УДК:519.767.6
Андреев А.М., Березкин Д.В., Симаков К.В.

Модель извлечения знаний из естественно-языковых текстов

ВВЕДЕНИЕ

Первым применением систем, извлекающих знания из естественно-языковых текстов, было выявление фактов в доступных on-line текстах. Извлекаемые факты представляли собой структурированное описание событий и явлений, излагаемых в анализируемых текстах, и имели конкретную тематическую направленность. Поскольку наибольший интерес к таким системам проявляли американские спецслужбы, то наиболее распространенными темами извлекаемых фактов были теракты, несчастные случаи и массовые беспорядки, имевшие место в горячих точках планеты [1], [2]. Кроме мониторинга содержимого Web, аналогичной обработке могла подвергаться и электронная почта на предмет выявления и предотвращения готовящихся преступлений. Извлекаемые факты состояли из структурных элементов, описывающих, например, участников события, их цели и средства, а также место события, его причины и последствия. Факты, как правило, заносились в базу данных, по анализу которой выявлялись закономерности между событиями, пополнялись списки участников однотипных событий, а также списки целей и средств.

В данной работе изложена модель извлечения, применение которой может быть найдено как в вышеперечисленных задачах, так и при решении более сложных проблем анализа текстов. Одним из таких применений является система выявления несоответствий и противоречий в документах. Концепция создания такой системы изложена в [4], а часть результатов, полученных при разработке этой системы, опубликована в [5] и [6]. В данной работе более детально изложены теоретические аспекты разработанной модели извлечения.

ПРЕДОПОСЫЛКИ К РАЗРАБОТКЕ МОДЕЛИ ИЗВЛЕЧЕНИЯ

Проведенный анализ существующих подходов к извлечению знаний из текстов ([7],[8],[9],[10] и др.) выявил повышенный интерес зарубежных исследователей к данной теме. Наиболее популярной является серия конференций MUC, проводимая в целях совершенствования методов компьютерной разведки. В связи с этим большинство существующих моделей и методов извлечения ориентированы на языки западной Европы и на восточные языки. В современных отечественных публикациях проблема извлечения знаний практически не затрагивается. В настоящий момент авторам известна лишь одна разработка [3], нацеленная на извлечение знаний из русскоязычных текстов и доведенная до практического использования. Недостатком [3] является отсутствие возможности обучения при составлении правил извлечения и как следствие - сложность адаптации к произвольным предметным областям.

Именно поэтому в данной работе была поставлена цель – разработать обучаемую модель извлечения, применимую для извлечения знаний из русскоязычных текстов. Для этого было необходимо учесть большой набор морфологических признаков, присущий многим русским словам, а также учесть неоднозначность морфологического разбора. Для обеспечения практической применимости в процедуре извлечения не должны использоваться сложные вычисления, поэтому было принято решение отказаться от использования результатов синтаксического анализа предложений.

ФРЕЙМОВОЕ ПРЕДСТАВЛЕНИЕ ЗНАНИЙ

Фреймы являются наиболее распространенным способом представления в задачах извлечения знаний [11]. Фрейм рассматривается как структура, с поименованными элементами – слотами. Фреймы и слоты наделяются семантикой, в зависимости от предметной области, в которой они используются [12]. Декларативную часть знаний о предметной области разделяют на две группы: аксиоматическую и фактическую. Аксиоматика для фреймового представления задает его схему, т.е. определяет состав каждого фрейма и об-

ласть значений слотов. В данной работе полагается, что она задается человеком-экспертом. Факты с точки зрения фреймовой модели – это конкретные экземпляры фреймов с означенными слотами. Экземпляры фреймового представления можно представить в виде алгебраической модели вида

$$FM = < F_e, S_e, T, R_{FST} >, \quad (1)$$

где F_e – множество экземпляров фреймов; S_e – множество слотов экземпляров фреймов; T – множество значений слотов; $R_{FST} \subseteq F_e \times S_e \times T$ - отношение, связывающее с экземпляром фрейма конкретный слот и его значение.

Извлечению подлежат экземпляры фреймов *FM*, полагая, что значения слотов из T представимы в виде последовательностей слов естественно-языковых текстов. Таким образом, задача извлечения знаний сводится к выявлению в анализируемом тексте некоторого экземпляра фрейма f_e и означиванию его слотов $\{s_e\}$ фрагментами из текста. Результатом извлечения являются тройки $(f_e, s_e, t_e) \in R_{FST}$, где t_e – значение слота выявленного фрейма, представленное текстовым сегментом, выделенным в исходном тексте.

МОДЕЛЬ ТЕКСТА

Для поставленной задачи необходимо, чтобы текст был представим в виде последовательностей сегментов. Минимальными элементами сегмента являются слова, представляющие собой последовательности символов алфавита естественного языка, а также знаки препинания. Данная модель текста представима в виде алгебраической системы вида

$$TM = < T, W, t_\varnothing, \bullet >, \qquad (2)$$

где T – множество текстовых сегментов, W – множество слов, $t_\varnothing$ - пустой текстовый сегмент, $\bullet$ - операция сцепления на T. В модели текста определены следующие свойства:

1. $\forall w \in W \Rightarrow w \in T$ - каждое слово является текстовым сегментом.

2. $\forall t_1 \in T \wedge \forall t_2 \in T \; \exists! t = t_1 \bullet t_2 \wedge t \in T$ - операция сцепления позволяет из произвольной пары текстовых сегментов сформировать новый текстовый сегмент.
3. $t_\varnothing \in T \wedge \forall t \in T \Rightarrow t = t_\varnothing \bullet t \wedge t = t \bullet t_\varnothing$ - пустой текстовый сегмент является нейтральным элементом по отношению к операции сцепления.
4. $t_1, t_2 \in T \wedge t_1 \neq t_\varnothing \wedge t_2 \neq t_\varnothing \Rightarrow t_1 \bullet t_2 \neq t_2 \bullet t_1$ - некоммутативность операции сцепления.

На основе приведенных свойств модели можно сделать следующие выводы:

1. $\forall t \in T \Rightarrow t = w_1 \bullet ... \bullet w_n : w_i \in W \wedge w_i \in t$ - любой текстовый сегмент может быть представлен в виде сцепления слов.
2. Поскольку слова являются неделимыми сегментами, то удобно измерять длину сегментов в словах, далее длину сегмента *t* будем обозначать N_t.
3. Длина пустого сегмента $t_\varnothing$ равна нулю, т.е. $N_{t\varnothing} = 0$.

КОМПОНЕНТЫ МОДЕЛИ ИЗВЛЕЧЕНИЯ

Модель извлечения определяется как алгебраическая система вида

$$EM = < V, P, R, p_\varnothing, \circ, <_v >, \qquad (3)$$

где V – множество правил извлечения, P – множество образцов текстовых сегментов, R – множество элементов образцов, $p_\varnothing$ - пустой образец, $\circ$ - операция сцепления образцов, $<_v$ - отношение порядка в контексте правила *v*. Эти компоненты обладают следующими свойствами.

1. $\forall p_1 \in P \wedge \forall p_2 \in P \; \exists p = p_1 \circ p_2$ - операция сцепления образцов позволяет формировать новые образцы на основе существующих.
2. $p_\varnothing \in P \wedge \forall p \in P \Rightarrow p = p_\varnothing \circ p \wedge p = p \circ p_\varnothing$ - пустой образец является нейтральным по отношению к операции сцепления.
3. $\forall r \in R \Rightarrow r \in P$ - каждый элемент образца сам по себе является образцом. По аналогии со словами модели *TM*, элементы образцов являются мини-

мально допустимыми образцами, неделимыми в том смысле, что они не могут быть представлены в виде сцепления других непустых образцов.

4. $\forall v \in V \Rightarrow v = p_b \circ p_c \circ p_a$ - правила извлечения представляют собой сцепление тройки образцов: префиксный, извлекающий и постфиксный.

На основе приведенных свойств модели можно сделать следующие выводы:

1. $\forall p \in P \Rightarrow p \in V$ - любой образец может быть представлен в виде правила. Тривиальным случаем является представление вида $p_\varnothing \circ p \circ p_\varnothing$, т.е. префиксный и постфиксный образцы являются пустыми.
2. $\forall r \in R \Rightarrow r \in V$ - любой элемент образца может быть представлен в виде правила извлечения.
3. $\forall p \in P \Rightarrow p = r_1 \circ ... \circ r_n$ - любой образец может быть представлен в виде сцепления элементов. Поскольку элементы являются неделимыми образцами, длину любого образца удобно измерять в составляющих его элементах, далее длину образца *p* будем обозначать N_p.

Данные следствия позволяют ввести единую для элементов, образцов и правил функцию покрытия $a : T \times V \to \{истина, ложь\}$, которая для любого правила и любого сегмента позволяет ответить на вопрос, покрывает ли данное правило данный сегмент. Чтобы дать интерпретацию функции покрытия для элементов образцов, рассмотрим структуру элемента

$$r_i = < c, e, l_1, l_2 >, \quad (4)$$

где $c \subseteq W$ – лексическое ограничение, $e \subset W$ – исключение лексического ограничения, l_1 и l_2 – минимальная и максимальная длина покрытия элемента. Лексическое ограничение *c* и его исключение *e* определяют множество слов *c\e={w}*, которые могут встречаться в текстовых сегментах T_{ri}*={t}*, покрываемых элементом r_i. Слова *{w}* берутся из множества *W* модели текста (2). Множество *c* определяет допустимые слова в сегментах T_{ri}, тогда как множество *e* задает исключения из *c*, т.е. слова, употребление которых запрещено в сегментах T_{ri}. Минимальная и максимальная длины покрытия l_1 и l_2 определяют допустимый диапазон длин текстовых сегментов T_{ri}. С учетом сказан-

ного функция покрытия для элементов образцов задается следующим образом

$$a(t,r)=\begin{cases} r =< c,e,l_1,l_2 > \\ (\forall w \in t \rightarrow w \in c \wedge w \notin e) \vee t = t_{\varnothing} \\ l_1 \le N_t \le l_2 \end{cases} \quad (5)$$

Таким образом, чтобы элемент *r* покрывал текстовый сегмент *t*, необходимо, чтобы все слова, сцепление которых образует *t*, принадлежали множеству слов, разрешенных лексическим ограничением элемента, не попадали в исключения, а длина текстового сегмента должна находиться в диапазоне $[l_1,l_2]$.

Функция покрытия для образца *p* задается следующим образом

$$a(t,p)=\begin{cases} p = r_1 \circ r_2 \circ ... \circ r_n \\ t = t_1 \bullet t_2 \bullet ... \bullet t_n \\ \forall i = 1..n \Rightarrow a(t_i, r_i) \end{cases} \quad (6)$$

Образец *p*, представленный сцеплением из *n* элементов, покрывает текстовый сегмент *t*, если этот текстовый сегмент может быть представлен в виде сцепления *n* сегментов $t_1 \ldots t_n$, так что t_i-ый сегмент покрывается соответствующим r_i-ым элементом. При этом допускается, что некоторые сегменты из $t_1 \ldots t_n$ могут быть пустыми. Выражение (6) позволяет судить об образцах, как о шаблонах фраз [4], состоящие из элементов, связанных отношением предшествования [5]. Роль образца как шаблона фразы заключается в том, что при извлечении информации из некоторого текста на основе данного образца элементы текста должны следовать друг за другом в том же порядке, в каком следуют друг за другом соответствующие им элементы образца. Т.е. для каждого элемента анализируемого текста должен присутствовать соответствующий ему элемент образца. Тексты, одинаковые по содержанию, но различные по форме написания относятся к разным образцам.

Функция покрытия для правила извлечения задается следующим образом. Правило извлечения *v*, представленное в виде сцепления трех образцов $p_b \circ p_c \circ p_a$, покрывает текстовый сегмент *t*, если этот сегмент представим в

виде сцепления трех сегментов $t_b \bullet t_c \bullet t_a$, каждый из которых покрывается соответствующим образцом правила (7).

$$a(t,v) = \begin{cases} v = p_b \circ p_c \circ p_a \wedge t = t_b \bullet t_c \bullet t_a \\ a(t_b, p_b) \wedge a(t_c, p_c) \wedge a(t_a, p_a) \\ t_c - \textit{результат извлечения} \end{cases} \quad (7)$$

При этом извлечению подлежит только часть t_c исходного сегмента, поскольку она покрывается извлекающим образцом правила. Именно t_c объявляется значением слота некоторого фрейма модели знаний, с которым связано данное правило. Формально связь между правилами и фреймами записывается в виде следующих утверждений:

1. с каждым слотом аксиоматической части фреймовой модели связан набор правил V_s такой, что любой текстовый сегмент, покрываемый одним из правил из V_s, принадлежит области значений данного слота;
2. множества правил извлечения для каждого слота уникальны и не пересекаются между собой.

Поясним семантику правил извлечения на примере. В программной реализации модели используется XML нотация для описания правил. Правило описывается XML-элементом <rule…>, содержащим дочерние элементы с тэгами <ct/> и <ex/>. XML-элементы <ct/> описывают элементы префиксного и постфиксного образцов, XML-элементы <ex/> описывают элементы извлекающего образца. Элементы имеют атрибуты set и len. Синтаксис записи значения атрибута len следующий: len="[l_1;l_2]", где l_1 и l_2 – числа, обозначающие верхнюю и нижнюю границу задаваемого диапазона. Атрибут set имеет синтаксис set="A\B", где A и B – записи, задающие множества лексических ограничений *c* элемента образца и *e* – исключений из *c*. Для $e=\varnothing$ вторая часть в записи set="A\B" отсутствует. Части A и B имеют одинаковый синтаксис, допускающий комбинации из следующих вариантов.

1. Перечисление допустимых к употреблению слов. Запись такого множества имеет вид: "(word$_1$|word$_2$|…|word$_n$)", где word$_i$ – i-ое слово множества.

2. Перечисление концевых буквосочетаний допустимых к употреблению слов. Запись имеет вид: “(*end$_1$|*end$_2$|…|*end$_n$)”, где end$_i$ – i-ое концевое буквосочетание слов множества.

3. Перечисление морфологических признаков слов. К морфологическим признакам относится часть речи и принятые в языке значения грамматических категорий. Для русского языка такими категориями являются падеж, число, род, лицо и др. Запись имеет вид: “{$z_1|z_2|\ldots|z_n$}”, где z_i – i-ый морфологический признак. Признаки связываются логической функцией «И». Таким образом, итоговое множество слов является пересечением множеств, соответствующих указанным в записи морфологическим признакам.

Возьмем в качестве примера текстовые сегменты: *«**Компания** nVidia **официально отложила день** выпуска видеокарты…»* и *«**Фирма** Apple **опровергла слухи** о том…»*. Оба примера представимы в виде $t_b \bullet t_c \bullet t_a$, где подчеркиванием выделены сегменты t_c, состоящие из одного слова и подлежащие извлечению. Значениями целевого слота являются названия компаний. Для первого примера $t_b = компания$, $t_c = nVidia$, $t_a = официально\ отложила\ день$. Для второго примера $t_b = фирма$, $t_c = Apple$, $t_a = опровергла\ слухи$. Запись правила, покрывающего эти примеры, представлена на рис 1. Правило состоит из пяти элементов и представимо в виде $p_b \circ p_c \circ p_a$. Образец p_b состоит из одного элемента **<ct len="[1;1]" set="{И|ЕД}"/>** и покрывает все текстовые сегменты длиной в одно слово, которое должно быть отнесено к категории единственного числа именительного падежа. Для данного примера такими сегментами являются $t_b = компания$ и $t_b = фирма$. Извлекающий образец p_c состоит из одного элемента <ex len="[1;1]" set="{eng}"/>.

```
<rule name=“company_1”>
    <ct len="[1;1]" set="{ И|ЕД }"/>
    <ex len="[1;1]" set="{eng}"/>
    <ct len="[0;1]" set="{нрч}"/>
    <ct len="[1;1]" set="{сов|пхд| глг|ЕД}"/>
    <ct len="[1;1]" set="{В}"/>
</rule>
```

Рис. 1. Пример правила извлечения в XML нотации.

Он покрывает все сегменты, состоящие из одного слова, записанные символами английского алфавита. В данном случае это сегменты: $t_c = nVidia$ и $t_c = Apple$. Постфиксный образец состоит из трех элементов, выделенных жирным шрифтом. Первый элемент **<ct len="[0;1]" set="{нрч}"/>** покрывает сегменты длиной от 0 до 1, слова которых должны относиться к категории наречий. Второй элемент **<ct len="[1;1]" set="{сов|пхд|глг|ЕД}"/>** покрывает однословные сегменты, которые должны относиться к категории переходных глаголов совершенного вида единственного числа. Последний элемент **<ct len="[1;1]" set="{В}"/>** покрывает однословные сегменты, у которого допустимо выделить винительный падеж. Поскольку минимальная длина покрытия первого элемента равна 0, в тексте могут не встречаться сегменты, покрываемые этим элементом. Так для первого примера t_1=«официально», t_2=«отложила», t_3=«день», тогда как для второго примера $t_1=t_\varnothing$, t_2=«опровергла», t_3=«слухи».

РЕШЕТКА ЛЕКСИЧЕСКИХ ОГРАНИЧЕНИЙ

Лексические ограничения элементов образцов и их исключения можно определять по-разному, начиная от поэлементного перечисления слов и заканчивая предопределенными классами. Особенное внимание необходимо уделить семантическим классам. Такая классификация слов может быть доступна во внешних источниках, таких как толковые словари [13] или тезаурусы. Например, во многих словарях в толкованиях слов можно встретить метки, относящие слова к той или иной сфере применения (медицина, юриспруденция, техника и др.)[14], которые можно использовать для более точного задания лексических ограничений. Другим примером источника семантических классов является русскоязычный тезаурус типа WordNet [15], в этом случае в качестве семантических классов можно напрямую использовать синсеты данного тезауруса [16].

В данной работе использовалось следующие способы задания лексических ограничений: определение класса слов с общими морфологическими признаками, определение посредством явного перечисления слов, определе-

ние класса слов с общим концевым буквосочетанием (обычно окончание и концевая часть основы слова).

Единым требованием для всех способов задания лексических ограничений и их исключений является возможность представить все их множество *C* в виде алгебраической решетки (8). Для этого множество *C* должно быть частично упорядоченным и на нем должны быть определены операции наименьшей верхней и наибольшей нижней границы.

$$CL = < C, \leq, \veebar, \barwedge > , \qquad (8)$$

Где $C \subseteq 2^W$ - множество лексических ограничений и их исключений, $\leq$ - отношение частичного нестрогого порядка на *C*, $\veebar$ - операция наименьшей верхней границы, $\barwedge$ - операция наибольшей нижней границы. Наименьшая верхняя граница $c_1 \veebar c_2$ для двух элементов c_1 и c_2 определяется как $\left(c_u = c_1 \veebar c_2\right) \wedge \forall c \in C : c \leq c_u \Rightarrow \left(c \leq c_1 \vee c \leq c_2\right)$. Наибольшая нижняя граница $c_1 \barwedge c_2$ для двух элементов c_1 и c_2 определяется, как $\left(c_l = c_1 \barwedge c_2\right) \wedge \forall c \in C : \left(c \leq c_1 \wedge c \leq c_2\right) \Rightarrow c \leq c_l$.

Для классификации слов с использованием морфологических признаков лексическое ограничение задается как $c = \bigcap_{z \in Z_i} z : \forall z \in Z_i \Rightarrow z \subset W$, где Z_i – множество признаков, используемых в естественном языке для классификации слов, с помощью которых задано лексическое ограничение *c*. Запись $z \subset W$ отражает тот факт, что каждый морфологический признак z описывает целый класс слов, так что морфологические признаки можно рассматривать как подмножества множества всех слов *W*. Отношение частичного порядка задается как $c_1 \leq c_2 \Leftrightarrow Z_2 \subseteq Z_1$. Операции верхней и нижней границ задаются следующим образом $c_1 = \bigcap_{z \in Z_1} z \wedge c_2 = \bigcap_{z \in Z_2} z \Rightarrow \left(c_1 \veebar c_2 = Z_1 \cap Z_2\right) \wedge \left(c_1 \barwedge c_2 = Z_1 \cup Z_2\right)$.

Для способа задания лексических ограничений путем непосредственного перечисления слов имеет место $c = W_i : W_i \subset W$, где W_i - перечисление

слов. Отношение частичного порядка задается как $c_1 \le c_2 \Leftrightarrow W_1 \subseteq W_2$. Операции верхней и нижней границ задаются следующим образом $c_1 = W_1 \wedge c_2 = W_2 \Rightarrow (c_1 \veebar c_2 = W_1 \cup W_2) \wedge (c_1 \barwedge c_2 = W_1 \cap W_2)$.

Требование к представлению множества *C* лексических ограничений и исключений в виде решетки *CL* связано с необходимостью обучения модели извлечения *EM* на примерах. Эта необходимость может быть строго доказана. На рис. 2 приведена диаграмма Хассе решеток для способов задания лексических ограничений при помощи пересечение морфологических признаков и непосредственного перечисления слов.

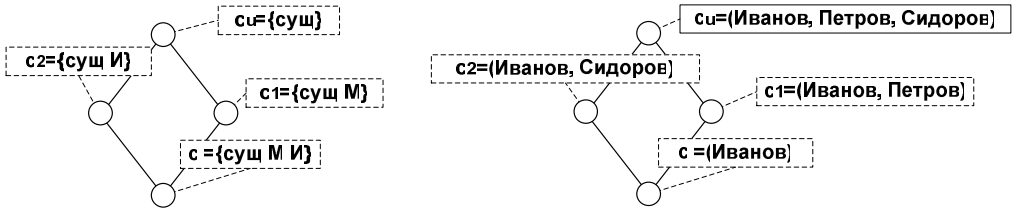

Рис. 2. Пример решеток лексических ограничений.

МЕТОД ИЗВЛЕЧЕНИЯ

Имея модель (3), метод извлечения сводится к поиску для сегмента $t \in T$ подмножества подходящих правил *v*. Проблемой данного подхода является вычислительная сложность, связанная с перебором имеющихся правил $v \in V$ и установлением факта покрытия *a(t,v)*. Проблема усугубляется тем, что один и тот же сегмент *t* может быть покрыт несколькими правилами. В данной работе предложено автоматное представление (9) правил извлечения.

$$EA = (W, Q, q_0, \delta, F), \quad (9)$$

где *W* – алфавит автомата, совпадающий с множеством слов модели извлечения *EM*, так что автомат переходит между состояниями при анализе одного слова анализируемого текстового сегмента, *Q* – множество состояний автомата, q_0 – начальное состояние, δ - функция переходов автомата $\delta : Q \times W \to Q$, *F* – множество конечных состояний.

Правила извлечения связаны с автоматом следующим образом:

1. $Q \subset 2^V : \forall q \in Q \; q \subset V$ - каждое состояние автомата представляет собой подмножество множества правил модели извлечения.

2. $\forall q \in Q \; \exists c \in C \Leftrightarrow c = c(q)$ - с каждым состоянием q автомата связано лексическое ограничение c, обозначаемое как $c(q)$.

3. $\forall v \in q \; \exists r_i = < c_i, e_i, l_1, l_2 > \in v : c(q) = c_i \setminus e_i \Leftrightarrow r_i = r(q,v)$ - каждое правило v, отнесенное к некоторому состоянию q, имеет такой элемент, что разница между лексическим ограничением и исключением этого элемента совпадают с лексическим ограничением $c(q)$ данного состояния. Этот элемент обозначается как $r(q,v)$.

4. $q_0 = V \wedge c(q_0) = W$ - с начальным состоянием автомата связаны все правила извлечения, а лексическое ограничение этого состояния совпадает со всем множеством слов модели извлечения.

Функция переходов обладает следующими свойствами:

1. $\forall w \in W \; \delta(q_i, w) = Q_i$ - при анализе произвольного слова w в общем случае существует множество состояний Q_i, в которые может быть выполнен переход из некоторого состояния q_i.
2. $\forall q_j \in Q_i \Rightarrow w \in c(q_j) \wedge \forall v_j \in q_j \; \exists v_i \in q_i : r(q_i, v_i) <_v r(q_j, v_j) \vee r(q_i, v_i) = r(q_j, v_j)$
   - каждое состояние q_j, в которое можно перейти из q_i по слову w, должно удовлетворять следующим свойствам:
   - слово w принадлежит лексическому ограничению $c(q_j)$ состояния q_j,
   - для каждого правила v_j, отнесенного к состоянию q_j, существует правило v_i, отнесенное к состоянию q_i, так что элементы $r(q_i, v_i)$ и $r(q_j, v_j)$ либо находятся в отношении предшествования в контексте общего правила, либо совпадают. В обоих случаях это означает, что $v_j = v_i$.

Первое свойство функции переходов отражает недетерминированность автомата *EA*, т.е. существуют сегменты, при разборе которых автомат выдаст более одного конечного состояния. В работе [17] предлагается преобразование недетерминированного автомата в детерминированный. В таком случае не-

обходимо гарантировать, что множества всех сегментов, покрываемых каждым правилом модели *EM*, не пересекаются. В общем случае этого сделать нельзя, поскольку эти множества не известны. В данной работе было принято решение использовать недетерминированный автомат. Для снятия неоднозначности, возникающей при анализе сегментов, был разработан критерий специфичности правил, позволяющий из всего множества правил, покрывающих анализируемый сегмент, выбрать наиболее предпочтительное. Данный критерий использует степень конкретности правила (10)

$$S(v,t) = 1 - \boldsymbol{P}(v \mid t) \qquad \textbf{(10)}$$

где $\boldsymbol{P}(v|t)$ – вероятность использовать где-либо правило v, при условии, если достоверно известно, что оно покрывает сегмент t. Таким образом, чем больше условная вероятность применения правила, тем менее оно конкретно. При анализе сегмента t правило v_1 является предпочтительней по отношению к правилу v_2, если степень конкретности правила v_1 больше степени конкретности правила v_2 применительно к этому сегменту.

Вероятность $\boldsymbol{P}(v|t)$ определяется следующим образом. Пусть при анализе сегмента t для выделения покрывающего его правила v автомат проходит последовательность состояний Q_t. Тогда имеет место следующее.

$$P(v \mid t) = \prod_{q \in Q_t} P(q) \qquad P(q) = \sum_{w \in c(q)} p(w) \quad \textbf{(11)}$$

где $\boldsymbol{p}(w)$ – вероятность встретить в тексте слово w. Для оценки $\boldsymbol{p}(w)$ используется приближение на основе количества употребления каждого слова w, отнесенного к общему количеству употребления всех слов в репрезентативной коллекции текстов предметной области $p(w) = \dfrac{N(w)}{\sum_{w_i \in W} N(w_i)}$, где $N(w)$ – количество употреблений слова w в корпусе текстов. Оценка (11) вероятности $\boldsymbol{P}(v|t)$ не является точной, поскольку не учитывает порядок следования элементов внутри правила. Учет порядка следования существенно усложнит модель (11), что приведет к противоположному результату - увеличению вычислительной сложности алгоритма извлечения.

РЕАЛИЗАЦИЯ МОДЕЛИ В СИСТЕМЕ ИЗВЛЕЧЕНИЯ ЗНАНИЙ

Реализация разработанной модели используется в системе извлечения знаний, функционирование которой отражено на рис. 3. Система извлечения выполняет анализ поступающих на вход текстов на предмет выявления в них фреймовых структур. Текст подлежит последовательной обработке модулями системы. Блок сегментирования определяет границы структурных элементов текста: слов, предложений, абзацев и т.д., в его работе используется сегментная модель текста (2).

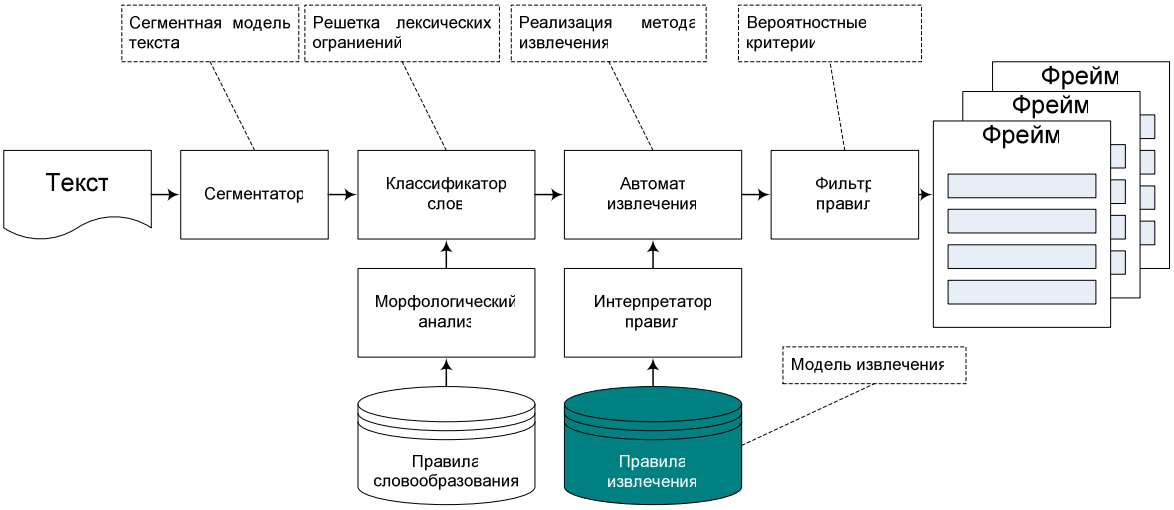

Рис. 3. Функциональная схема системы извлечения знаний.

Блок классификации слов использует решетку лексических ограничений (8) для представления отдельных слов и их морфологических признаков в виде лексических ограничений модели извлечения. Автомат реализует модель извлечения, представленную правилами, хранящимися в виде файлов на внешнем носителе. Финальная сборка фреймов выполняется блоком фильтрации путем отбора наиболее предпочтительных правил, среди конкурирующих за общие текстовые сегменты правил. В качестве критерия фильтрации используется выражение (10) .

ЗАКЛЮЧЕНИЕ

В работе предложена модель извлечения значений слотов фреймовой модели из естественно-языковых текстов. Отличительными особенностями подхода являются: работоспособность модели извлечения при условии неод-

нозначности морфологического разбора и высокая производительность метода извлечения, достигаемая за счет использования простой структуры правил, позволяющей реализовать модель извлечения на основе конечного автомата.

Разработанные методы могут быть использованы в задачах, связанных с обработкой слабоструктурированных текстов. В частности возможно применение при мониторинге потока новостей для извлечения конкретных данных по интересующим событиям. Другой областью применения является наполнение онтологий, когда в качестве источника знаний выступают репрезентативные тексты предметной области.

В настоящий момент разработанные модели и методы используются в Системе семантического контроля текстов документов для поиска несоответствий в текстах стенограмм заседаний Совета Федерации Федерального Собрания Российской Федерации. Модели извлечения в рамках данной системы показывает показатели точности и полноты 0.8 и 0.9 соответственно.

Кроме того, предложенная модель извлечения используется в Интеллектуальной системе исправления ошибок в почтовых адресах клиентов банка, выполняющая в настоящий момент коррекцию в 89% случаев.

ЛИТЕРАТУРА

[1] Kim J.T., Moldovan D.I. PALKA: a system for lexical knowledge acquisition. ACM'1993.

[2] Turmo J., Ageno A., Catala N. Adaptive information extraction. ACM Computing Surveys, Vol. 38, No. 2.

[3] http://www.rco.ru/

[4] Андреев А.М., Березкин Д.В., Симаков К. В. Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. 6-ая Всероссийская научная конференция RCDL'2004.

[5] Андреев А.М., Березкин Д.В., Рымарь В. С. Использование технологии Semantic Web в системе поиска несоответствий в текстах документов. 8-ая Всероссийская научная конференция RCDL'2006.

[6] Андреев А.М., Березкин Д.В., Симаков К. В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения. 8-ая Всероссийская научная конференция RCDL'2006.

[7] Huffman S.B. Learning to extract information from text based on user-provided examples, ACM, 1996.

[8] Dejean H. Learning Rules and Their Exceptions, Journal of Machine Learning Research 2, 2002.

[9] Califf M., Moony R.J. Bottom-Up Relational Learning of Matching Rules for Information Extraction, Journal of Machine Learning Research 4, 2003.

[10] Sakurai S., Suyama A. Rule Discovery from Textural Data based on Key Phrase Patterns, ACM, 2004.

[11] Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. – М.: Радио и связь, 1992.

[12] Hahn A. Knowledge mining from textual sources. ACM'1997.

[13] Bruce R., Guthrie L. Genus Disambiguation: a study in weighted preference. Computation Linguistics, COLING, 1992.

[14] Rigau G., Atserias J., Agirre E., Combining unsupervised lexical knowledge methods for word sense disambiguation, ACM, 1996.

[15] Сухоногов А.М., Яблонский С.А. Разработка русского WordNet. 6-ая Всероссийская научная конференция RCDL'2004.

[16] Claveaue V., Sebillot P. Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming. Journal of Machine Learning Research 4, 2003.

[17] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Архитектура инструментальных средств систем извлечения информации из текстов. Труды международной конференции "Программные системы: теория и приложения", Переславль-Залесский, М.: Физматлит, 2004.