ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №4(27), 2015, с. 401–418

УДК 004.89:004.912

Ю. П. Сердюк

Базовая архитектура, методы и алгоритмы системы извлечения темпоральной информации из текстов на естественном языке

Аннотация. В данной статье представляется базовая архитектура системы извлечения темпоральной информации из текстов на естественном языке. Определяются основные структурные компоненты такой архитектуры, а также методы и алгоритмы, которые в них реализуются. В частности, выделяется этап извлечения информации о темпоральных элементах — событиях и темпоральных ссылках в тексте. Подчеркивается необходимость использования синтаксических зависимостей между словами обрабатываемого предложения, а также семантических ролей групп слов для установления отношений порядка между темпоральными элементами, извлеченными из текста. Отдельной важной компонентой предлагаемой архитектуры является модуль логического вывода, использующего статистическую информацию. Соответственно, показывается необходимость использования методов машинного обучения и различных корпусов лингвистических данных (аналогичных WordNet, SemCor, TimeBank и др.) для успешного решения общей задачи.

Ключевые слова и фразы: Извлечение информации, темпоральные элементы, машинное обучение.

Введение

Задача извлечения информации из неструктурированных источников, включая тексты естественном языке (ЕЯ), решается исследователями уже на протяжении более двух десятков лет. Естественно, что главным источником таких данных является World Wide Web (WWW) [1, 2]. В рамках WWW наиболее популярными источниками, из которых делаются попытки автоматически извлекать значимую информацию, являются, в первую очередь, новостные ленты. Тем не

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

менее, в последнее время важными источниками для которых решается задача извлечения информации, являются источники, которые часто не присутствуют явно в Web, а именно, различного рода медицинские и деловые документы.

Ранние работы по извлечению информации, в первую очередь, концентрировались на выделении из текстов статических фактов с представлением их в виде отношений. Примерами таких отношений являются:

преподаватель (`Николай_Крылов`, `СПбГУ`),
президент (`Билл_Клинтон`, `США`).

Однако, почти сразу стало очевидно, что большинство извлекаемых фактов не являются по своей природе статическими, поскольку они обычно являются истинными только в пределах некоторого интервала времени (в частности, Билл Клинтон был президентом США с *20 января 1993 года* по *20 января 2001 года*).

Основными сущностями, которые важны при извлечении темпоральной информации из ЕЯ-текстов, являются

- события и
- темпоральные ссылки (указатели).

Эти оба вида сущностей далее в тексте будут называться *темпоральными элементами*. Для таких элементов почти всегда можно указать

- время начала (`beginning time`) и
- время окончания (`ending time`)

события или временной ссылки. Такого рода точечные временные отметки являются базой как для упорядочивания (во времени) темпоральных элементов, так и для проведения рассуждений над событиями и темпоральными ссылками для вывода новых фактов о них.

Соответственно, решение общей задачи извлечения темпоральной информации из ЕЯ-текста может быть разбито на три последовательных этапа:

(1) извлечение информации о событиях и темпоральных ссылках,
(2) установление зависимостей между темпоральными элементами,
(3) упорядочение темпоральных элементов и получение новых отношений между ними с использованием логического вывода.

Первый этап общей задачи — извлечение информации о событиях и темпоральных ссылках — на настоящий момент довольно хорошо исследован и развит. Результаты этой работы реализованы в нескольких практических системах таких как Evita [3], SUTime [4], TRIPS/TRIOS [5]. В частности, разработан язык формальных спецификаций TimeML [6], который представляет собой формализм для представления информации о событиях и темпоральных ссылках, присутствующих в тексте на естественном языке. Язык TimeML является по своей структуре XML-подобным языком, в котором, в частности, имеется тег `EVENT` с соответствующими атрибутами для представления информации об извлеченном событии, а также тег `TIMEX3` со своими атрибутами для представления информации об извлеченной темпоральной ссылке. Отметим, что на данном первом этапе, для решения задачи извлечения информации о событиях и темпоральных ссылках, используются как чисто лингвистические методы (например, разбор входных ЕЯ-предложений на части речи и определение семантических ролей глаголов), так и статистические методы (методы машинного обучения). Более подробно этап извлечения информации о событиях и темпоральных ссылках будет описан в разделе 1.

После того, как информация о событиях и темпоральных ссылках извлечена, наступает второй этап — установление зависимостей между темпоральными элементами. Под зависимостью между темпоральными элементами понимается отношение (темпорального) порядка

$$t(e_1) \leqslant t(e_2),$$

где $t(e_i)$ есть время начала или время окончания темпорального элемента e_i, $i = 1, 2$.

Установление таких зависимостей обычно происходит за два шага: вначале выделяются синтаксические зависимости между отдельными словами предложения, которые определяют некоторое упорядочение темпоральных ссылок на временной оси; затем эти синтаксические зависимости между словами преобразуются в отношения (темпорального) порядка между темпоральными элементами, в состав которых входят эти слова. И снова, и на этом этапе, используются методы машинного обучения (в частности, при определении зависимостей между словами с использованием семантических ролей). Более подробно этап установления зависимостей между темпоральными элементами описан в разделе 2.

Заключительным этапом работы системы извлечения темпоральной информации является этап глобального логического вывода на множестве концевых точек (точек начала и окончания) темпоральных элементов. Результатом этого этапа является упорядочение темпоральных элементов на оси времени.

Для получения более правдоподобных результатов, совпадающих с реальными ожиданиями, часто применяют форму логического вывода, использующего вероятности. Одним из таких формализмов являются *логические сети Маркова* (ЛСМ) [7], которые представляют собой вероятностное расширение логики 1-го порядка. ЛСМ есть гибкий способ встроить знания человека о фактах реального мира, которые не являются «жесткими», т.е., имеют некоторую вероятностную природу, в механизм логического вывода. Формально, ЛСМ есть множество формул логики 1-го порядка, снабженных весами, представляющими собой вещественные числа. Вычисление значений весов и их привязывание к формулам осуществляется, обычно, на основе процесса обучения на множестве тренировочных данных. Примером лингвистического корпуса аннотированной информации о событиях и темпоральных ссылках, который может быть использован для такого обучения, является TimeBank [8]. Более подробное описание этапа логического вывода на множестве темпоральных элементов представлено в разделе 3.

В разделе 4 мы кратко формулируем результаты, изложенные в данной статье, и представляем направления дальнейшей работы.

1. Извлечение информации о событиях и времени

Темпоральные ссылки имеют смысл и могут интерпретироваться только в связи с событиями, о которых идет речь в ЕЯ-тексте и с которыми они ассоциированы. Соответственно, одной из важнейших подзадач в общей задаче извлечения темпоральной информации является максимально полное и корректное извлечение информации о событиях.

В ранних подходах к решению задачи извлечения информации о событиях, часто использовалось множество заранее фиксированных типов отношений (событий), которые предполагалось извлекать из текста на естественном языке. Такой подход, в общем случае, нечувствителен к контексту, в рамках которого описывается или интерпретируется то или иное событие. В частности, в подходе на основе шаблонов (предустановленных типов событий) очень трудно учесть

различного рода модальные определители (квалификаторы), относящиеся к описываемому событию. К такого рода определителям относятся, например, сообщения о чем-либо (reporting), слова, выражающие намерения или попытки (intending/attempting) и т.п. Поэтому важными требованиями к системе извлечения информации о событиях являются

(1) отсутствие ограничений, связанных с предустановленным списком типов событий,
(2) применимость к любой предметной области.

Не пытаясь дать полное и точное определения понятия «событие», отметим лишь одно очень важное его свойство, которое является тем более адекватным в рамках задачи извлечения темпоральной информации. А именно, «событие» есть такая сущность, которой всегда может быть приписана либо некоторая временная точка, либо интервал на оси времени. Как следствие, событие всегда может быть упорядочено на оси времени по отношению к другим событиям и временным указателям.

Основой идеей, позволяющей выполнить требования к системе извлечения информации о событиях, отмеченные выше, а именно, требования к отсутствию предустановленного списка типов извлекаемых событий, а также применимость к любой предметной области, является идея поиска описаний событий в тексте на естественном языке по определенным *синтаксическим формам*, с помощью которых обычно выражаются события.

В качестве синтаксических форм, выражающих события, можно выделить следующие синтаксические формы [**3**], которые инвариантны для большинства европейских языков, включая русский язык:

(1) предложения, в которых участвует глагол в некотором времени («*рыбаки давно* покинули *это место*»),
(2) предложения с глаголами в неопределенной форме («восходить *на Эверест*»),
(3) именные группы («*быстрый* рост *автомобильной промышленности*», «*многотысячные антивоенные* демонстрации»),
(4) существительные, которые непосредственно обозначают события («*гибридная* война»),
(5) группы слов, ассоциированные с прилагательным («*полностью* вооруженный»).

Подчеркнем, что данные синтаксические формы являются первичными объектами, которые выделятся из ЕЯ-текста, и из которых в дальнейшем будут уже непосредственно извлекаться события.

Каждое извлекаемое событие характеризуется некоторыми его свойствами, имеющими грамматический характер. К настоящему времени, для представления грамматических свойств событий выработан стандарт, а именно, список атрибутов и состав их значений, формализованных в виде структуры тега `EVENT` языка разметки темпоральных выражений TimeML [6]. Устоявшимися на практике атрибутами и их значениями для событий являются:

(1) класс события:
   (a) состояние («*быть президентом*»),
   (b) явление, случай («*путешествие*», «*перелет*»),
   (c) сообщение («*сказано утвердительно*»),
   (d) событие, выражающее намерение («*я попытался*»),
   (e) событие, связанное с восприятием, ощущением («*я заметил*»);
(2) полярность (положительная или отрицательная) — указывает на то, имело ли место данное событие или, наоборот, не произошло;
(3) модальность (выражаемая вспомогательными модальными глаголами «*мочь*», «*уметь*», «*хотеть*», «*желать*» и т.п. или наречиями типа «*вероятно*», «*по всей видимости*» и др.);
(4) время (прошедшее, настоящее, будущее);
(5) вид (совершенный/несовершенный).

Ниже приведен пример представления о некотором извлеченном событии, которое записано на языке TimeML:

«*5 октября 1993 года Ельцин* `<EVENT class = "OCCURRENCE" polarity = "POS" pos = "VERB" tense = "PAST" aspect = "NONE">`*запретил*`</EVENT>` *некоторые из политических левых и националистических организаций и газет, которые* `<EVENT class = "OCCURRENCE" polarity = "POS" pos = "VERB" tense = "PAST" aspect = "PERFECTIVE">`*поддержали*`</EVENT>` *парламент.*»

Для распознавания событий в тексте на естественном языке, обычно используется некоторая форма синтаксического разбора входного предложения, в частности, выделение и анализ частей речи. В качестве частей речи, которые являются возможными кандидатами на события, обычно рассматривают

(1) глаголы (ходить, видеть, работать, говорить),

(2) существительные (прибытие, вспышка, война) и
(3) прилагательные (перевернутый, обозначенный, искривленный, вооруженный).

Для извлечения событий в рамках этих трех категорий используются различные стратегии, уже сильно зависящие от конкретного естественного языка.

Тем не менее, при выделении событий, выражаемых существительными, приходится решать общую задачу для многих языков, а именно, задачу разрешения неоднозначностей — является ли данное существительное «событийным» или «несобытийным». Перед этапом разрешения такого рода неоднозначностей, для существительных — кандидатов на события, осуществляется их лексический анализ с привлечением лингвистических корпусов данных. Например, для английского языка для решения этой задачи используются банки WordNet [9], SemCor [10] и TimeBank [8]. Сам этап разрешения неоднозначностей обычно реализуется с применением методов машинного обучения, а именно, некоторого классификатора (например, байесовского), обученного на данных из некоторого аннотированного корпуса. Например, в системе Evita [3] для разрешения неоднозначностей в интерпретации событийных и несобытийных существительных используется классификатор, обученный на данных из корпуса SemCor.

Кроме извлечения информации о событиях, отдельной задачей является задача извлечения информации непосредственно о темпоральных ссылках (указателях). Эта задача является более простой по сравнению с задачей извлечения событий, поскольку понятие времени легче поддается формализации, чем понятие события, в частности потому, что темпоральные ссылки, в большинстве случаев, включают в себя числовые выражения.

Соответственно, извлечение информации о темпоральных ссылках обычно реализуется в виде трех последовательных этапов:

(1) извлечение числовых выражений, относящихся к темпоральным ссылкам;
(2) отыскание в тексте простых темпоральных выражений с помощью правил (шаблонов) над словами и числовыми выражениями;
(3) извлечение полных темпоральных выражений с использованием составных шаблонов

(см. реализацию такого подхода в системе SUTime [4]).

На выходе этапа извлечения информации о темпоральных ссылках получается исходный текст с соответствующими аннотациями присутствующих в нем темпоральных выражений. Для такого рода аннотаций имеется стандарт TIMEX3, который частью языка аннотаций TimeML [6]. Отметим, в языке TimeML имеется три вида средств аннотации:

- для аннотации событий,
- для аннотации темпоральных ссылок (указателей) и
- для аннотации отношений (упорядочений) над событиями и темпоральными ссылками.

Таким образом, TimeML является довольно богатым языком, который предназначен для представления информации, получаемой не только на этапе извлечения темпоральной информации, но также и на этапе ее *интерпретации*.

Формат TIMEX3 является расширением более раннего формата TIMEX2, который предназначен для представления нормализованных значений времени и дат, а также для представления временных интервалов. Для более строгого представления информации о времени и датах, оба формализма опираются на стандарт ISO 8601, который используется для представления значений атрибута `value` в этих языках аннотаций.

Состав и структура атрибутов для представления информации о темпоральном выражении постоянно развиваются. Формат TIMEX3, в частности, содержит более богатый список атрибутов, чем формат TIMEX2. Одними из главных атрибутов формата TIMEX3 являются

type (тип),
value (значение),
mod (модификатор).

Например, простое темпоральное выражение «*не более, чем 60 дней*» на языке TIMEX3 будет представлено как

```
<TIMEX3 type = "DURATION" value = "P60D"
mod = "EQUAL_OR_LESS">не более, чем 60 дней</TIMEX3>.
```

Наиболее важным атрибутом в представлении информации о темпоральном выражении, который классифицирует данное выражение, является атрибут `type`. В языке TIMEX3 для него предусмотрено четыре возможных значения: `"DATE", "TIME", "DURATION", "SET"`.

Тем не менее, конкретные системы реализации, которые решают задачу извлечения темпоральной информации из ЕЯ-текста и

представления ее в некотором заданном формате, обычно расширяют выходной формат своими собственными дополнительными атрибутами. Среди наиболее развитых систем такого рода следует отметить системы GUTime [11], SUTime [4], HeidelTime [12] и TRIPS/TRIOS [5]. Для такого рода систем в текущей литературе установилось общее название «темпоральные таггеры».

Следуя общей стратегии *поэтапного* извлечения инфориации о темпоральных ссылках [2], в частности, для более четкого отделения

- этапа первичного извлечения информации

от

- этапа извлечения информации с использованием средств интерпретации,

более удобно иметь два формата представления информации о темпоральных ссылках:

(1) формат, аналогичный TIMEX2 или TIMEX3, предназначенный для представления первичной информации, извлеченной из темпорального выражения, и
(2) более богатый формат (включающий в себя первый формат), предназначенный для полного представления информации о темпоральных ссылках, полученной, в частности, уже и с применением различных методов интерпретации.

Отметим, что для получения полной информации о темпоральных ссылках, в методах интерпретации, в свою очередь, могут использоваться

(1) информация о событиях, извлеченная из данного ЕЯ- текста,
(2) модель (онтология) времени и
(3) база знаний о реальном мире (common-sense knowledge base).

2. Установление зависимостей между темпоральными элементами

После этапа извлечения темпоральных элементов (событий и времен), ключевым шагом к их упорядочению является этап установления зависимостей между парами (e_1, e_2) темпоральных элементов. Здесь под зависимостями понимаются синтаксические зависимости между словами предложения, которые задают темпоральные отношения между элементами, в которые входят указанные слова. Типичный пример синтаксической зависимости, которая определяет темпоральное отношение, демонстрирует следующее предложение:

«Австралия стала независимой с 1901 г.»

В этом предложении есть синтаксическая зависимость вида *предлог_с* (независимой, 1901 г.) Данное отношение означает, что независимость (Австралии) «произошла» в некоторый момент времени в 1901г. и имеет место быть с того времени.

Зависимости такого вида (и многие другие) являются результатом работы синтаксического анализатора ЕЯ-предложений. Примером наиболее развитого синтаксического анализатора такого рода, существующего, кроме английского, и еще для некоторых языков, является Stanford Parser [**13**]. Его составной частью является подсистема Stanford Dependencies (SD) [**14**], которая выводит около 80 видов синтаксических зависимостей между словами предложения. В частности, для предложения *«Steve Jobs revealed the iPhone in 2007»* подсистемой SD будут выведены следующие основные зависимости:

nsubj (`revealed`, `Jobs`),
det (`iPhone`, `the`),
dobj (`revealed`, `iPhone`),
preposition_in (`revealed`, `2007`).

Неформально, зависимость *nsubj* задает отношение между подлежащим и сказуемым предложения. Зависимость *det* определяет связь между главным словом именной группы и определяющим словом (необязательно артиклем). Зависимость *dobj* задает отношение между сказуемым и дополнением. Зависимость *preposition_in* связывает, в данном случае, глагольную группу с темпоральным указателем.

Таким образом, на этапе синтаксического анализа, система извлечения темпоральной информации рассматривает синтаксическое дерево разбора и выведенные из него зависимости. Из этих зависимостей выбираются те, которые связаны с темпоральными ссылками (типа preposition_in, которая приведена выше). Далее, выбранная зависимость, определенная для слов предложения, переводится в зависимость между темпоральными элементами в соответствии с простым правилом: если слова ω_1 и ω_2 находятся в некоторой зависимости и они принадлежат текстуальным выражениям e_1 и e_2, представляющим собой темпоральные элементы, соответственно, то аналогичная синтаксическая зависимость переносится и на элементы e_1 и e_2. Также на этом этапе рассматривается и случай, когда некоторое событие e не имеет синтаксических связей с другими элементами в предложении. В этом случае задается специальная зависимость «сходства»,

которая приравнивает данное событие e к ближайшему к нему элементу в синтаксическом дереве разбора. Это делается для того, чтобы в результате работы всей системы было обнаружена и извлечена информация о максимальном количестве событий, описываемых в ЕЯ-тексте, совместно с наиболее точным и правдоподобным множеством темпоральных ограничений (упорядочений) над этими событиями.

На этом же этапе, в дополнение к синтаксическим зависимостям, выводятся зависимости, получаемые с использованием семантических ролей (глаголов) [**15**]. Например, в предложении *«В большинстве стран, восстановление после Великой Депрессии началось в конце 1931 – начале 1933гг.»*, глагол *началось* будет выделен как событие, и для него будут выведены аргументы "Тема/Предмет" («восстановление после Великой Депрессии») и темпоральный аргумент, представляющий собой обстоятельство времени (*в конце 1931 – начале 1933 гг.*).

Таким образом, алгоритм работы системы на данном подэтапе будет состоять в

(1) рассмотрении каждого глагола, идентифицированного как событие,
(2) отыскании для этого глагола его темпорального аргумента (если таковой существует) и
(3) определения зависимости между глаголом и темпоральным аргументом на основе предлога, использованного в данном аргументе.

В общем виде, зависимости, получаемые с использованием выделения семантических ролей (Semantic Role Labeling) задаются в форме $srl\_dependency(e_1, e_2)$, где e_1, e_2 есть некоторые темпоральные элементы.

3. Упорядочение темпоральных элементов

Третьим, и заключительным, этапом работы системы извлечения темпоральной информации из ЕЯ-текстов, является этап, на котором для каждой выведенной пары темпоральных элементов (которыми, напомним, могут быть либо события, либо времена (темпоральные ссылки)) определяется (выводится) отношение порядка относительно оси времени. Исходной информацией для этого этапа являются зависимости между темпоральными элементами, полученные на предыдущем этапе.

Методом для получения такого отношения порядка обычно является логический вывод в некоторой его форме. Поскольку мы имеем дело с естественным языком, выражения которого, в общем случае, являются неполными, неточными, предположительными и т.п., то в данном случае обычно применяют некоторую модификацию классической логики первого порядка, учитывающую, в определенной степени, такие неполноту и неточность. Одним из популярных формализмов такого рода, используемым в задачах извлечения информации из ЕЯ-текстов, являются логические сети Маркова (ЛСМ) [7]. Неформально, ЛСМ являются вероятностным расширением логики первого порядка. С формальной точки зрения, ЛСМ есть множество формул логики первого порядка, снабженных весами, которые представляют собой просто вещественные числа.

Например, утверждение на естественном языке

«Курение вызывает рак легких»

в логике первого порядка представляется предложением

$$\forall x(Sm(x) \implies Ca(x))$$

или, в эквивалентной форме (в которой внешние кванторы всеобщности опущены)

$$\neg Sm(x) \vee Ca(x)$$

где предикат $Sm(x)$ обозначает, что индивид x курит, а предикат $Ca(x)$ обозначает, что индивид x более раком легких). В ЛСМ, к последнему виду формулы просто добавляется весовой коэффициент:

$$\neg Sm(x) \vee Ca(x), \qquad 1.5$$

Получение весов для формул реализуется с помощью методов машинного обучения с использованием корпусов лингвистических данных. Так, например, в системе TIE [16], в качестве реализации ЛСМ, используется система Alchemy [17]. Обучение системы Alchemy реализовано с использованием банка аннотированных темпоральных выражений TimeBank [8].

Соответственно, для вывода отношений упорядочения на темпоральных элементах на основе ранее полученных синтаксических зависимостей, ЛСМ будет содержать, в частности, следующие типичные формулы:

(1) $dependency(e_1, e_2) \implies after(point(e_1), point(e_2))$;

это правило означает, что если для темпоральных элементов e_1 и e_2 имеется синтаксическая зависимость, то временная точка $point(e_1)$ идет после временной точки $point(e_2)$, где выражение $point(e)$ есть либо время начала, либо время окончания темпорального элемента e; конкретные значения выражений $point(e_1)$, $point(e_2)$ определяет конкретный вид синтаксической зависимости $dependency(e_1, e_2)$;

(2) $srl\_after(p, q) \implies after(p, q)$;

данное правило переводит синтаксическую зависимость "after", полученную с использованием семантических ролей, в отношение порядка $after$ для темпоральных элементов p и q;

(3) $after(p, q) \wedge after(q, r) \implies after(p, r)$;

данное правило реализует отношение транзитивности между темпоральными элементами;

Из такого вида правил, отражающих связи между синтаксическими зависимостями и отношениями упорядочения на темпоральных элементах, строится так называемая базовая сеть Маркова. Сначала в формулы ЛСМ вместо переменных подставляются константы, представляющие собой извлеченные из данного ЕЯ-предложения события и времена. Затем строится сама сеть, узлами которой являются атомарные выражения из конкретизированных формул. В данной сети, пара узлов соединена дугой, если соответствующие узлам атомарные выражения входят в какую-либо формулу из множества формул ЛСМ, полученных на основе обрабатываемого ЕЯ-предложения.

Для базовых сетей Маркова определена специальная процедура вывода [7], которая, в общем случае, позволяет отвечать на вопросы вида

«Какова вероятность истинности формулы F1 при условии истинности формулы F2?»

Таким образом, в результате применения процедуры вывода к формулам, в которые входят отношения упорядочения, получаются новые отношения упорядочения с некоторыми коэффициентами вероятности. Соответственно, для получения окончательного результата об упорядоченности темпоральных элементов, применяется оценка этих отношений с использованием выбранного порога значения вероятности: все отношения порядка с вероятностями, превышающими указанный порог, будут считаться истинными и выдаваться в качестве результата работы всей системы извлечения темпоральной информации из ЕЯ-текстов.

4. Заключение

В данной работе описан некоторый прототип системы извлечения темпоральной информации из ЕЯ-текста — ее базовая архитектура и основные методы и алгоритмы ее работы. Данный прототип ориентирован на главную задачу, решаемую при извлечении темпоральной информации — определении упорядоченности темпоральных элементов (событий и времен) на оси времени. Конечно, для решения других задач, связанных с темпоральной информацией, извлекаемой из ЕЯ-текста, могут понадобиться дополнения и расширения указанных архитектуры, методов и алгоритмов. Исследования в этом направлении являются одним из путей нашей дальнейшей работы.

В данной статье, были выделены три этапа решения общей задачи извлечения темпоральной информации из ЕЯ-текста: извлечение событий и времен, установление (синтаксических) зависимостей между ними и их упорядочение во времени. Для решения каждой из этих подзадач, ориентированных на тексты на русском языке, понадобятся соответствующие компоненты: синтаксический анализатор, система определения семантических ролей и др. Изучение возможностей использования уже существующих компонент с соответствующим функционалом одновременно с разработкой собственных — также есть одно из направлений нашей будущей работы.

В заключение отметим, что, в отличие от английского языка, для которого уже разработаны очень мощные и свободно доступные упомянутые компоненты (в том числе, большое количество разнообразных лингвистических банков данных), для русского языка, состав и степень развитости соответствующих компонентов, на настоящий момент, очень ограничены.

Список литературы

[1] S. Brin, “Extracting patterns and relations from the World Wide Web”, *WebDB Workshop at 6th International Conference on Extending Database Technology*, EDBT’98, pp. 172–183. ↑401

[2] D. E. Appelt. “Introduction to information extraction”, *Journal AI Communications*, **12**:3 (1999), pp. 161–172. ↑401,409

[3] R. Sauri, R. Knippen, M. Verhagen, J. Pustejovsky, “Evita: A Robust Event Recognizer for QA Systems”, *Proceedings of HLT/EMNLP 2005*, pp. 700–707. ↑403,405,407

[4] A. X. Chang, C. D. Manning. “SUTime: a Library for Recognizing and Normalizing Time Expressions”, 8th International Conference on Language Resources and Evaluation LREC 2012, pp. 23–25. ↑403,407,409

[5] N. UzZaman, J. F. Allen, “TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text”, *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval’10, pp. 276–283. ↑403,409

[6] 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*, eds. I. Mani, J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, pp. 545–557. ↑403,406,408

[7] M. Richardson, P. Domingos. “Markov logic networks”, *Machine Learning*, **62**:1 (2006), pp. 107–136. ↑404,412,413

[8] J. Pustejovsky et al., “The TIME-BANK Corpus”, *Proceedings of Corpus Linguistics 2003*, pp. 647–656. ↑404,407,412

[9] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998. ↑407

[10] G. A. Miller et al., “Using a semantic concordance for sense identification”, *Proceedings of ARPA Human Language Technology Workshop 1994*, pp. 240–243. ↑407

[11] I. Mani, G. Wilson, “Processing of News”, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, ACL2000, pp. 69–76. ↑409

[12] J. Strötgen, M. Gertz, “HeidelTime: High quality rule-based extraction and normalization of temporal expressions”, *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval’10, 2010, pp. 321–324. ↑409

[13] D. Chen, C. D. Manning, “A Fast and Accurate Dependency Parser using Neural Networks”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015 (25–29 October 2014, Doha, Qatar), pp. 740–750. ↑410

[14] M.-C. de Marneffe, B. MacCartney, C. D. Manning. “Generating Typed Dependency Parses From Phrase Structure Parses”, 5th International Conference on Language Resources and Evaluation LREC 2006, pp. 449–454. ↑410

[15] P. Koomen, V. Punyakanok, D. Roth, W. Yih, “Generalized inference with multiple semantic role labeling systems”, *Proceedings of the Computational Natural Language Learning Conference*, CoNLL’05, pp. 181–184 (shared task paper). ↑411

[16] X. Ling, D. S. Weld, “Temporal Information Extraction”, *Proceedings of the 25th National Conference on Artificial Intelligence*, AAAI 2010, pp. 1385–1390. ↑412

[17] S. Kok et al.. *The Alchemy system for statistical relational AI*, Technical Report 2(6), Department of Computer Science and Engineering, University of Washington, 2005. ↑412

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

Об авторе:

Юрий Петрович Сердюк

Старший научный сотрудник ИПС РАН. В область научных интересов входят математическая искусственный интеллект, математическая логика, включая темпоральную и модальную логики, а также теоретические вопросы параллельного программирования

e-mail: yury@serdyuk.botik.ru

Пример ссылки на эту публикацию:

Ю. П. Сердюк. «Базовая архитектура, методы и алгоритмы системы извлечения темпоральной информации из текстов на естественном языке», *Программные системы: теория и приложения*, 2015, **6**:4(27), с. 401–418.

URL: http://psta.psiras.ru/read/psta2015_4_401-418.pdf

Yury Serdyuk. *Basic architecture, methods and algorithms of system for temporal information extraction from natural language texts.*

Abstract. The given paper presents a basic architecture of the system which intended for temporal information extraction from natural language texts. We present the main structural parts of this architecture along with the corresponding methods and algorithms. Also we distinguish three stages in the temporal information extraction: recognizing the events and times, identifying the dependencies between them and final ordering of temporal elements. Correspondingly, logical inference, machine learning methods and linguistics banks of data are necessary in all these stages. (*In Russian*).

Key Words and Phrases: Information extraction, temporal elements, machine learning.

References

[1] S. Brin, "Extracting patterns and relations from the World Wide Web", *WebDB Workshop at 6th International Conference on Extending Database Technology*, EDBT'98, pp. 172–183.

[2] D. E. Appelt. "Introduction to information extraction", *Journal AI Communications*, **12**:3 (1999), pp. 161–172.

[3] R. Sauri, R. Knippen, M. Verhagen, J. Pustejovsky, "Evita: A Robust Event Recognizer for QA Systems", *Proceedings of HLT/EMNLP 2005*, pp. 700–707.

[4] A. X. Chang, C. D. Manning. "SUTime: a Library for Recognizing and Normalizing Time Expressions", 8th International Conference on Language Resources and Evaluation LREC 2012, pp. 23–25.

[5] N. UzZaman, J. F. Allen, "TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text", *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval'10, pp. 276–283.

[6] 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*, eds. I. Mani, J. Pustejovsky, R. Gaizauskas, Oxford University Press, Oxford, 2005, pp. 545–557.

[7] M. Richardson, P. Domingos. "Markov logic networks", *Machine Learning*, **62**:1 (2006), pp. 107–136.

[8] J. Pustejovsky et al., "The TIME-BANK Corpus", *Proceedings of Corpus Linguistics 2003*, pp. 647–656.

[9] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998.

[10] G. A. Miller et al., "Using a semantic concordance for sense identification", *Proceedings of ARPA Human Language Technology Workshop 1994*, pp. 240–243.

[11] I. Mani, G. Wilson, "Processing of News", *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, ACL2000, pp. 69–76.

[12] J. Strötgen, M. Gertz, "HeidelTime: High quality rule-based extraction and normalization of temporal expressions", *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval'10, 2010, pp. 321–324.

[13] D. Chen, C. D. Manning, “A Fast and Accurate Dependency Parser using Neural Networks”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015 (25–29 October 2014, Doha, Qatar), pp. 740–750.

[14] M.-C. de Marneffe, B. MacCartney, C. D. Manning. “Generating Typed Dependency Parses From Phrase Structure Parses”, 5^{th} International Conference on Language Resources and Evaluation LREC 2006, pp. 449–454.

[15] P. Koomen, V. Punyakanok, D. Roth, W. Yih, “Generalized inference with multiple semantic role labeling systems”, *Proceedings of the Computational Natural Language Learning Conference*, CoNLL’05, pp. 181–184 (shared task paper).

[16] X. Ling, D. S. Weld, “Temporal Information Extraction”, *Proceedings of the* 25^{th} *National Conference on Artificial Intelligence*, AAAI 2010, pp. 1385–1390.

[17] S. Kok et al.. *The Alchemy system for statistical relational AI*, Technical Report 2(6), Department of Computer Science and Engineering, University of Washington, 2005.

Sample citation of this publication:

Yury Serdyuk. “Basic architecture, methods and algorithms of system for temporal information extraction from natural language texts”, *Program systems: theory and applications*, 2015, **6**:4(27), pp. 401–418. (*In Russian*).

URL: http://psta.psiras.ru/read/psta2015_4_401-418.pdf