Анализ текстов: лингвистика, семантика, прагматика в рамках когнитивного подхода

Харламов А.А., Ермоленко Т.В.

В работе рассматривается когнитивный подход к анализу лингвистической информации человеком. Рассматриваются процессы обработки информации разных лингвистических уровней: морфологического, синтаксического, семантического для отдельного предложения, и, наконец, семантического и прагматического – для всего текста в целом. Результатом обработки текста на прагматическом уровне является цепочка расширенных предикатных структур предложений текста. Использование когнитивного подхода к анализу текстов продемонстрировано на конкретном примере.

• автоматическая обработка текстов • когнитивный подход • морфологическая обработка • синтаксическая обработка • семантическая обработка • прагматическая обработка • цепочка расширенных предикатных структур

The cognitive approach to the linguistic information analysis of human is considers in this paper. The processes of information processing of different linguistic levels: morphological, lexical, syntactic, semantic for a separate sentence, and finally, semantic and pragmatic for all of the text as a whole are described. The result of text processing on a pragmatic level is an extended predicate structures chain of sentences of the text. There is given an example of cognitive approach using for a specific text.

• automatic processing of text • cognitive approach • morphological processing • syntactic processing • semantic processing • pragmatic processing • extended predicate structures chain.

ВВЕДЕНИЕ

В настоящий момент два основных подхода главенствуют в автоматическом анализе текстов: лингвистический и статистический. Первый дает точный анализ смысла отдельных предложений текста [1], второй – позволяет сформировать семантическое представление целого текста [2]. Они не очень дружно живут вместе – практически не существует работ, описывающих их совместное применение, что объясняется существенным различием механизмов их реализации. В первом случае это – чистая лингвистика, во втором – чистая математика. Тем не менее, их объединение могло бы позволить получить семантические представления целого текста с помощью быстрых алгоритмов статистического анализа с точностью, характерной для лингвистического анализа.

Существует возможность примирения лингвистического и статистического подходов к анализу текстов. Для этого воспользуемся представлениями об обработке информации (в том числе, текстовой) человеком. В двух

словах, обработка специфической информации в мозге человека сводится к накоплению ее в колонках коры полушарий большого мозга [3], и к ее ранжированию в гиппокампе [4]. В колонках коры формируются и хранятся словари образов событий (квази-слов из квази-текстов, в том числе – и обычных текстов) различной частоты встречаемости различных модальностей. В гиппокампе происходит ранжирование этих представлений их весовыми характеристиками, характеризующими значимость этих представлений в рамках отдельных ситуаций (квази-текстов).

1. ОБРАБОТКА ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ ЧЕЛОВЕКОМ

Рассмотрим процедуру структурной обработки информации в коре головного мозга человека, дополненную ранжированием понятий в рамках ситуации, характерной для гиппокампа. Это необходимо для понимания механизма объединения статистического и лингвистического подходов к анализу текстов. Далее мы сможем интерпретировать этапы обработки текстовой информации разных уровней (лингвистической – морфология, лексика, синтаксис), и надлингвистической (семантической и прагматической) в терминах структурного анализа, с естественными переходами с одного уровня обработки на другой уровень.

Нейроны колонок коры в совокупности моделируют многомерное пространство и осуществляют отображение входных сенсорных последовательностей в траектории в этом пространстве [9, 2].

Пусть мы имеем *n*-мерное сигнальное пространство R^n и в нем единичный гиперкуб $G^n \in R^n$. Для дальнейшего изложения введем некоторые обозначения и определения.

Обозначим через $G(n, N)$ – множество последовательностей длины N, элементы которых – точки пространства R^n – являются вершинами единичного гиперкуба G^n. Здесь $G(1, N) \in R^n$ – множество последовательностей длины N (N – произвольное натуральное число), элементами которых являются бинарные числа.

Определение 1. Траектория – это последовательность

$$\hat{A}: \hat{A} \in G(n, N) \ \forall n, N>1. \quad (1)$$

Действительно, если последовательно соединить точки, являющиеся элементами последовательности $\hat{A}$, получим траекторию в пространстве R^n.

Определение 2. N-членный фрагмент – это фрагмент длины n последовательности $A \in G(1, N)$.

Введем преобразование F_n одномерной последовательности в траекторию $\hat{A}$ в многомерном пространстве R^n (2):

$$F_n : G(1,N) \to G(n, N+1-n), F_n(A) = \hat{A}, \quad (2)$$

где $A = (a(t) : a(t) \in \{0,1\})_{t=1}^{N}$, $\hat{A} = (\hat{a}(t) : \hat{a}(t) = (a(t+i-1), i = \overline{1,n}))_{t=1}^{N+1-n}$, то есть $\hat{A}$ – это последовательность векторов $\hat{a}_n$ в многомерном пространстве.

В общем случае входная последовательность A может содержать одинаковые n-членные фрагменты, что приводит к возникновению точек самопересечения траектории.

Обратное преобразование к (2) вычисляется согласно (3):

$$F_n^{-1} : G(n,N) \to G(1, N+1-n), F_n^{-1}(\hat{A}) = A, \quad (3)$$

где $\hat{A} = (\hat{a}(t) : \hat{a}(t) = (a(t+i-1), i = \overline{1,n}))_{t=1}^{N}$, а

$$A = \left\{ a(i) : a(i) = \begin{cases} \hat{a}_1(i), 1 \le i \le N \\ \hat{a}_{i+1-N}(N), N < i < N+n \end{cases} \right\}_{i=1}^{N+n-1}.$$

Обратное преобразование существует во всех точках траектории, кроме точек ее самопересечения, где оно должно быть доопределено.

Формирование поуровневых словарей. Механизм памяти, чувствительный к числу прохождений заданной вершины в заданном направлении, является инструментом для анализа входной последовательности с точки зрения повторяющихся ее частей. Как было показано выше, одинаковые фрагменты последовательности отображаются преобразованием F_n в одну и ту же часть траектории $\hat{A}$ в многомерном пространстве R^n.

Словарь формируется на основе анализа множества последовательностей $\{A_k\}$, в каждой из которых с помощью суперпозиции $H_h RMF_n$ (отображением последовательностей класса $\{A_k\}$ преобразованием F_n в *n*-мерное пространство, запоминанием числа прохождений траекторией конкретной точки в памяти нейрона M, считыванием R содержимого памяти всех нейронов, и применением к ним порогового преобразования H_h) выделяются входящие в нее не менее h раз (где h – значение порога порогового преобразования H_h) подпоследовательности $\{B_j\} \subset A_k$. Таким образом, преобразование $H_h RMF_n$ при взаимодействии с входным множеством $\{A_k\}$ генерирует словарь $\{\hat{B}_j\}$, характеризующий траектории, соответствующие подпоследовательностям B_j входного множества в пространстве данной мерности R^n:

$$\{\hat{B}_j\} = H_h RMF_n(\{A_k\}) . \quad (4)$$

В зависимости от величины порога h порогового преобразования H слова $\hat{B}_j$ словаря могут быть деревьями или графами, содержащими циклы.

Формирование синтаксических последовательностей. Сформированный словарь может быть использован для детектирования старой информации (слов словаря $\hat{B}_j$) в потоке новой (во входной последовательности $\tilde{A}$, отличающейся от последовательностей множества $\{A_k\}$, сформировавших словарь). Для этого необходимо поглощение фрагментов траектории $\hat{\tilde{A}}$ входной последовательности $\tilde{A}$, соответствующих словам словаря $\{\hat{B}_j\}$, и пропускание новой относительно словаря информации (их связей).

Для решения задачи детектирования преобразование F_n^{-1} модифицируется для придания ему детектирующих свойств. Модификация $F_{n,C}^{-1}$ состоит в том, чтобы выделить из входной последовательности $\tilde{A} \in G(1, N)$, содержащей наряду со старой информацией входящих в неё слов словаря (3), некоторую новую информацию (информацию о связях элементов словаря во входной последовательности). Использование преобразования $F_{n,C}^{-1}$ позволяет сформировать так называемую синтаксическую последовательность или последовательность аббревиатур C, характеризующую связи слов словаря $\{\hat{B}_j\}$ в последовательностях множества $\{A_k\}$. Обозначим через $\{B_j\}$ множество подпоследовательностей, соответствующих всем цепям слов $\hat{B}_j$ словаря (4). Тогда:

$$F_{n,C}^{-1}(\hat{\tilde{A}}, \{\hat{B}\}) = C \quad (5)$$

$$C = \left\{ c(t) : c(t) = \begin{cases} 0, \ \textit{если} \ \exists l,k : (\hat{\tilde{a}}(l),\ldots,\hat{\tilde{a}}(l+k)) \in \{\hat{B}\}, \ l \le t \le l+k \\ \tilde{a}(t); \ \textit{иначе} \end{cases}, \ t = 1,\ldots,N \right\}$$

$$\{C\} = F_{n,C}^{-1}(F_n(\tilde{A}), H_h RM(\{\hat{A}\})) = F_{n,C}^{-1}(F_n(\tilde{A}), \{\hat{B}\}) \quad (6)$$

Таким образом, отображение $F_{n,C}^{-1}$ позволяет устранить из входной последовательности $\tilde{A}$ некоторую информацию, содержащуюся в словаре $\{\hat{B}_j\}$. В результате реализуется структурный подход к обработке информации: сначала выявляются структурные элементы, потом – связи между ними. Синтаксическая последовательность C, содержащая только новую, по отношению к словарю данного уровня, информацию, становится входной для следующего уровня. На следующем уровне, подобно описанному выше, из множества синтаксических последовательностей $\{C\}$ формируется словарь $\{\hat{D}\}$ и множество синтаксических последовательностей следующего уровня $\{E\}$. Таким образом, мы имеем стандартный двухуровневый элемент многоуровневой иерархической структуры. Такая обработка с выделением поуровневых словарей происходит на всех уровнях. Словарь следующего уровня является, в этом случае, грамматикой для предыдущего уровня, так как его элементами являются элементы связей между словами словаря предыдущего уровня.

2. ПОЛНЫЙ ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ ПРЕДЛОЖЕНИЙ ТЕКСТА

Первичная обработка текста. Задачами первичной обработки являются очистка текста от нетекстовой информации, а также корректная обработка таких единиц текста как аббревиатуры, инициалы, заголовки, адреса, номера, даты, указатели времени.

Последующий лингвистический анализ текста включает в себя графематический, морфологический [5] и синтаксический уровни обработки.

Графематический уровень обработки. На графематическом уровне анализа текст сегментируется на слова и предложения. Единицей графематического анализа для слов является цепочка символов, выделенная с двух сторон пробелами. Выделенная цепочка символов подвергается последовательной обработке эвристическими правилами: отсекаются знаки пунктуации, проверяется наличие гласных внутри цепочки, чередование верхнего и нижнего регистров и т.д. Аналогично можно сформулировать правила для выделения в тексте предложений.

Морфологический анализ. Полные словоформы анализируются на морфологическом уровне лингвистического анализатора, цель которого разбить все множество словоформ на подмножества по признаку принадлежности к той или иной лексеме, и по возможности однозначно определить их грамматические характеристики $\{B_i\}_1 = \{m_i\}$.

Большая часть лексем текста представляет неизменный фундамент языка и охватывается словарем в пределах 100 тысяч слов. Другая, более редкая, но не менее важная составляющая лексикона, постоянно пополняется и в принципе не имеет четко очерченных границ, прежде всего в части имен собственных и словообразовательных вариантов известных слов. Поэтому для морфологического анализа используются методы, как с декларативной, так и с процедурной ориентацией [5].

Для декларативного морфологического анализатора используется полный словарь всех возможных словоформ для каждого слова. При этом каждая словоформа снабжена полной и однозначной морфологической информацией, куда входят как постоянные, так и переменные морфологические

параметры. Задача морфологического анализа сводится к поиску нужной словоформы в словаре. Если слово не найдено, используются процедурные методы, где каждое слово разделяется на основу и аффикс, и словарь содержит только основы слов вместе со ссылками на соответствующие строки в словаре аффиксов.

Синтактико-семантический анализ отдельного предложения. Семантико-синтаксический анализ предложения выявляет информацию о связях слов в группах и между группами $\{B_k\}_1=\{r_k\}$, где r_k – предикативная связь субъекта с главным объектом, а $r_k \mid k>1$ – все остальные типы связей, и проводится в несколько этапов: осуществляется фрагментация предложения, объединение однородных фрагментов, установление иерархии между фрагментами разных типов, объединение фрагментов в простые предложения, построение внутри фрагментов простых синтаксических групп, выявление предикативного минимума каждого из простых предложений, выделение остальных членов простого предложения, являющихся актантами выявленного предиката, построение синтаксических групп, в которых актант предиката – главное слово [6].

Синтаксические правила задают отношения между словами (сегментами) в предикативном виде. В зависимости от типа сегментов и типа подчинительного союза с помощью эвристических правил реализуется несколько операций объединения над ними: подчинение, однородность, импликация, присоединение. В результате осуществляется разбиение сложных предложений на простые предложения, связанные сочинительными, или подчинительными союзами.

Следующий шаг – построение простых синтаксических групп внутри каждого простого предложения и выделение предикатного ядра. К простым синтаксическим группам относятся группы на атрибутивном уровне, группы с предлогом и сравнительные конструкции.

Множество простых предложений русского языка задается перечнем минимальных структурных схем предложений, описывающих предикативный минимум предложения.

Во всех сегментах предложения, не являющихся вложенными и однородными, проводится последовательный поиск подходящего шаблона минимальной структурной схемы предложения. В соответствие с найденным шаблоном, каждому главному члену предложения присваивается соответствующее значение.

Далее решается задача получения расширенной предикатной структуры простых предложений и заполнения валентных гнезд предиката [7]. Выделение остальных членов простого предложения (остальных семантически значимых объектов и атрибутов) проводится с помощью последовательного сравнения слов предложения с актантной структурой глагола, для чего используется словарь валентностей глаголов [8].

Введем понятие звёздочки [9]. Синтаксическую структуру типа:

$$d = <c_i <c_j>> = U_j <c_i\ c_j>, \qquad (7)$$

где c_i – главное слово, $<c_j>$ – множество зависимых слов, семантическихе признаки слова c_i, будем называть «звёздочкой». Такое название вполне объяснимо, поскольку дерево зависимости для структур подобного рода представляет собой граф типа звезда.

В случае реализации полной лингвистической обработки для каждого простого предложения можно построить расширенную предикатную структуру, которая после небольших преобразований сводится тоже к звездочке

$d = \cup_j <c_i\ r_k\ c_j>$, где c_i – предикат, c_j – его актанты. В звездочке, построенной из расширенной предикатной структуры, к паре <главное слово, зависимое слово> добавляется связь между ними, размеченная одним из k типов отношений «предикат-актант» [7].

3. СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТА

Анализ лексической семантики. Статистический анализа текста сводится к выявлению частоты встречаемости p_i слов в тексте – этап формирования словаря слов – лексического анализа, и к выявлению частоты попарной встречаемости p_{ij} слов в смысловых фрагментах текста – этап формирования словаря попарной встречаемости слов в предложениях текста. Попарная встречаемость характеризует смысловую сочетаемость слов в языке [10].

В качестве критерия для определения наличия семантической связи между парой понятий используется частота их совместной встречаемости в предложениях текста. Превышение частотой некоторого порога позволяет говорить о наличии между понятиями ассоциативной (семантической) связи, а совместные вхождения понятий в предложения с частотой, меньше порога, считаются просто случайными.

В простых случаях статистического анализа текста, для того чтобы анализ был более устойчивым, а полученные результаты – более интерпретируемыми, словоформы слов приводятся к их корневым основам. При этом формируются словарь корневых основ $\{B_k\}_2$, и словарь попарной сочетаемости корневых основ $\{B_m\}_4$. Выделенные таким образом корневые основы служат далее в качестве элементов для построения ассоциативной (однородной семантической) сети.

Формирование ассоциативной сети целого текста. Словарь попарной встречаемости корневых основ $\{B_m\}_4$ позволяет построить сеть, характеризующую смысловую структуру текста. Получается так называемая ассоциативная (однородная семантическая) сеть N как совокупность несимметричных пар понятий (корневых основ) $<c_i\ c_j>$, где c_i и c_j – понятия (корневые основы), связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста, например, в предложении) $<c_i\ c_j> = B_i \in \{B_i\}_4$:

$$N = \cup_i <c_i\ c_j>. \quad (8)$$

При этом пары корневых основ связываются через одинаковые корневые основы: $<c_1\ c_2>*<c_2\ c_3>$, где (*) означает присоединение справа. В результате получается цепочка $<c_1\ c_2\ c_3>$, к которой далее присоединяются другие пары. При этом возможны ветвления и вхождения, то есть, строится действительно сеть.

Если предварительно объединить все пары слов с одинаковым первым словом в звездочку $d = <c_i <c_j>> = \cup_j <c_i\ c_j>$ (где c_i – главное слово, $<c_j>$ – множество его семантических), то можно сказать, что сеть может быть построена и объединением всех звездочек. Звездочки из корневых основ связываются через одинаковые понятия (корневые основы): $<c_1 <c_j>>*<c_{j1} <c_k>>$, где (*) также означает присоединение справа. В результате получается цепочка звездочек. При этом также возможны ветвления и вхождения, то есть, ассоциативная сеть N строится и таким способом тоже:

$$N = \cup_i <c_i <c_j>>. \quad (9)$$

Переранжирование понятий. Элементы семантической (ассоциативной) сети $N = \cup_i <c_i <c_j>>$. и их связи имеют числовые характеристики,

отражающие их относительный вес в данной предметной области – семантический вес. При достаточно представительном множестве текстов, описывающих предметную область, значения частот встречаемости понятий действительно отражают соответствующие семантические (субъективно оцениваемые) веса. Однако, для небольших обучающих выборок, в частности, при анализе отдельного текста, не все частотные характеристики соответствуют действительным семантическим весам – важности понятий в тексте. Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого "семантического сгущения". В результате такого анализа наибольший вес приобретают понятия, обладающие мощными связями и находящиеся как бы в центре "семантических сгущений". При этом на каждой итерации переранжирования понятия, связанные с понятиями, имеющими большой вес, свой вес увеличивают. Другие его равномерно теряют.

Для этого сформированное первоначально статистическое представление текста – сеть слов с их связями переранжируется с помощью итеративной процедуры, аналогичной алгоритму сети Хопфилда, что позволяет перейти от частотного портрета текста к ассоциативной сети ключевых понятий текста:

$$w_i(t+1) = (\sum_{\substack{i \\ i \neq j}} w_i(t) w_{ij}) \sigma(\overline{E}) . \qquad (10)$$

здесь $w_i(0) = p_i$, $w_{ij} = p_{ij}/p_j$ и $\sigma(\overline{E}) = 1/(1+e^{-k\overline{E}})$ – функция, нормирующая на среднее значение энергии всех вершин сети E, где p_i – частота встречаемости i-го слова в тексте, p_{ij} – частота совместной встречаемости i-го и j-го слов в фрагментах текста (предложениях). Полученная числовая характеристика слов – их смысловой вес – характеризует степень их важности в тексте.

4. ОБЪЕДИНЕНИЕ ПОДХОДОВ: СЕМАНТИЧЕСКИЙ И ПРАГМАТИЧЕСКИЙ АНАЛИЗ ЦЕЛОГО ТЕКСТА

Наконец, мы подошли к объединению подходов. Если выявить расширенную предикатную структуру предложения, привести ее к виду звездочки, а потом из этих звездочек построить семантическую сеть, и переранжировать ее вершины, мы получим возможность более точно описывать смысл текста. А потом перейти от статического представления смысла текста (семантической сети) к его динамическому представлению (к прагматике). Отличие звездочки, построенной из пар понятий, от звездочки, построенной из расширенной предикатной структуры, заключается в замене связей ассоциативного типа между понятиями звездочки на связи любых мыслимых семантических типов (присутствующих в предикатных структурах).

4.1. Семантический анализ целого текста

Для этого необходимо реализовать полную лингвистическую обработку, то есть на морфологическом уровне выявить вся морфологическую информацию о словах $\{B_i\}_1=\{m_i\}$, а на синтаксическом – информацию о связях слов в группах и между группами $\{B_k\}_3=\{r_k\}$, где r_1 – предикативная связь субъекта с главным объектом, а r_i $i>1$ – все остальные типы связей. Структуры синтаксического уровня при этом учитываются в словаре шаблонов простых синтаксических структур и в словаре валентностей глаголов [8].

Формирование звездочки с размеченными отношениями. В случае реализации полной лингвистической обработки для каждого простого

предложения можно построить расширенную предикатную структуру (см. предыдущий раздел), которая после небольших преобразований сводится к звездочке:

$$d = \cup_j <c_i\ r_i\ c_j>. \quad (11)$$

Для этого из расширенной предикатной структуры удаляется предикат, субъект через предикативную связь связывается с главным объектом. Остальные объекты и атрибуты присоединяются к субъекту как актанты предиката присоединялись к предикату. Правда, в отличие от звездочки с простыми ассоциативными связями, в звездочке, построенной из расширенной предикатной структуры, вместо пар понятий (корневых основ) используются тройки $<c_i\ r_i\ c_j>$, где между парой понятий имеется связь, размеченная одним из типов отношений.

Формирование частотного портрета текста. И в этом случае строится частотный портрет текста, то есть выявляются частоты p_i встречаемости корневых основ понятий (полученных в результате морфологического анализа), и частоты p_{ij} их попарной встречаемости в предложениях текста. И, наконец, частоты встречаемости переранжируются в смысловые веса с использованием итеративной процедуры, похожей на алгоритм искусственной нейронной сети, предложенной Хопфилдом. В результате итеративной процедуры переранжирования наибольшие веса получают понятия, связанные с наибольшим числом других понятий с большим весом, то есть те понятия, которые стягивают на себя смысловую структуру текста. Полученные таким образом смысловые веса ключевых понятий показывают значимость этих понятий в конкретном тексте.

4.2. Прагматический анализ текста

Формирование реферата текста. Рассмотрим далее, что можно сделать с текстом и полученной из него неоднородной семантической сетью. Поскольку понятия – вершины семантической сети конкретного текста – в процессе анализа оказываются ранжированными их смысловыми весами, мы можем воспользоваться этим для выявления наиболее значимой для текста части предложений. Мы можем вычислить весовые характеристики предложений текста как сумму весов включенных в предложение лексем. Далее, мы можем удалить предложения текста, вес которых превышает заданный порог. Это будет похоже на квази-реферат текста. Связность текста может быть нарушена, но предложения, в нем содержащиеся, будут нести основной смысл текста.

Формирование цепочек звездочек. Необходимо заметить, что семантическая сеть текста одновременно включает в себя все понятия текста. Но если спроецировать предложения текста на эту сеть, то мы получим последовательность понятий сети, которые следуют друг за другом последовательно во времени. Эти понятия включены во фрагменты текста, которые либо являются описаниями чего-либо (князь Андрей ехал мимо дуба: «Старый дуб, весь преображенный, раскинувшись шатром сочной, темной зелени, млел, чуть колыхаясь в лучах вечернего солнца …» Л.Н. Толстой, Война и мир), либо описывают алгоритм реализации чего-либо («Опустим в колбу с кислородом железную проволоку с кусочком угля на конце …». http://files.school-collection.edu.ru/dlrstore/deb6e939-f8c8-fea7-fe24-7b2c80013fd7/index.htm). И в том и в другом случае в этих последовательностях предложений есть как существенно важные для предметной области предложения, так и второстепенные.

Отдельные предложения этих фрагментов описывают отдельные фрагменты ситуации. Действительно, расширенная предикатная структура простого распространенного предложения содержит не больше семи актантов глагола – предиката. Это столько, сколько удерживает в акте внимания кратковременная память человека (7 ± 2). Расширенной предикатной структуре соответствует, после описанных выше преобразований, звездочка $d = \cup_j <c_i\ r_i\ c_j>$. Тогда цепочка расширенных предикатных структур (11) содержит смысл этих фрагментов – описаний, или алгоритмов:

$$D = (d_i \mid i = \overline{1, N}). \qquad (12)$$

Модель предметной области как множество классов цепочек звездочек. Подберем корпус текстов таким образом, чтобы он описывал некоторую предметную область. В этом случае предложения текстов корпуса, превышающие пороговое значение (квази-рефераты текстов корпуса) включают в себя (с некоторой заданной ошибкой) содержание предметной области. Смысл этих последовательностей предложений может быть представлен последовательностями расширенных предикатных структур этих предложений. То есть последовательности расширенных предикатных структур (и цепочек соответствующих звездочек) являются моделью предметной области D_i:

$$M = \cup_i D_i. \qquad (13)$$

Кластеризация цепочек звездочек текстов модели предметной области. Это множество цепочек звездочек $\{D_i\}$ является избыточным: в текстах, описывающих предметную область, могут быть смысловые повторы. Чтобы сформировать минимальное описание смысла предметной области, проведем кластеризацию множества цепочек по степени их похожести. Учитывая различную степень полноты описания смысла отдельного предложения в конкретном предложении, будем использовать нечеткое сравнение звездочек. То есть будем считать похожими звездочки, имеющие несовпадающими только некоторое (не больше заданного порога) количество семантических признаков. При этом все множество звездочек разобьется на классы, в которых можно выбрать некоторым образом представителя класса. Цепочки звездочек в процессе кластеризации могут как разбиваться на более мелкие подцепочки, так и рекомбинировать в более крупные. Тогда множество цепочек звездочек – представителей классов – будет минимальным описанием модели предметной области.

Классификация текстов. При наличии множества моделей предметных областей $\{M_i\}$, входной текст можно отнести к конкретной предметной области подсчитывая степень пересечения модели текста (множества цепочек звездочек, соответствующего этому тексту $\{D_i\}$) с моделями предметных областей $\{M_j\}$: $CS_i = \{D_i\} \cap \{M_j\}$. При этом отнесение к классу осуществляется вычислением $\arg\max CS_i$.

5. ПРИМЕР ПРАГМАТИЧЕСКОГО АНАЛИЗА ТЕКСТА

Рассмотрим прагматический анализ текста с привлечением описанных выше механизмов, позволяющий выявить цепочки предикатных структур предложений текста, существенно важные для представления смысла этого текста. Покажем, как выявляются эти предложения (квазиреферат текста), формируется расширенная предикатная структура, и как выглядит цепочка таких структур для некоторого фрагмента квазиреферата. Чтобы интерпретация цепочки была более понятной, из расширенных предикатных структур оставим только их наиболее важные части: (субъект-предикат– главный объект).

5.1. Формирование частотного портрета текста

Для примера возьмем текст учебника Т.И. Трофимовой «Курс физики», Москва, «Высшая школа», 2001. Фрагмент частотного портрета этого текста представлен в Таблице 1. Здесь после удаления из текста не несущих смысла слов, и стемминга оставшихся слов, подсчитывается частота появления в тексте корневых основ оставшихся после удаления слов.

Таблица 1

Фрагмент таблицы частоты встречаемости корневых основ

	Семантически значимый объект или атрибут	Частота встречаемости
1	поверхн (поверхность)	8
2	замкнут систем (замкнутая система)	7
3	величин (величина)	6
4	движен материальн точк (движение материальной точки)	3
5	плоскост (плоскость)	7

Здесь «движение материальной точки» является устойчивым словосочетанием, и потому рассматривается как единое понятие.

После формирования семантической сети частоты встречаемости корневых основ пересчитываются в их смысловые веса, что позволяет в дальнейшем вычислить смысловые веса предложений.

5.2. Формирование квазиреферата текста

Если взять текст того же учебника Т.И. Трофимовой «Курс физики», Москва, «Высшая школа», 2001 (ниже приведен лишь его фрагмент):

«… Первый закон Ньютона: всякая материальная точка (тело) сохраняет состояние покоя или равномерного прямолинейного движения до тех пор, пока воздействие со стороны других тел не заставит ее изменить это состояние. Первый закон Ньютона выполняется не во всякой системе отсчета, а те системы, по отношению к которым он выполняется, называются инерциальными системами отсчета. …».

и удалить из текста предложения, имеющие смысловой вес менее заданного порогового значения, то останется квазиреферат текста, фрагмент которого представлен в Таблице 2.

Таблица 2

Существенные предложения в порядке их следования в тексте – квазиреферат (фрагмент)

	Предложение	Вес
1	Первый закон Ньютона: всякая материальная точка (тело) сохраняет состояние покоя или равномерного прямолинейного движения до тех пор, пока воздействие со стороны других тел не заставит ее изменить это состояние.	99
2	Первый закон Ньютона выполняется не во всякой системе отсчета, а те системы, по отношению к которым он выполняется, называются инерциальными системами отсчета.	97
3	Инерциальной системой отсчета является такая система отсчета, относительно которой материальная точка, свободная от внешних воздействий, либо покоится, либо движется равномерно и прямолинейно.	99

5.3. Формирование расширенной предикатной структуры предложения

Для примера возьмем предложение из того же учебника: *«Механика – часть физики, которая изучает закономерности механического движения и причины, вызывающие или изменяющие это движение».*

Не будем демонстрировать подробности лингвистического механизма извлечения расширенной предикатной структуры из предложения. Покажем конечный результат. Единственное замечание: предложение разбивается на простые составляющие *«Механика – часть физики»* и *«Механика изучает закономерности механического движения и причины, вызывающие или изменяющие это движение»*.

Для первой части расширенная предикатная структура имеет очень простой вид: «Механика (субъект) – включена в (предикат) – физику (главный объект)».

5.4. Формирование цепочек звездочек

Из предложений квазиреферата выявляются их расширенные предикатные структуры, которые формируют те самые цепочки, характеризующие прагматику текста или целой предметной области, описываемой корпусом текстов.

Для простоты восприятия ниже приведена цепочка только существенно важной части предикатных структур (субъект←предикат→главный объект). Если в предикатной структуре объект отсутствует, то в цепочке на его месте стоит NUL. Остальные члены расширенных предикатных структур опущены.

1) точка←сохраняет→состояние; 2)

не заставит→её
воздействие←
изменить→состояние

3) закон←выполняется→NUL; 4) системы←называются→NUL;
5) система←является→NUL; 6) точка←(покоится, движется)→NUL

ЗАКЛЮЧЕНИЕ

В работе изложен подход, объединяющий статистические и лингвистические методы анализа текстов, статистическая и прагматическая обработка текста с помощью предложенного подхода демонстрируется на конкретных примерах. Совместное использование быстрых и независимых от языка статистических алгоритмов обработки теста и лингвистических баз знаний в виде словарей валентности позволяют получить семантические представления целого текста с точностью, присущей лингвистическому подходу. Такое объединение позволяет рассматривать прагматический уровень как динамическое представление на фоне статистического, соответствующего семантическому уровню, что приводит к более точной классификации текстов по сравнению с использованием только статистического представления. Предложенное в работе понимание прагматики текста в общем случае не является общераспространенным. Более того, пожалуй, затруднительно дать какое-либо более или менее общепринятое определение прагматики текста. Поэтому мы оставляем за собой право называть представления обработки этого уровня прагматическими, хотя бы потому, что они надстраиваются над семантикой. И готовы дискутировать эту точку зрения. Тем не менее, такое представление оказывается достаточно конструктивным для реализации реальных механизмов автоматической обработки текстовой информации и, как все другие, имеет право на существование.

ЛИТЕРАТУРА

1. *Леонтьева Н.Н.* Автоматическое понимание текстов. Системы, модели, ресурсы – М.: «Academia», 2006
2. *Харламов А.А.* Нейросетевая технология представления и обработки информации (естественное представление знаний). – М.: «Радиотехника», 2006

3. *Kharlamov A.A., Raevsky V. V.* Networks constructed of neuroid elements capable of temporal summation of signals. /In «Neural Information Processing: Research and Development», Jagath C. Rajapakse and Lipo Wang, Editors, Springer-Verlag, May, 2004, ISBN 3-540-21123-3. Pp. 56-76

4. *Виноградова О. С.* Гиппокамп и память. – М.: «Наука», 1975

5. *Харламов А.А., Ермоленко Т.В., Дорохина Г.В., Журавлев А.О.* Предсинтаксический анализ русско-английских текстов // Программная инженерия, № 10, 2013, – С. 37 – 47

6. *Харламов А. А., Ермоленко Т. В.* Разработка компонента синтаксического анализа предложений русского языка для интеллектуальной системы обработки естественно-языкового текста// Программная инженерия № 7, 2013. – С. 37-47

7. *Kharlamov A.A., Ermolenko T.V.* Semantical Text Analysis on the Basis of Detecting of Key Predicate Structures. Proceedings of the 14-th International Conference «Speech and Computer SPECOM'2011», – М.:, 2011. – Pp. 383 – 388

8. *Дорохина Г. В.* Автоматическое выделение синтаксически связанных слов простого распространенного неосложненного предложения / Г.В. Дорохина, Д. С. Гнитько // «Сучасна інформаційна Україна: інформатика, економіка, філософія»: матеріали доповідей конференції, 12 13 травня 2011 року, Донецьк, 2011. Т. 1. – С. 34-38

9. *Харламов А.А., Раевский В.В.* Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды. Речевые технологии, N 3, 2008. – С. 27-35

10. *Рахилина Е.В.* Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари, 2000

Сведения об авторах

Харламов Александр Александрович,
доктор технических наук, учредитель и директор компании ООО НПИЦ «МИКРОСИСТЕМЫ», старший научный сотрудник Института высшей нервной деятельности и нейрофизиологии РАН, автор технологии смысловой обработки текстов TextAnalyst, признанной лучшей по данным ведомственной экспертизы в реализации построения рефератов. Область научных интересов: нейроинформатика, распознавание речи, анализ текстов, распознавание изображений, семантические представления, искусственные нейронные сети.
E-mail: kharlamov@analyst.ru

Ермоленко Татьяна Владимировна,
кандидат технических наук, начальник отдела распознавания речевых образов Института проблем искусственного интеллекта МОНМС и НАН Украины (г. Донецк), доцент кафедры компьютерных технологий физико-технического факультета Донецкого национального университета, доцент кафедры программного обеспечения интеллектуальных систем факультета компьютерных наук и технологий Донецкого национального технического университета. Область научных интересов: технологии искусственного интеллекта, распознавание речи, идентификация диктора, автоматический анализ текстов.
E-mail: naturewild71@gmail.com