

АНАЛИЗ И КЛАССИФИКАЦИЯ ХАРАКТЕРИСТИК ВРЕМЕНИ И СРОКОВ В РОССИЙСКИХ НОРМАТИВНЫХ ПРАВОВЫХ АКТАХ

ANALYSIS AND CLASSIFICATION OF TIME AND TERMS CHARACTERISTICS IN RUSSIAN LEGAL ACTS

Цибульский В.В. (VTsibulsky@it.ru), Ежов А.С. (nsib@space.ru), Поляков Г.А. (geopol@bk.ru), Феклистов В.В. (VFeklistov@it.ru)

Высшая школа экономики, Москва, Россия

Ключевые слова: именованные сущности, характеристики времени и сроков, структурно-семантическая классификация, российские правовые акты, автоматизация бизнес-процессов

Цель данной статьи - постановка задачи анализа и классификации характеристик времени и сроков в российских правовых актах, регулирующих административную деятельность. Представлено описание метода анализа временных характеристик, основанного на выделении устойчивых словосочетаний, характеризующих сущности "время" и "срок", и описание средств, с помощью которых проводился анализ, а также указаны их достоинства и недостатки. Результатом проведенного анализа явилась структурно-семантическая классификация временных характеристик административной деятельности, позволившая свести к минимуму количество вариантов их описаний и, как следствие, поставить задачу автоматизации контроля исполнения хода административных процедур.

Tsibulsky. V. V. (VTsibulsky@it.ru), Ezhov A.S. (nsib@space.ru), Polyakov G.A. (geopol@bk.ru), Feklistov V. V. (VFeklistov@it.ru)

Higher School of Economics, Moscow, Russian Federation

The paper describes a problem of analysis and classification of time and terms characteristics in Russian legal acts regulating administrative activity. We presented a model and computing environment for extraction of named entities for these characteristics and specified its merits and demerits. The created model of the "time and terms" named entities in this type of documents showed a large difference from the entities described in the literature. On the base of the analysis we developed a new structural-semantic classification of time and terms characteristics of the administrative activity. This classification allowed to reduce a quantity of variants in its descriptions and to formulate a task of automation for controlling the course of administrative procedures.

Key words: named entities, time and terms characteristics, structural and semantic classification, Russian legal acts, business process automation

Введение

В процессе перевода государственных услуг и функций в электронный вид, т.е. создания на базе предписаний таких нормативных актов, как административные регламенты, автоматизированной системы управления, возникает задача алгоритмизации контроля исполнения сроков и определения реакции системы на их истечение.

Одной из ключевых проблем при решении такой задачи для разработчиков является отсутствие закрепленной в государственном праве устойчивой классификации характеристик времени и сроков (далее – временных характеристик), и, как следствие свободы употребления языка, представление сроков и характеристик времени в правовых актах в разнообразном (не нормализованном для автоматизации) виде.

В своих исследованиях авторы не пытались решить задачу юридической классификации и нормализации понятий сроков и характеристик времени. Задачей исследования явилось выявление устойчивых словосочетаний, определяющих понятия временных характеристик в текстах нормативных правовых актов, утверждающих административные регламенты, и проведения классификации и нормализации терминологии методами лингвистического анализа.

В результате проведенных работ была также осуществлена постановка задачи автоматизации контроля сроков исполнения операций, процедур и регламентов оказания государственных услуг и исполнения государственных функций.

Методика решения проблемы

Задача практического моделирования для выявления возможных форм выражения сроков в правовой лексике представляла собой задачу не обязательно точного, но в значительной мере полного выделения сущностей определенного вида в ограниченном числе однородных документов. Технологии выделения именованных сущностей, появление которых часто связывают с работой Лиз Пау 1991 года, где было предложено использовать набор правил для выделения названий компаний в тексте, получили развитие в последующие годы. Уделяя наибольшее внимание выделению именованных сущностей, таких как имена PERSON, ORG – названия организаций и LOC – названия мест, разные авторы отмечали, что классификация сущностей этим не ограничена. Так широкий набор сущностей, таких как: наименования национальностей, названия событий, и другие были введены как новый тип MISC - разное.

В классификации Satoshi Sekine которая продолжает развитие классификации (Named Entity set) предложенной MUC (Grishman et al., 1996), и классификации, разработанной IREX (Sekine et al., 2000), и последующей расширенной иерархии (Extended Named Entity hierarchy) содержащей уже 150 типов сущностей (Sekine et al., 2002). временные сущности классифицируются по 11 типам.

Выражения времени, которые следует выделять как сущности расширенной иерархии именных сущностей, должны отвечать на вопросы: Time: "When (TIME) was that?" или Period: "How long (PERIOD) was that?"

По этому определению такие выражения, как (1) *сегодня*, (2) *два дня тому назад*, (3) *в прошлом году* не задают точной даты. В отличие от них, неполная дата (4) *1 января* определяет точную дату месяца, хотя год и не определен и поэтому будет выделяться как сущность.

Следующие примеры показывают отличие описываемой модели от расширенной классификации. Например, в следующих примерах: (5) *Бабье лето было дождливым, – весь период или дожди* и (6) *Бабье лето началось поздно, – теплая погода установилась после обычной даты*, - одна и та же словарная фраза может быть отнесена как к типу **Момент**, так и к типу **Период**. Фраза (7) *в тот же день* в тексте нормативного акта определенно устанавливает срок до окончания рабочего дня, несмотря на то, что она не определяет точную дату по С.Секин.

Кроме того, в этой классификации выделяются сущности типа "возраст" (AGE), которые нередко выражаются теми же длительностями, что и правовые сроки: (8) *15 лет и 6 месяцев*. Особенность языкового выражения единиц времени и географических координат также следует принять во внимание.

По аналогии с термином "именованная сущность" (named entity), определяемой как объект определенного типа, имеющий имя, название или идентификатор, мы будем говорить о временной сущности как объекте, имеющем измеримую временную характеристику.

Моделирование распознавания фрагментов правового текста, содержащего временные характеристики описываемого процесса, проводилось в предположении, что эти фрагменты представляют отдельную неразрывную последовательность токенов: слов, чисел и некоторых знаков препинания. Модель учитывала возможные способы выражения сущности времени в русском языке и давала возможность при нахождении в тексте простых элементов обнаруживать сложные структуры, описывающие как моменты времени, так и периоды. Поскольку основной задачей являлось теоретическое исследование возможности обнаружения сроков, пригодных для программирования электронных регламентов, определенность длительности сроков и моментов времени являлась необходимым условием для последующего практического использования.

Поэтому модель предполагала важным наличие у выделяемой сущности, характеризующей срок, количественного (измеряемого) атрибута или указания на него. В этой связи такие применимые в гражданском праве конструкции, как "разумный срок", а также "соразмерный", "согласованный", "краткий" или "ограниченный" сроки не могут быть использованы в электронном административном регламенте для запуска программного таймера, которому требуется указание точного времени отсчета.

В ходе реализации модели выделялись следующие типы сущностей:

Момент – календарная дата и/или точное время.

Срок - промежуток времени, имеющий не менее двух характеристик из следующих: длительность, момент начала или момент окончания.

Отсылка – фрагмент в тексте, относящийся к времени с неопределенными характеристиками, указывающий на источник такой информации вне сопряженного текста.

Такие сущности в текстах представлены фрагментами, которые обычно содержат слова, которые могут быть собраны в несколько частных словарей по различию их ролей в этом фрагменте. Например, обозначения единиц времени составили один словарь, а прилагательные, обозначающие разновидности времени – другой. Так, секунда, минута, час, неделя, год, десятилетие, столетие и т.д. собранные в одном словаре, позволили выделить маркеры времени, на основе которых стало возможным обнаруживать в тексте более крупные фрагменты текста, не содержащие не относящихся к обозначению временных характеристик слов. Под **маркерами времени** понимаются слова, значение которых содержит тему времени: его единицы и их

дериваты, указание на взаимное расположение во времени – например, *одновременно, позднее*; даты и сроки.

Сущности **Момент** и **Срок**, встречающиеся в текстах, выделялись известными способами. Так сезоны и части суток выделялись по словарю, цифровые даты - по регулярному правилу, а буквенно-цифровые даты по комбинированному правилу с использованием словаря месяцев и регулярных правил для допустимых дат.

Собранная в словари лексика использовалась при дальнейшем построении более сложных правил для выделения несловарных сущностей в тексте. Используемая лингвистическая модель предполагала, что фраза, описывающая срок, имеет структуру, в которой каждый элемент может выполнять свою роль указания на течение времени. Для таких элементов составлены простые правила выделения в тексте. Например, фрагменты : (9) *не позднее чем за*, содержащий четыре слова, и (10) *в срок до*, содержащий два предлога и маркер, выполняют роль ограничителей срока. Эти фрагменты, часто используемые для верхнего и иногда нижнего ограничения срока могут быть выделены из текста отдельным правилом, учитывающим частный словарь и общелексический. Тогда редуцированный вариант (11) *до* и более развитый (12) *за сутки вплоть до* тоже могут быть выделены как сущности того же простого типа.

К простым правилам отнесены, например, правила и словари, выделяющие в тексте следующие основные типы сущностей:

TimeBorder – словарная сущность, временная граница: начало, окончание, завершение и т.д.

TimeDuration – словарь: в течение, не менее, превышающий, не превышающий.

TimerLex1 – словарь маркеров времени: период, срок, дата, год, завершение, продолжительность.

Ru-XX-Month – словарь месяцев. Январь- декабрь.

TimePreposition – словарная сущность, предлог, употребляемый при указании момента времени: в, до, после, с.

Days – словарная сущность, разновидности производственного времени: финансовый, календарный отчетный, рабочий.

DayNum – регулярное правило, группа цифр, могущая представлять число месяца. Ничто не мешает в этом правиле цифры теми же алгоритмами записывать в виде последовательности числительных либо римских цифр. Правило может срабатывать как на один так, и на несколько токенов. Проверка на допустимость числа для даты здесь не проводится.

Hour – для сравнения с предыдущим: словарная сущность: целые часы. Эта же сущность может быть выделена регулярным выражением.

TimeShifter – словарная сущность, признак сдвига срока. Так, фрагмент (12) , *следующий за* представляет сущность с ролью отсылки на другую точку во времени и такого рода сущности собраны в независимое правило. Этот сдвиг временной привязки может быть также выражен вариантами (13) *со дня*, (14) *от первой даты* и т.д.

DocLink – словарная сущность, признак отсылки на другой документ.

С использованием подготовленных правил выделения простых сущностей в ходе моделирования были сформированы сложные правила, допускающие вариации отдельных элементов в их составе.

К ним отнесены:

TimerFull – сложная совокупность простых сущностей, представляющая собой полное указание на сущность типа срок.

TimerSentence – предложение, содержащее срок.

Ранее в порядке другого исследования были подготовлены правила для выделения сущностей, не описанных в расширенной классификации именованных сущностей, но общеупотребимых в правовой сфере.

FullExtraPointers – ссылка по тексту на внешний правовой акт.

ListGroups – группа ссылок на внешние документы. Эта конкретная сложная сущность является объединением последовательных ссылок в виде перечисляемых реквизитов документов. В правовых текстах такие ссылки встречаются часто и сгруппированы по изданиям, правовым актам и их разделам. Даты изданий, содержащиеся в ссылках, маскируются и не принимаются в рассмотрение при анализе сроков.

На рисунке 2 представлено правило, отвечающее реальной фразе из документа на рисунке 1. Сложная сущность Period формируется как последовательность простых сущностей, в том числе, словарных и определенного числа произвольных токенов. Рисунок 3 показывает как сконфигурировано это правило.

Например, сущность DayNum – необязательна в правиле, она может быть выражена как числом, так и числительным. А Token должен быть запятой, непосредственно следующей за предыдущим словом из словаря обозначений времени TimeLex1.

не позднее 12:00 второго рабочего дня, следующего за днем

Рис. 1. Пример моделируемой фразы о времени

Избранный подход позволил сформировать ограниченное число сложных правил, которые неплохо выделяют сроки (и их временные характеристики) из значительного количества правовых русскоязычных текстов.

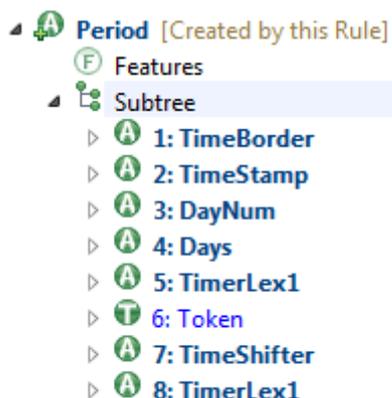


Рис. 2. Пример сложного правила выделения сущности как последовательности простых аннотаций

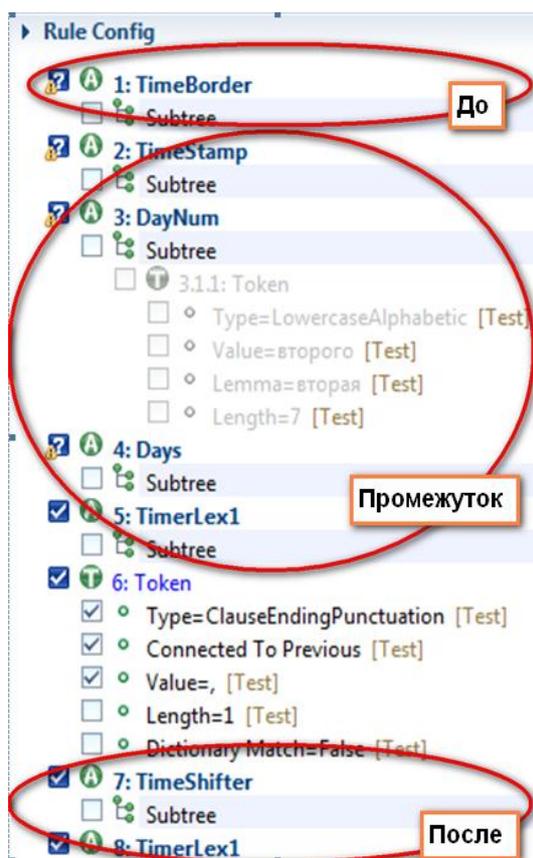


Рис. 3. Пример структуры сложной сущности

Разработка правил выделения сроков проводилась в прототипе среды подготовки лингвистических ресурсов компании IBM для продукта Content Analytics, носящем в настоящее время название LanguageWare Resource Workbench (LRW) и поставляемом исключительно в составе основного продукта и который можно было легально использовать партнерам компании разработчика для академических (исследовательских и научных) целей.

Настроенный в ходе тестирования на наборе документов комплект правил выделения сущностей типа Срок был перенесен в открытую среду обработки неструктурированной информации UIMA, где было выполнено поточное выделение аннотаций – аналогов сущностей из всех доступных текстов административных регламентов, принимаемых органами государственной власти Российской Федерации.

В среде UIMA, поддерживаемой сообществом как проект Apache, для выделения сущностей из текстов используются так называемые аннотаторы - программы, которые можно формировать из набора простых java-классов, выполняющих последовательные операции по обработке текста.

Эти классы могут применять к тексту шаблоны формального языка поиска и осуществления манипуляций с подстроками Regex.

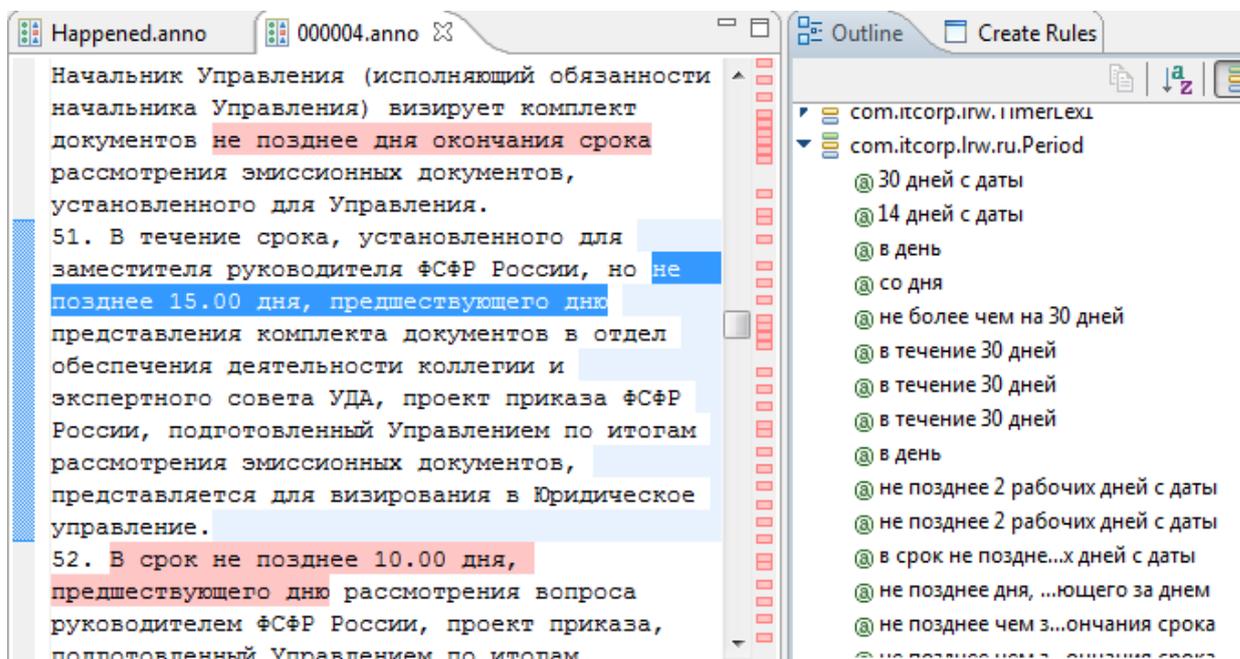


Рис. 4.. Результаты тестирования правила

Пример документа, полученного для просмотра в типовом редакторе, показан на рисунке 5. Ярко-зеленым подсвечен текст абзаца, где встречается фрагмент, выделенный аннотатором UIMA как сущность Period. Золотистый цвет выделяет саму сущность типа "Срок".

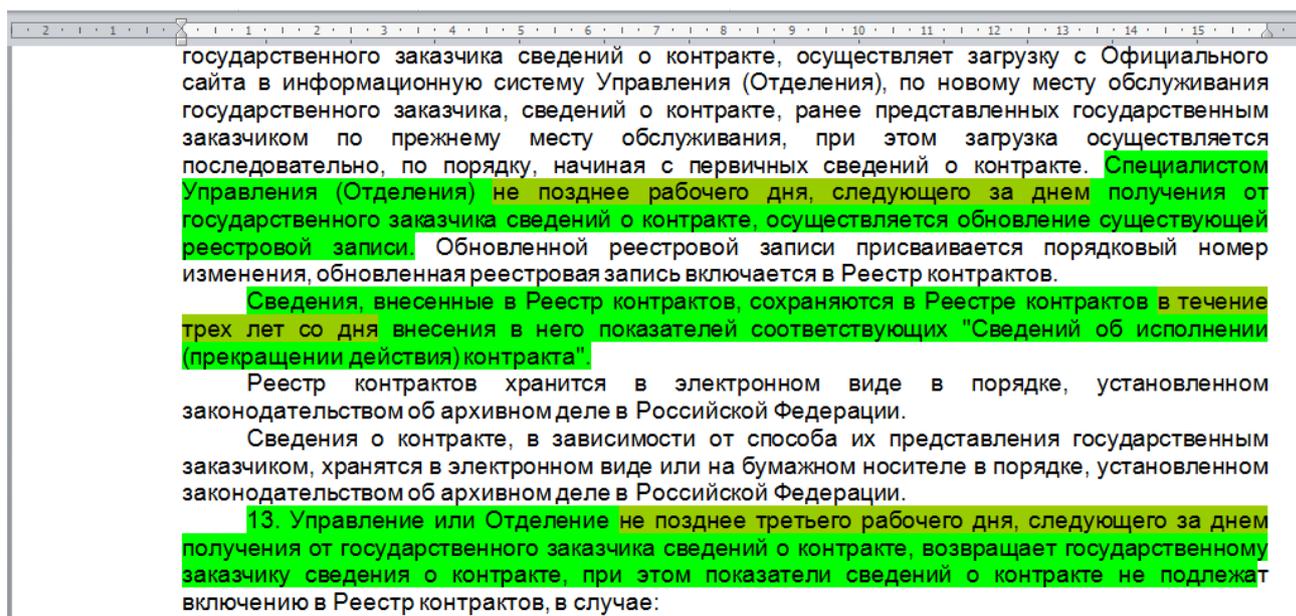


Рис. 5. Пример текста регламента в редакторе с выделенными фрагментами, касающимися сроков

Полученные фрагменты текстов подвергли анализу и статистической обработке.

Для быстрой автоматизированной разметки текстов были применены архитектура управления неструктурированной информацией UIMA и АРМ подготовки лингвистических ресурсов LRW. Эти

инструменты вполне подходили для решения нашей задачи, мы имели возможность общаться с разработчиками и применять несколько следующих друг за другом версий систем. При постоянном совершенствовании программной среды, тем не менее, мы столкнулись с ограничениями, мешающими широкому применению этих инструментов для разметки русскоязычных текстов.

Архитектура UIMA использует язык Java для создания приложений и ограничена его характерными особенностями, что нередко проявляется в нехватке памяти и сложности отладки многоступенчатой обработки текста. Архитектурой UIMA не предусмотрено создание инструмента для вовлечения в работу по формированию лексических ресурсов лингвистов без уверенных навыков программирования. UIMA не предоставляет с дистрибутивом общих русскоязычных ресурсов. Потенциал развития технологии еще далеко не исчерпан.

LRW дополняет производительную среду исполнения, предоставляемую UIMA, и дает возможность создать готовый программный код для работы с лексикой непрограммисту. Этот код, упакованный в аннотатор, и снабженный словарями и базами правил может быть широко использован в приложениях, поддерживающих архитектуру UIMA. Программисты критикуют его за невозможность рефакторинга. Лингвистам не нравится невозможность замены лингвистического процессора на отличный от предоставляемого разработчиком.

С точки зрения применения правил выделения сущностей недостатком является невозможность отмены маскирования многословными терминами терминов меньшей длины. Проблема с отсутствием частеречной разметки для русского языка решена и в коммерческом продукте с начала 2012 года уже применяется так называемый POS-tagging для всех основных европейских языков. Компания разработчик – поставляет свои лингвистические ресурсы с LRW в закрытом виде и быстрое формирование частных словарей на их основе затруднено, несмотря на возможность удобного импорта. Это создает необходимость вовлечения в работу большого числа лингвистов. Для любых языков справедлива настоятельная рекомендация не включать в словари многословные термины и это еще более актуально для русского языка с согласованностью слов в словосочетаниях. Ручное заполнение словарей решает эту проблему, но не гарантирует уверенность в пересечении с закрытыми словарями.

LRW использует особый диалект языка регулярных выражений Regex и это заметно ограничивает различение сущностей написанных разными регистрами и в расширенной кодировке. Сильно ограничена также привязка к структуре текста, что особенно заметно на правовых текстах. LRW поддерживает идею UIMA последовательного применения аннотаторов к тексту с возможностью перестраивания цепочки обработки по условию, но на практике изменения в отдельных поздних этапах могут влиять на результат аннотирования на ранних. Поэтому, нередко настройка длинных цепочек становится искусством.

Применение современного инструмента подготовки текста дает значительное преимущество даже на правовом тексте. Узкая задача достаточно полно решается с использованием маскирования сущностей и небольшим набором правил. Допускается построение классификации сущностей ad hoc, с настройкой по корпусу, и последовательное использование для обработки текстов независимых классификаций сущностей. Без особых трудностей создаются сложные сущности, специфичные для специальных текстов.

Понимая, что описанная выше среда разработки является частью лингвистического процессора вопросно-ответной системы Watson производства IBM, следует предположить, что ее технологические ограничения должны быть преодолены при разработке аналогичных систем.

Классификация сроков и характеристик времени

Итак, как было указано, исследования в области анализа и классификации выражения временн х характеристик и сроков ранее в основном проводились с точки зрения логики в рамках гражданского права, и классификация гражданско-правовых сроков достаточно хорошо разработана. Однако лингвистический анализ выражения временных характеристик ранее не проводился, если иметь в виду структурно-семантические структуры и способы выражения этих характеристик.

Проведенные нами анализ и моделирование, которые были описаны в предыдущем разделе, дают возможность сформулировать лингвистические принципы формирования реализуемой сущности "времени". С точки зрения семантики структуру выражения характеристик времени и сроков в документах можно описать следующим образом.

М

х характеристик:

1. **Момент времени:** указание календарной даты (день, месяц, год, иногда с уточнением до часа, минут и секунд), а также дня недели. Однако подобный тип выражения времени встречается редко как самостоятельная сущность, чаще подобная сущность входит в качестве составной части в другие типы. Данный тип можно отнести к классу "**МОМЕНТ**".

2. **Период времени:** указание или задание промежутка времени, исчисляемого годами, месяцами, неделями, днями и часами, изредка минутами.. Данный тип можно отнести к классу "**ПРОМЕЖУТОК**".

3. **Ограничение срока:** указание на ограничение времени определенным моментом времени "в будущем". Данный тип можно отнести к классу как "ДО".

4. **Отсылка к моменту времени,** начинающемуся после некоторой временной точки "в будущем". Данный тип можно отнести к классу "ПОСЛЕ".

В реальных текстах эти элементарные типы могут использоваться самостоятельно, образуя простые конструкции выражения характеристик, как показано в нижеприведенных примерах, взятых из реальных регламентных документов:

(15) ... приказ ... от 11 февраля 2008 года; (16) ... в день окончания срока...; (17) ... в четверг... ("МОМЕНТ")

(18) ... проведения публичного мероприятия в течение 1 рабочего дня...; (19) ... 10 рабочих дней...; (20) ... за прошедший календарный год...; (21) ... отчетность в срок, указанный в конкурсной документации...; (22) ... в течение определенного срока... ("ПРОМЕЖУТОК")

(23) ... до момента...; (24) ... до начала...; (25) ... до последнего дня...; (26) ...до ранее установленного срока... ("ДО")

(27) ... после 23 часов...; (28) ... после дня окончания срока...; (29) ...не ранее 15 дней...; (30) ...с момента регистрации... ("ПОСЛЕ")

Те же самые элементарные типы могут объединяться и формировать более сложные семантические структуры выражения времени и сроков. Теоретически эти элементарные типы могут объединяться логическими операциями AND и OR как с себе подобными, так и с другими семантическими типами, формируя в результате составные конструкции для выражения временных характеристик. Далее тоже примеры, не имеющие отношения к постановке задачи. Во всяком случае, многие из них.

(31) ... пять дней до даты окончания срока (ПРОМЕЖУТОК + ДО)

(32) ... пять рабочих дней со дня заключения контракта (ПРОМЕЖУТОК + ПОСЛЕ)

(33) ... суббота, воскресенье - выходные дни (МОМЕНТ + МОМЕНТ)

(34) ... от 3 до 90 рабочих дней в зависимости от наличия технической возможности (ПОСЛЕ + ДО)

(35) ... на следующий рабочий день за днем (ПОСЛЕ + ПОСЛЕ) и т.д.

По правилам комбинаторики из двух типов можно получить 20 комбинаций, однако не все возможные структуры встречаются в реальных текстах. Характер документов (регламенты) наложил свой отпечаток на возможность сочетаемости различных типов. В частности, регламентирующий характер документов практически исключает использование логической операции OR. Но, тем не менее, при анализе необходимо учитывать такую возможность. И действительно, в анализируемых текстах встретились примеры подобного типа:

(36) ... часов и (или) ранее 8 часов...

(37) ... максимальный срок выполнения действия - тридцать минут (в случае ...) или один рабочий день (в случае ...)....

Но семантическая картина выражения характеристик времени и сроков не ограничивается простыми конструкциями из четырех элементарных типов и максимально возможными двадцатью двусоставными конструкциями. В текстах возможны и реально встречаются комбинации из большего количества простых типов, что также необходимо принимать во внимание при формировании конечного набора структурных классов. Например, не исключены следующие варианты (логические операции и иерархия отношений не указаны):

(38) ... не ранее 15 дней и не позднее 10 дней до дня проведения... (ДО + ПОСЛЕ + ДО)

(39) ... считается день, следующий за днем исполнения документа (в рабочих днях) ... (МОМЕНТ + ПОСЛЕ + МОМЕНТ);

(40) ...в течение 3 месяцев до или 3 месяцев после даты проведения выборов (ПРОМЕЖУТОК + ДО + ПРОМЕЖУТОК + ПОСЛЕ + МОМЕНТ)

... в течение 15 (пятнадцати) дней от даты заключения настоящего договора... (ПРОМЕЖУТОК + ПОСЛЕ + МОМЕНТ) и т.п.

Некоторая сложность выделения подобных типов связана и с тем, что в реальных ситуациях иногда бывает трудно понять, где проходит грань между понятиями "МОМЕНТ" и "ПРОМЕЖУТОК". На самом деле, практически любое обозначение временной характеристики включает или подразумевает протяженность во времени. Например, в контексте (41) суббота, воскресенье - выходные дни подразумеваются 2 промежутка, однако для их обозначения используются тип элементарной характеристики который мы определяем как "МОМЕНТ".

Таким образом, семантика выражения характеристик времени и сроков может быть описана как множество элементарных или составных конструкций. При этом составные структуры могут быть иерархически структурированы, и это необходимо учитывать при решении поставленной задачи.

Заключение

В данной работе мы представили модель и вычислительную среду для выделения именованных сущностей, содержащих характеристики времени. Исследованию была подвергнута коллекция административных регламентов органов исполнительной власти РФ, содержащих описание процессов, исполняемых должностными лицами в порядке оказания государственных услуг. Модель сущностей типа "время" и "сроки" в силу особенности последующего практического применения серьезно отличается от описанной в литературе. Обнадеживающие по полноте результаты дают основания для подтверждения гипотезы о конечности типизации реально применяемых форм выражения временных сроков и моментов, что является необходимым условием успешности автоматизации такого рода бизнес-процессов. Проведенный анализ и моделирование распознавания фрагментов правового текста, содержащего временные характеристики описываемого процесса, дали возможность сформулировать лингвистические принципы формирования сущностей "время" и "срок". Структурно-семантическая классификация сущностей «время» и «срок», а также логические комбинации, которые могут быть построены из разных категорий сущностей, позволяют сделать предварительные выводы о том, что набор таких комбинаций конечен, если рассматривать классы в контексте исключительно разделов, описывающих собственно административные процедуры. Данное утверждение подлежит дальнейшему исследованию.

Тем не менее, уже на данном этапе может быть поставлена задача автоматизации контроля исполнения административных процедур на уровне классов категорий сущностей. А именно, необходимо создать следующие сервисы, обеспечивающие отсчет времени и формирования необходимых сигналов при его истечении, или начале.

Категория	Входные параметры	Выполняемые действия
МОМЕНТ	Дата, время	Сигнал при наступлении момента
ПРОМЕЖУТОК	Дата, время начала Дата, время окончания	Запуск таймера в срок начала. Остановка таймера в срок окончания. Сигнал при окончании.
ДО	Текущие дата и время Дата, время окончания	Запуск таймера. Остановка таймера в срок окончания. Сигнал при окончании

Категория «ПОСЛЕ» проконтролирована быть не может, т.к. не имеет окончания.

Данная спецификация упрощена и предварительна. При этом результатов проведенного исследования явно недостаточно для решения задачи автоматического контроля, т.к. за его рамками остались, в частности некоторые постулаты ГК, кроме того, в некоторых ведомствах существует иная практика определения конечных сроков. Поэтому механизмы, контролирующие время, должны также уметь обращаться с календарем с учетом специфики календарей некоторых субъектов РФ.

Целью дальнейших исследований описания временных характеристик является также автоматизированное формирование нормализованных значений описаний с целью их дальнейшего использования при анализе вновь созданных правовых и регулирующих документов на предмет выявления заложенных в них противоречий. Кроме того, для решения задачи автоматизации необходимо выявлять из текстов нормативных правовых актов и классифицировать ещё целый ряд сущностей, к которым, в частности относятся: административные процедуры; административные действия; сведения о должностных лицах, ответственных за выполнение административных действий; предмете административных действий. Что также является целью дальнейших исследований.

Данное исследование проводилось при финансовой поддержке Правительства Российской Федерации (Минобрнауки России) в рамках договора № 13.G25.31.0096 о «Создании высокотехнологичного производства кросс-платформенных систем обработки неструктурированной информации на основе свободного программного обеспечения для повышения эффективности управления инновационной деятельностью предприятия в современной России».

Литература

1. Rau L. F. Extracting company names from text. *Proceedings of the Seventh Conference on Artificial Intelligence Applications CAIA-92*. Miami Beach, FL, 1991, Vol. I: Technical Papers, pp. 29–32.
2. Grishman R., Sundheim B. Message understanding conference-6: a brief history. *Proceedings of the 16th conference on Computational linguistics*. Stroudsburg, PA, USA, 1996/Vol. 1, COLING '96, pp. 466–471.

3. Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. Stroudsburg, PA, USA, 2003, Vol. 4, CONLL '03, pp. 142–147.
4. Sekine S. and Ishara H. IREX: IR & IE evaluation project in japanese. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*. 2000.
5. Sekine S., Sudo K., Nobata Ch. Extended Named Entity Hierarchy *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)* .2002.
6. Гражданский кодекс Российской Федерации (часть первая) от 30.11.1994 N 51-ФЗ (ред. от 30.11.2011), "Собрание законодательства РФ", 05.12.1994, N 32, ст. 3301, "Российская газета", N 238-239, 08.12.1994.
7. Постановление Правительства РФ от 31.10.2009 N 879 "Об утверждении Положения о единицах величин, допускаемых к применению в Российской Федерации", "Собрание законодательства РФ", 09.11.2009, N 45, ст. 5352.
8. Постановление Правительства РФ от 16 мая 2011 г. N 373 "О разработке и утверждении административных регламентов исполнения государственных функций и административных регламентов предоставления государственных услуг", "Собрание законодательства РФ", 30.05.2011, N 22, ст. 3169.