синтаксиса, а далее — к извлечению именованных сущностей, отношений между ними. Также возможно извлечение информации о некоторых типах событий. Метод извлечения информации о событиях подробно описан в работе [11]. Модуль выявления временных сущностей устроен очень просто. В отдельном блоке правил фактически задается множество шаблонов, в которые могут укладываться самые частотные временные выражения. Временные указатели, как и упоминания о событиях помечаются аннотациями, в атрибуты которых записывается информация о данных объектах извлечения.

Для возможной простейшей интерпретации временных указателей каждая аннотация, соответствующая текстовому фрагменту, снабжается специальным атрибутом. Система таких атрибутов задает элементарную лингвистическую модель времени. В системе ИСИДА-Т это деление временных указателей на точки, интервалы и периоды.

Примеры временных указателей со значением «*точки*»: *12 января 2013 года*, *сегодня*, *накануне*, *на этой неделе*, *через месяц* и т.п.

Примеры временных указателей со значением «*интервалы*»: *в течение двух месяцев*, *три года*, *несколько лет* и т.п.

Примеры временных указателей со значением «*периоды*»: *по понедельникам*, *каждый месяц*, *раз в 10 лет* и т.п.

Такой простой способ выявления временных указателей позволяет получить довольно высокие показатели по точности (94.88%) и

полноте (97.85%). В силу специфики устройства временных выражений они редко омонимичны невременным конструкциям. Кроме того, способов выражения информации о времени в языке, в принципе, ограниченное количество, поэтому такой способ выявления временных указателей вполне себя оправдывает.

Анализ новостных текстов показал, что в рамках одного предложения с упоминанием ситуации отставки или назначения возможны следующие комбинации глаголов, описывающих событие, и временных указателей.

(1) В предложении нет указания на время.

> В Воронеже участковый *уволен* за получение взятки.

> Министр обороны Сергей Шойгу *назначил* своим советником бывшую телеведущую Марию Китаеву.

(2) В предложении один глагол в личной форме (и это форма глагола «*назначить*» или «*уволить*») и одно указание на время.

> *Летом 2012 года* она была *назначена* на должность советника губернатора Московской области по информационной политике.

(3) В предложении несколько глаголов в личной форме или глагольных форм (одна из этих форм — форма глагола «*назначить*» или «*уволить*»), к которым может синтаксически относиться временной указатель. В предложении одно предикатное слово (назначение или отставка) и один временной указатель.

> *Вчера* кабинет министров Украины *уволил* Владимира Козака, который *занял* пост министра инфраструктуры, с должности гендиректора государственной администрации железнодорожного транспорта Украины, *говорится* в соответствующем постановлении правительства.

(4) В предложении несколько глаголов в личной форме или глагольных форм (одна из этих форм — форма глагола «*назначить*» или «*уволить*»), к которым может синтаксически относиться временной указатель. В предложении одно предикатное слово (назначение или отставка) и два или несколько временных указателя.

> Калинин был *назначен* главой СБУ *3 февраля 2012 г.*, а *до этого* он *в течение трех лет занимал* должность начальника Управления государственной охраны Украины.

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

Проблема с соотнесением временного указателя и упоминания ситуации актуальна для случаев (3) и (4). Мы предполагаем, что указания на время, которые могут относиться к существительным, не подходят для описания событий отставки и назначения. Назначение или отставка — события точечного типа, значит, они могут быть соотнесены только с временными указателями с атрибутом «*точка*».

4. Исследование фактического материала, предварительные выводы

При отсутствии полного синтаксического анализа предложения необходимо выработать некоторые правила, которые будут определять, соотносится ли данное указание на время с упоминанием ситуации заданного типа. Эти правила применяются ко всем глагольным формам предикатов данного предложения.

В ходе исследования фактического материала было выработано несколько таких правил:

Правило 1

Указание на время должно иметь атрибут «*точка*».

Это правило основано на том, что такой тип событий, как назначение или отставка — точечный, поэтому не может характеризоваться указанием на время «интервального» или «периодического» типов.

> Президент Владимир Путин *14 января назначил* нынешнего главу республики Коми Вячеслава Гайзера временным руководителем региона.

14 января — временной указатель с атрибутом «*точка*». Строится отношение «`$время_события`».

> Тренера клуба НХЛ *уволили* после *16 лет* работы.

16 лет — временной указатель с атрибутом «*интервал*». Отношения нет.

Все остальные правила касаются только подходящих по семантике временных указателей (с атрибутом «*точка*»).

Правило 2

Указание на время не может быть отделено одной запятой или сочинительным союзом «и» от упоминания ситуации.

> Аваков *уволил* Луцюка после того, как *накануне* в результате беспорядков и пожара в Доме профсоюзов на Куликовом поле Одессы погибли несколько десятков человек.

> Шарипов сделал заявление, из-за которого его *уволили*, *20 ноября 2013 года*.

Временной указатель «*20 ноября 2013 года*» относится к глаголу «*сделал*».

При этом две запятые между указанием на время и упоминанием ситуации допустимы.

> *В то же время*, как сообщает Интерфакс, исполняющий обязанности президента Украины Александр Турчинов *уволил* главу Службы безопасности Украины по Донецкой области Валерия Иванова.

Правило 3

Через две запятые можно «перешагнуть», если это придаточное со словом «*который*», а также, если это причастный или деепричастный оборот или после второй запятой нет больше предикатного слова, к которому можно было бы отнести временной указатель.

> *В 2011 году* по рекомендации главы сельской администрации Любови Валеевой, с которой у него наладились хорошие отношения, Фарбера *назначили* директором ДК.

Правило 4

В случае однородных сказуемых, соединенных союзами *«а»*, *«и»* (морфологические характеристики проверяются у каждого предиката), когда в предложении один указатель на время (при этом он стоит в первой части предложения, предпочтительнее — на первом месте в предложении) — он приписывается обоим предикатам.

> *В конце февраля* крымский парламент отправил в отставку правительство республики и *назначил* главой Совета министров лидера движения «Русское единство» Сергея Аксенова.

> Глава администрации Белгорода *провел сегодня* внеочередное заседание муниципалитета и *назначил* нового первого зама.

Правило 5

Если в предложении с двумя глаголами есть два подходящих указания на время, то для каждому предикату приписывается ближайшее к нему слева.

> *В начале февраля* Михаил Рыбаков *пошел* на повышение и *еще через два месяца* был *назначен* на пост гендиректора этого совместного предприятия.

Правило 6

Если в предложении встречаются подряд два указания на время, причем второе выделено запятыми, то правило (1) не действует и оба временных выражения соотносятся с упоминанием ситуации, если это не противоречит остальным правилам.

> *Накануне, 1 марта,* Денис Березовский был *назначен* главой ВМС Украины указом исполняющего обязанности президента Александра Турчинова.

Если контекст не удовлетворяет ни одному из вышеприведенных правил, то ситуации ставится в соответствие ближайший временной указатель (если он существует).

Для каждого из приведенных выше случаев были составлены правила на диалекте языка CPSL, по которым между указанием на время и упоминанием ситуации автоматически строится отношение «`$время_события`».

5. Результаты тестирования

Для проверки работы правил была размечена тестовая коллекция. Использовались тексты коллекции Situations-1000 [11], в которых размечены ситуации отставки и назначения. Из этих текстов

были выбраны предложения, содержащие упоминания ситуаций отставки и назначения, описываемые с помощью маркеров, которые выражены словами «*назначить*» и «*уволить*». Рассматривались только такие предложения, в которых указанные ситуации были выявлены на предыдущих этапах анализа. Временные маркеры анализировались только самостоятельные. Временные предлоги, которые обозначают отношения между событиями в тексте, в анализ не включаются (*после, когда, перед* и т.п.).

Во всех предложениях, в которых встретились упоминания ситуаций отставки или назначения, обозначенные глаголами «*назначить*» или «*уволить*», а также временные указатели, были размечены отношения «`$время_события`» между глагольной словоформой и временным указателе. Всего получилось 637 эталонов разметки. Далее к тестовой коллекции были применены описанные выше правила. Были получены следующие результаты: точность 90.9%, полнота 92.62%, F-мера 91.75%.

Заключение

В настоящей работе была предпринята попытка установить некоторые формальные правила, по которым возможно было бы автоматическое построение отношения между выявленными в тексте упоминаниями событий определенного типа и временными указателями. Как показали результаты тестирования, такие правила работают достаточно эффективно на ограниченном материале в рамках одного предложения.

В результате анализа текстов выяснилось, что временные указатели для событий отставки и назначения встречаются в том же самом предложении в меньшей части случаев. Поэтому в дальнейшем предполагается расширить сферу поиска временных указателей для упоминаний событий (выйти за пределы предложения), а также попытаться сформулировать правила для автоматического определения последовательности событий в тексте. Кроме того, будет продолжена работа над усовершенствованием лингвистической модели времени, что позволит интерпретировать извлеченные из текста и привязанные к событиям временные указатели.

Благодарности. *Автор благодарит А. В. Подобряева за помощь в обработке данных.*

Список литературы

[1] Appelt D. E., “Introduction to information extraction”, *Journal AI Communications*, **12**:3 (1999), pp. 161–172 ↑ 231.

[2] Pustejovsky J., Ingria B., Sauri R., Castano J., Littman J., Gaizauskas R., Setzer A., Katz G., Mani I., “The Specification Language TimeML”, *The Language of Time: A Reader Mani*, eds. Pustejovsky J., Gaizauskas R., Oxford University Press, 2005 ↑ 232.

[3] Weiser S., Laublet P., Minel J.-L., “Automatic Identification of Temporal Information in Tourism Web Pages”, The International Conference on Language Resources and Evaluation (LREC) (2008) ↑ 232.

[4] Хорошевский В. Ф., «OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов», Труды конференции КИИ-2004 (Тверь, 2004) ↑ 232.

[5] Крапивный Ю. Н., Касаткина Г. В., «Проблемы и методы извлечения темпоральных знаний из текстов на естественном языке», *Электротехнические и компьютерные системы*, 2013, №09 (85), URL http://storage.library.opu.ua/online/periodic/ee_85/170179.pdf ↑ 232.

[6] Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л. и др., *Лингвистическое обеспечение системы ЭТАП-2*, Наука, М., 1989 ↑ 233.

[7] *Abbyy Compreno*, http://www.abbyy.ru/isearch/compreno/ ↑ 233.

[8] Ефименко И. В., «Время в мультиязычных коллекциях документов: лингвистическая модель и ее реализация в среде GATE», *Девятая Всероссийская конференция по искусственному интеллекту КИИ-2004*. т. 2, Физматгиз, Тверь–Москва, 2004, с. 525–532 ↑ 233.

[9] Ефименко И. В., *Модель времени в системах извлечения знаний из письменного дискурса*, Диссертация на соискание учёной степени кандидата филологических наук, М., 2007, URL http://rcdl.ru/doc/2012/paper45.pdf ↑ 233.

[10] Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В., «Извлечение информации из текста в системе ИСИДА-Т», *Труды XI Всероссийской научной конференции RCDL’2009*, КарНЦ РАН, Петрозаводск, 2009, с. 247–253 ↑ 235.

[11] Власова Н. А., «Извлечение информации о ситуациях отставок-назначений в новостных текстах. Опыт разметки коллекции. Результаты тестирования», *Труды XV Всероссийской научной конференции RCDL’2013 "Электронные библиотеки: перспективные методы и технологии, электронные коллекции"* (Ярославль, 2013), с. 145–154, URL http://rcdl2013.uniyar.ac.ru/doc/full_text/s4_2.pdf ↑ 235, 239.

[12] Situations-1000, *Размеченная коллекция новостных текстов на русском языке, содержащих информацию о назначениях и отставках лиц*, http://ai-center.botik.ru/Airec/index.php/ru/collections/33situations-1000, Исследовательский центр искусственного интеллекта, ИПС им. А. К. Айламазяна РАН, 2014 ↑ .

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

Об авторе:

Наталья Александровна Власова

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

e-mail: nathalie.vlassova@gmail.com

Образец ссылки на эту публикацию:

Н. А. Власова. *Об одной проблеме автоматического извлечения временной информации из русскоязычных текстов* // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. 5, № 4(22), с. 231–242.

URL http://psta.psiras.ru/read/psta2014_4_231-242.pdf

Natalya Vlasova. *On one problem of automatic information extraction from Russian texts.*

Abstract. In this paper we consider a problem of matching temporal expressions and target events. The target events are appointments and resignations. We introduce a set of rules to make this matching. We give the results on a test collection of Russian news texts. (*In Russian*).

Key Words and Phrases: Key Words and Phrases: automatic information extraction, temporal expressions, events, time aspect, test corpora.