УДК 681.3

АНАЛИЗ ТЕКСТОВЫХ СООБЩЕНИЙ В СИСТЕМАХ МОНИТОРИНГА ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

И. С. Лебедев,
канд. техн. наук, доцент
Санкт-Петербургский государственный университет
Ю. Б. Борисов,
аспирант
Санкт-Петербургский государственный университет информационных технологий, механики и оптики

Описываются модели формализации естественно-языковых сообщений для систем мониторинга информационной безопасности открытых вычислительных сетей. Рассматриваются особенности обработки и анализа сообщений.

Ключевые слова *— формализация естественного языка, обработка сообщений, вычисление информационных структур.*

Введение

В условиях социальных преобразований, происходящих в мире, возникает необходимость непрерывного наблюдения за различными информационными событиями. Интеграция глобальных вычислительных сетей в огромное количество сфер деятельности человека обуславливает появление информационных ресурсов, отражающих политические, социальные, экономические новости. Сообщения блогеров, комментаторов лент новостных агентств и порталов, участников «Живого Журнала» содержат информацию о личных отношениях к происходящему в общественной жизни. Вследствие чего возникает задача автоматизированной обработки информации с целью определить и проанализировать политический, социальный, экономический спектр мнений.

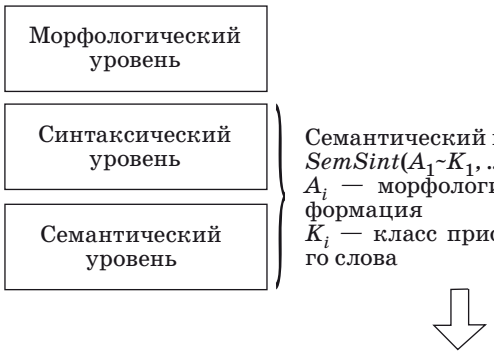
Существующая легкость использования информационного пространства, предоставляемого глобальными вычислительными сетями, участившееся применение различных ресурсов сети Интернет для проведения всевозможных PR-акций, информационных компаний, направленных на решение политических, экономических, идеологических задач, наносит определенный урон хозяйствующим субъектам и требует анализа огромного количества текстов для выявления внешних и внутренних источников информационных угроз.

Однако сложность методов, позволяющих в автоматическом режиме идентифицировать структуру и значение обрабатываемых естественно-языковых сообщений, заставляет производить их обработку с применением «ручных» технологий [1]. Вместе с тем высокая степень интеграции и использования ПЭВМ наряду с внедрением информационных технологий дает возможность разрабатывать и реализовывать в информационных системах более эффективные методы и алгоритмы вычисления слабоструктурированных данных [2].

Формализации естественно-языковых конструкций

Аналитические модели описания естественного языка (ЕЯ) в большинстве случаев являются узкоспециализированными и сложными с точки зрения адаптации под конкретные виды задач обработки текстовой информации открытых компьютерных сетей. Для повышения качества обработки документов на ЕЯ в предметной области обнаружения информационных угроз необходимо решить вопрос о формализации семантической составляющей.

Одним из подходов, который может быть применен для обработки относительно коротких текстовых сообщений, является семантическая модель ЕЯ профессора СПбГУ В. А. Тузова [3]. В ней выделяется три уровня: морфологический, семантико-синтаксический, семантический (рис. 1):



■ ***Рис. 1.*** *Семантическая модель языка В. А. Тузова*

$$M = <W, Se, K>, \quad (1)$$

где W — множество словоформ; Se — множество семантических шаблонов; K — множество классов.

Особенностью предложенной В. А. Тузовым модели ЕЯ является объединенный семантико-синтаксический уровень. Каждое слово обладает морфологическими и семантико-синтаксическими характеристиками, на основе которых строится семантический предикат.

Общий шаблон описания словоформы в словаре Тузова можно представить в следующем виде:

$$W(Z1{:}!\text{Им}\{K_1\}_g, Z2{:}!\text{Род}\{K_2\}_g, Z3{:}!\text{Дат}\{K_3\}_g, Z4{:}!\text{Вин}\{K_4\}_g, Z5{:}!\text{Тв}\{K_5\}_g, Z6{:}!\text{Пред}\{K_6\}_g),$$

где $\{K_1\}_g$ ... $\{K_6\}_g$ — набор классов, соответствующий данной словоформе.

Однако семантический словарь Тузова, применяемые для решения аналогичных задач словари Шведовой, Ефремовой, лингвистические базы данных компаний АОТ, RCO и др. очень сильно отличаются по структуре, количеству классов, числу входящих в них слов. Вследствие чего подобные продукты должны быть подвержены дополнительной адаптации под конкретную задачу анализа текста, связанной с уточнением состава и вида (например, древовидный или линейный) классификатора словоформ. Использование словарных баз данных (БД) в большинстве случаев требует знаний лингвиста и может быть сложным для специалиста в области информационной безопасности, которому необходимо настроить фильтр, осуществляющий контент-анализ текстовых сообщений.

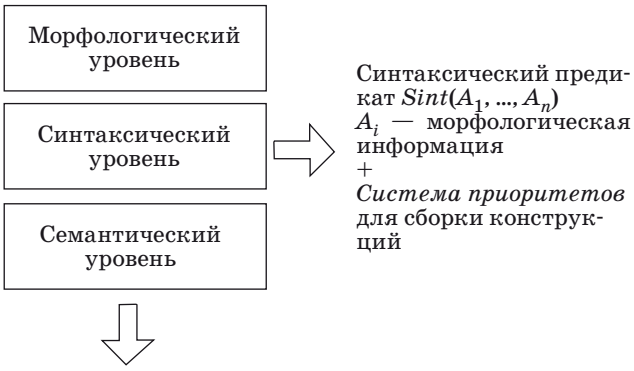
Модель ЕЯ Тузова предполагает возможность анализа любого предложения естественного (русского) языка. Формирование применяемой в ней семантической БД происходило путем автоматизированной обработки различных, в том числе художественных, текстов. Учитывая «произвольный» порядок слов, где, например, образующее связь с существительным прилагательное может быть отделено от него оборотами и находиться в любых частях предложения, для построения структуры ЕЯ-конструкции необходим перебор всех аргументов на предмет вычисления возможности образования связей. С другой стороны, несмотря на поддержку и развитие данной модели определенные сложности при вычислении результата анализа предложения происходят, когда встречаются неоднозначные словоформы, что влияет на построение информационных объектов текста [4, 5]. Реализация систем, базирующихся на приводимой модели, требует значительных затрат на поддержку.

Части недостатков лишена адаптированная модель, предназначенная для поиска определенной тематической информации [6]. В ней, аналогично семантической модели Тузова, также выделяются уровни — морфологии, синтаксиса и семантики. Однако последние отделены друг от друга. Синтаксический уровень содержит информацию о связях между словами, а семантический определяет правила анализа, синтеза и обработки полученных конструкций (рис. 2):

$$M = <W, Si, Ks>, \quad (2)$$

где Si — множество синтаксических шаблонов, $Si \in Se$; Ks — множество классов, $Ks \in K$.

Особенность приводимой модели состоит в использовании масштабируемых предикатов описания информации аргументов словоформ предметно-ориентированных словарных БД ЕЯ, что позволяет осуществлять идентификацию, сравнение конструкций и построение управляющих правил обработки на уровне связей.



■ ***Рис. 2.*** *Адаптированная модель языка*

Масштабируемый предикат по своему составу идентичен семантическому предикату предыдущей модели. Однако вместо семантического класса в нем используются классы идентификационного множества, влияющие на тип и семантическое значение ЕЯ-конструкции в рамках тематики предметной области.

Рассмотрим подход к их построению на основе вычисления структуры предложения и особенности использования.

Вычисление структуры предложения

В нашем случае анализ стилистики текстов блогов, лент новостных агентств показывает почти полное отсутствие «длинных» предложений, которые встречаются у русских классиков. Среднее количество слов в таких сообщениях около 10, что подтверждается данными статистических исследований, опубликованных на сайтах, посвященных классической лингвистике. Прилагательные и уточняющие существительные в родительном и творительном падежах, обороты, идентифицируемые словом «который», причастия не разбросаны по тексту сообщений, а тяготеют к базовым, образующим конструкцию с существительным. Оценка обработки источников текстовой информации сети Интернет может быть осуществлена через подходы, основанные на ошибках первого и второго рода. Для этого словарные БД адаптируют под конкретную предметную область. Ограничения предметной области позволяют избавиться от значительного количества неоднозначных словоформ и использовать для идентификации часто встречающихся последовательностей терминов синтаксический анализатор. Описание одного из решений для синтаксического анализатора можно найти на сайте компании АОТ (www.aot.ru). Принцип действия алгоритма состоит в упорядоченном последовательном переборе около 40 правил.

Однако при анализе текста в системах мониторинга основную часть информации предоставляют существительные. Обнаружение этих частей речи с последующим присоединением к ним подчиненных прилагательных, наречий, причастий позволяет при образовании связи не тратить ресурсы на вычисление типа образовавшейся конструкции. Приводимый алгоритм использует описания словоформ частей речи, основанные на шаблоне, содержащем синтаксическую информацию о потенциальных связях:

$$W(Z1{:}!\text{Им}, Z2{:}!\text{Род}, Z3{:}!\text{Дат}, Z4{:}!\text{Вин}, Z5{:}!\text{Тв}, Z6{:}!\text{Пред}).$$

В предикате конкретной словоформы лишние связи удаляются. Например, для подавляющего числа существительных синтаксический шаблон будет выглядеть следующим образом:

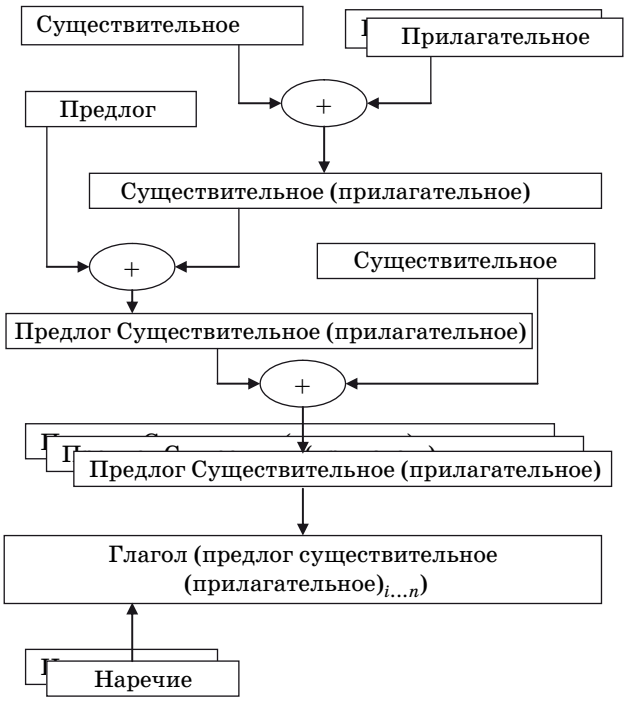
$$W(Z1{:}!\text{Род}).$$

Типовые шаблоны частей речи, особенности их использования приведены в работе [6]. Наибольший приоритет отдается анализу возможности образования связей между двумя ближайшими словоформами.

Рассмотрим упрощенный алгоритм свертки предложения, не акцентируя внимание на таких частях речи и предложения, как числительные, союзы, частицы, причастия, деепричастия, подчиненные предложения. В простом распространенном предложении могут содержаться (или не содержаться) следующие части речи: глаголы, существительные, прилагательные, наречия. На рис. 3 показана последовательность шагов свертки предложения.

1. Присоединение подчиненных прилагательных к существительным.

На этом шаге основная информация берется из морфологического описателя словоформы. При первом просмотре предложения слева направо ищутся ближайшие, согласующиеся по падежу, роду и числу, прилагательные и существительные. Так как прилагательное может находиться справа от существительного, то необходим аналогичный просмотр справа налево, на котором осуществляется попытка присоединения



■ ***Рис. 3.*** *Упрощенный алгоритм свертки предложения*

оставшихся прилагательных, не вошедших в конструкцию.

Ввиду ограниченности объема не будем останавливаться на отдельных ситуациях, когда прилагательные не согласуются по морфологической информации со своими существительными, например:

Средства и методы — проверенные.

Подобных ситуаций конечное количество, и они поддаются довольно строгому описанию и формализации.

2. Присоединение предлогов к конструкциям существительных и прилагательных. Особенностью шага является то, что предлог всегда находится слева от конструкции существительного. Основная информация для реализации свертки — это синтаксический описатель предлога и морфологический описатель конструкции существительного. Информация по предлогу содержит падеж и класс присоединяемого существительного.

3. Присоединение конструкций существительных к другим объектам осуществляется на основании анализа синтаксического описателя левой конструкции и морфологического и синтаксического описателя правой конструкции. Производится слева направо. Вне зависимости от описаний объекты существительных в родительном падеже присоединяются к конструкциям, стоящим слева.

4. Все созданные конструкции подставляются в предикат глагольной функции на основании своей синтаксической информации.

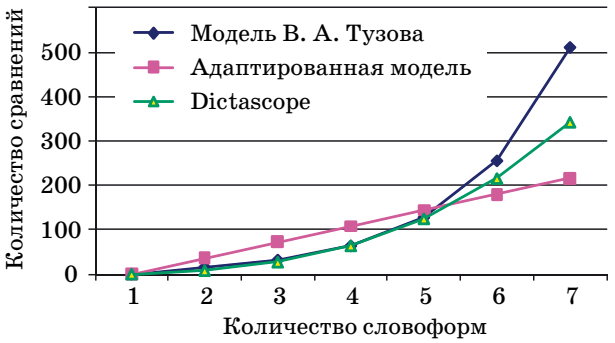
5. Наречия и собранные конструкции, не вошедшие в описатель глагола, приписываются к нему со своим семантико-грамматическим типом.

Следует отметить, что русский язык является довольно регулярным и исключения из правил составляют не более 10 %.

Причастные, деепричастные обороты, подчиненные предложения, начинающиеся со слова *который*, составные конструкции типа *если ... то*, вложенные предложения отделяются перед анализом. Над ними выполняются действия алгоритма свертки, а затем полученные конструкции присоединяются к основному предложению.

В зависимости от стилистических особенностей текстов предметной области, при отсутствии грамматических ошибок синтаксический анализатор выдает 60–80 % адекватных структур.

Первоначальное получение структуры и наложение на нее семантической информации позволяет уменьшить вычислительную сложность и избавиться от лавинообразного роста зависимости количества анализа связей от количества словоформ конструкций (рис. 4). (Оценка модели *Dictascope* приводится согласно публикациям [7, 8].)



■ ***Рис. 4.*** *Зависимость количества проверок связей от количества словоформ*

Полученная структура является основой для вычисления идентификационного множества классов предикатов адаптированной модели.

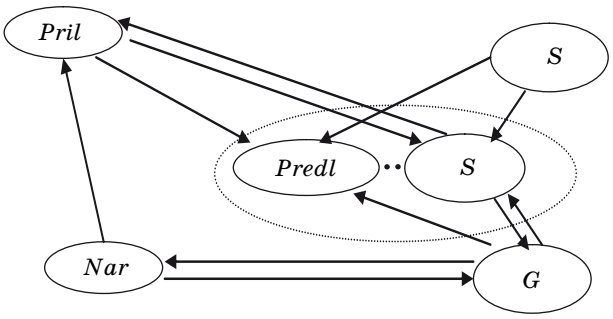
Построение идентификационного множества классов аргументов

Для реализации анализа текстовой информации в системе мониторинга необходимо изначально настроить идентификационное множество классов $k_1 \ldots k_n$ в БД с позиции тематики обрабатываемого текста. Для этого применимы анализаторы различных разработчиков. В результате обработки синтаксическим анализатором предложение приобретает вид функциональной записи, содержащей структуру и связи между его конструкциями:

$$F(w_i \rightarrow \{s\}_i), \qquad (3)$$

где w_i — слова в предложении, каждому из которых соответствует свой набор связей $\{s\}_i$ с другими словами.

Структура, представленная на рис. 5, позволяет формализовать связи, которые образуют другие части речи относительно предложно-падежной формы существительного. Вершины этого



■ ***Рис. 5.*** *Связи между частями речи относительно предложно-падежной формы существительного*

графа составляют глагол G, прилагательное Pril, предлог Predl, существительное S, наречие Nar. Каждая стрелка в графе определена совокупностью вопросов, которую можно задать от различных частей речи к предложно-падежной форме существительного или от нее.

Первая группа — падежные вопросы. Она практически однозначно определяется предложно-падежной формой и поддается формализации на уровне синтаксического шаблона. Вторая группа — смысловые вопросы. Для их формализации требуется классификатор существительных, описывающих семантическую принадлежность.

Прогон тематических текстов через синтаксический анализатор позволяет построить информационные структуры и провести их статистический анализ на предмет вычисления термов предметной области. Частота встречаемости слова, содержащие его лексические конструкции дают информацию для построения классификатора, уточнения синонимов. Особенностью подхода является то, что в основу классификатора может быть положен синтаксический анализатор и словарная БД стороннего разработчика.

Следующий этап — создание предикативного описания словоформ, базирующегося на «новом» классификаторе.

В случае, когда полученным классам можно поставить в соответствие, например, классификатор Тузова, возникает возможность доработать описание его словаря. Таким образом, при поиске, например, текстов экстремистской направленности значение слова МОЧИТЬ в его словаре необходимо преобразовать в два предиката:

МОЧИТЬ N%~МОКРЫЙ$12/113/15(Z1: ДОЖДЬ$122153\
ЖИВОЙ$124~!Им,Z2: !Тв,Z3: !Вин,Z4: НЕЧТО$1~!вПред,Z5:
НЕЧТО$1~!До)

→

МОЧИТЬ *G*(*Z*1:!Им, *Z*2:!Род, *Z*3:!Дат, *Z*4:!Вин, *Z*5:!Тв, *Z*6:!Пред)
МОЧИТЬ N%~УБИВАТЬ$1010 (*Z*1*:* ЖИВОЙ$348.352~!Им, *Z*2:!Род, *Z*3:!Дат, *Z*4: ЖИВОЙ$348.352~!Вин, *Z*5:!Тв, *Z*6:!Пред)

Конструкция, основанная на втором предикате, будет нести для системы мониторинга больше информации, чем конструкция, использующая первый предикат. Возможность подставить первый аргумент второго предиката определяется морфологической информацией и принадлежностью к классу ЖИВОЙ, участвующей в образовании связи словоформы. С другой стороны, класс ЖИВОЙ$124 в исходном предикате является «очень общим» для конкретной задачи. Для уменьшения вероятности ложной тревоги необходимо убрать часть подклассов в словаре Тузова (например, «*животные*», «*растения*») и детализировать подклассы, описывающие значения «*человек*», «*соц.группа*» и т. д.

Таким образом, адаптированная модель ЕЯ использует в описаниях словоформ масштабируемые предикаты связей, аргументы которых содержат информацию о морфологических характеристиках и классах идентификаторов присоединяемых слов, что позволяет унифицировать описания, упростить их структуру.

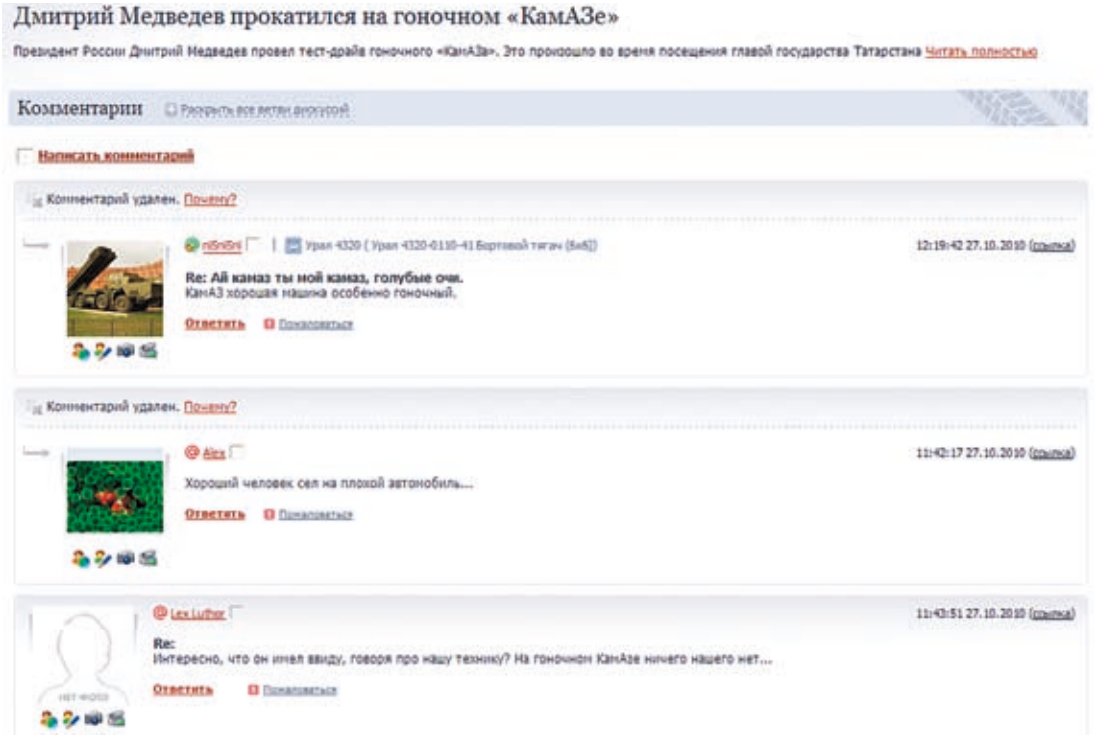
Анализ сообщений блогов и комментариев

Обеспечение экономической, социальной, политической безопасности обуславливает необходимость аудита информационного поля, одной из задач которого является анализ реакции пользователей на различные события.

Современные системы обработки комментариев направлены на получение эмоциональной оценки сообщений. Для этого применяются подходы, основанные на статистическом анализе, в котором словоформы сообщений сопоставляются с семантическими шкалами, например *хорошо-плохо*. Каждому слову такой шкалы ставится в соответствие числовое значение. Количество словоформ семантической шкалы в комментариях позволяет оценить общий эмоциональный фон. Однако в процессе ведения «дискуссий» часть идентификаторов может относиться не к обсуждаемому событию, а к другим объектам. Например, второй комментарий на рис. 6 показывает, что прилагательное *хороший* относится к существительному *человек*, а прилагательное *плохой* определяет существительное *автомобиль*. В случае простого наложения шкалы *хорошо-плохо* приводимые словоформы, характеризующие эмоциональную окраску, будут влиять друг на друга. Если построить структуру ЕЯ-конструкции, то становится очевидным, что определяются различные информационные объекты.

Учитывая стиль и особенности написания комментариев в сети Интернет, заключающиеся в использовании специфических выражений, синтаксических ошибках при построении фраз и предложений, необходимо отметить, что в автоматическом режиме не всегда удается построить адекватную структуру анализируемого сообщения [9]. В этом случае необходим универсальный подход к созданию конструкций ЕЯ на уровне синтаксических связей. В данной задаче обработка информации может основываться на вычислении трех видов элементов: *объектов, характеристик* и *действий* [10].

Поэтому модель, ложащуюся в основу получаемой информационной структуры, можно описать следующим образом:

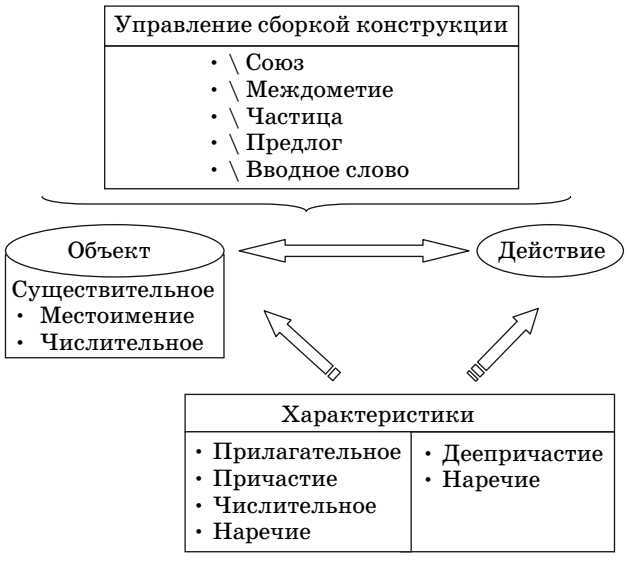


■ ***Рис. 6.*** *Пример комментариев сети Интернет*

$$M = <W, H>, \quad (4)$$

где H — характеристики: $H = \{O|D|C\}$, здесь O — объект; D — действие; $C = \{C_o, C_d\}$ — словоформы, характеризующие объекты (C_o) и действия (C_d).

Универсальная структура представления ЕЯ на примере русского (рис. 7) состоит из объектов, действий, характеристик и слов, осуществляющих управление сборкой конструкции.



■ ***Рис. 7.*** *Универсальная структура представления ЕЯ*

Если рассмотреть простое распространенное предложение на любом ЕЯ, то можно сопоставить полученные морфологические идентификаторы согласно описанной ниже системе.

1. Объекты предложения — существительные.

2. Действие — глагол со своей группой, которая определяется структурой графа предложения.

3.1. Характеристики объектов — прилагательные, причастия, наречия, подчиненные существительные.

3.2. Характеристики действий — наречия, деепричастия.

4. Управляющие слова — простые и составные предлоги, союзы, знаки препинания.

Подготовительный этап для простейшего алгоритма создания структуры информационных объектов предложения на основе морфологического анализа состоит из следующих шагов:

1) поиск объектов предложения;

2) поиск управляющих слов;

3) поиск ближайших характеристик объектов предложения;

4) проверка на возможность образования групп объектов;

5) определение действий;

6) поиск характеристик действий.

Для реализации алгоритма необходимо точно определить роль словоформы в предложении и создать систему приоритетов выбора последовательности частей речи.

Задача, решаемая с помощью данной модели, состоит в том, чтобы при обработке текстов сообщений с неправильным синтаксисом постараться получить отдельные связанные ЕЯ-конструкции, на основе которых определить информационный объект, его характеристики, свойства и действия. Модель является упрощением предыдущих, описанных в статье, ее достоинство заключается в том, что предложенный подход по созданию структуры универсален для большинства ЕЯ, быстро реализуем без существенных затрат на морфологическом и синтаксическом уровне.

В практической реализации данная модель применена в рамках задач мониторинга и создания рейтинга высказываний по событиям, обсуждаемым в сети Интернет.

Заключение

Подход к выбору аналитических моделей представления естественного языка в системах мониторинга, обрабатывающих ЕЯ-сообщения, основывается на обеспечении требуемых характеристик (адекватности, полноты, точности) представления и отражения текстовой информации в базы данных и базы знаний.

Задачи обработки текстовой информации, стилистические особенности документов позволяют определить уровни формализации моделей представления ЕЯ и систематизировать совокупность требуемых характеристик.

Анализ стилистических особенностей обрабатываемой текстовой информации предметной области при мониторинге сообщений позволяет упростить структуру и сложность применения внешних и внутренних управляющих правил, обеспечивающих построение ЕЯ-конструкции.

Степень детализации свойств вычисляемой информации зависит от структуры представления предметной области в БД информационной системы.

Литература

1. **Боярский К. К., Каневский Е. А., Лезин Г. В.** Концептуальные модели в базах знаний // Научно-технический вестник СПбГИТМО (ТУ). Вып. 6. Информационные, вычислительные и управляющие системы. 2002. С. 57–62.
2. **Ермаков А. Е., Плешко В. В.** Семантическая интерпретация в системах компьютерного анализа текста // Информационные технологии. 2009. № 6. С. 2–7.
3. **Тузов В. А.** Компьютерная семантика русского языка. — СПб.: Изд-во СПбГУ, 2004. — 400 с.
4. **Леонтьева Н. Н.** Роль связей в семантической разметке корпуса текстов // Тр. Междунар. конф. «Корпусная лингвистика — 2004». СПб.: Изд-во СПбГУ, 2004. С. 195–206.
5. **Лебедев И. С.** Способ формализации связей в конструкциях текста при создании естественно-языковых интерфейсов // Информационно-управляющие системы. 2007. № 3. С. 23–26.
6. **Лебедев И. С.** Построение семантически связанных информационных объектов текста // Прикладная информатика. 2007. № 5 (11). С. 83–89.
7. **Разработка** пилотной версии системы синтаксического анализа русского языка: Отчет о НИОКР/ ВНТИЦ; Руководитель работы В. В. Окатьев; Инв. № 02200803750. СПб., 2008. http://www.vntic.org.ru (дата обращения: 15.11.2010).
8. **Окатьев В. В., Ерехинская Т. Н., Скатов Д. С.** Модели и методы учета пунктуации при синтаксическом анализе предложений русского языка // Материалы Междунар. конф. «Диалог 2009», Бекасово, 27–31 мая 2009 г. М.: РГГУ, 2009. Вып. 8 (15). С. 423–429.
9. **Ронжин А. Л.** Особенности автоматического распознавания разговорной русской речи // Анализ разговорной русской речи: Тр. первого междисциплинарного семинара АРЗ-2007. СПб.: ГУАП, 2007. С. 42–55.
10. **Лебедев И. С.** Построение шаблонов кода по текстам спецификаций // Информационно-управляющие системы. 2009. № 5. С. 39–43.