УДК 004.89:004.93

А. А. Харламов¹, Т. В. Ермоленко²

¹Институт высшей нервной деятельности и нейрофизиологии РАН, г. Москва
117485, г. Москва, ул. Бутлерова, 5а

²Донецкий национальный университет, г. Донецк
83001, г. Донецк, ул. Университетская, 24

АНАЛИЗ ТЕКСТОВ: ЛИНГВИСТИКА, СЕМАНТИКА, ПРАГМАТИКА В РАМКАХ КОГНИТИВНОГО ПОДХОДА

A. A. Kharlamov¹, T. V. Yermolenko²

¹Institute for Higher Nervous Activity and Neurophysiology RAS, c. Moscow
117485, c. Moscow, Butlerova str., 5a

²Donetsk National University, c. Donetsk
83001, c. Donetsk, University str., 24

TEXT ANALYSIS: LINGUISTICS, SEMANTICS, PRAGMATICS IN THE COGNITIVE APPROACH

О. О. Харламов¹, Т. В. Ермоленко²

¹Інститут вищої нервової діяльності та нейрофізіології РАН, м. Москва
117485, м. Москва, вул. Бутлерова, 5а

²Донецький національний університет, м. Донецьк
83001, м. Донецьк, вул. Університетська, 24

АНАЛІЗ ТЕКСТІВ: ЛІНГВІСТИКА, СЕМАНТИКА, ПРАГМАТИКА В РАМКАХ КОГНІТИВНОГО ПІДХОДУ

В статье рассматривается когнитивный подход к анализу лингвистической информации человеком. Рассматриваются процессы обработки информации разных лингвистических уровней: морфологического, лексического, синтактико-семантического для отдельного предложения, и, наконец, семантического и прагматического – для всего текста в целом. В качестве примера упомянутой обработки выявляется представление прагматического уровня – цепочка расширенных предикатных структур конкретного текста.

Ключевые слова: автоматическая обработка текстов, когнитивный подход, морфологическая обработка, лексическая обработка, синтаксическая обработка, семантическая обработка, прагматическая обработка, цепочка расширенных предикатных структур.

This paper considers the cognitive approach to linguistic information analyzed by the human. The processes of information processing are studied at various linguistic levels: morphological, lexical, syntactic and semantic levels for separate sentences, and finally, semantic and pragmatic levels for the text as a whole. As an example of the said processing, representation of the pragmatic level is identified as a chain of extended predicate structures of a particular text.

Key words: automatic text processing, cognitive approach, morphological processing, lexical processing, syntactic processing, semantic processing, pragmatic processing, chain of extended predicate structures.

Розглядається когнітивний підхід до аналізу лінгвістичної інформації людиною, а також процеси обробки інформації різних лінгвістичних рівнів: морфологічного, лексичного, синтактико-семантичного для окремого речення, і, нарешті, семантичного та прагматичного – для всього тексту в цілому. Як приклад зазначеної обробки виявляється подання прагматичного рівня – ланцюжок розширених предикатних структур конкретного тексту.

Ключові слова: автоматична обробка текстів, когнітивний підхід, морфологічна обробка, лексична обробка, синтаксична обробка, семантична обробка, прагматична обробка, ланцюжок розширених предикатних структур.

Введение

В настоящий момент два основных подхода главенствуют в автоматическом смысловом анализе текстов: лингвистический и статистический. Первый – дает очень точный анализ смысла 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 (квазислов из квазитекстов, в том числе – и обычных текстов) различной частоты встречаемости различных модальностей. В гиппокампе происходит ранжирование этих представлений, характеризующее значимость этих представлений в рамках отдельных ситуаций (квазитекстов).

1 Обработка лингвистической информации человеком

Рассмотрим 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.

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

Пусть мы имеем n -мерное сигнальное пространство R^n и в нем единичный гиперкуб $G^n \in R^n$.

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):

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

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 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):

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

где

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

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

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 $H_h R M F_n$ (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 $H_h R M F_n$ 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 :

$$\{\hat{B}_j\} = H_h R M F_n(\{A_k\}) \tag{4}$$

В зависимости от величины порога h порогового преобразования H слова словаря $\hat{B}_j$ 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 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. 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:

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

$$C = c(t): c(t) = \begin{cases} 0, & \text{если } \exists l, k: (\tilde{a}(l), \dots, \tilde{a}(l+k)) \in \{\hat{B}\}, l \leq t \leq l+k, \\ & \tilde{a}(t), \text{ иначе} \end{cases}, t=1, \dots, N, \\ \{C\} = F_{n,C}^{-1}(F_n(\tilde{A}), H_n RM(\{\hat{A}\})) = F_{n,C}^{-1}(F_n(\tilde{A}), \{\hat{B}\}) \quad (6)$$

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 элементы структуры, 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.

Анализ текста. При анализе текста на этапе морфологической обработки формируется словарь первого уровня – $\{B_j\}_1$ – словарь флексий. Потом формируется словарь второго уровня – $\{B_k\}_2$ – словарь корневых основ. На следующих этапах формируются: словарь третьего уровня – $\{B_l\}_3$ – словарь флективных структур синтаксических групп, и словарь четвертого уровня – $\{B_m\}_4$ – словарь попарной встречаемости в тексте корневых основ. Совместная встречаемость характеризуется ассоциативной связью между этими словами – иначе она означает смысловую корректность предложения («Зеленые идеи яростно спят»).

Введем понятие звездочки [6]. Синтаксическую структуру типа:

$$d = \langle c_i \langle c_j \rangle \rangle = \cup_j \langle c_i c_j \rangle, \quad (7)$$

где c_i – главное слово, $\langle c_j \rangle$ – множество зависимых слов, семантические признаки слова c_i , будем называть «звездочкой».

2 Полный лингвистический анализ предложений текста

В процессе полной лингвистической обработки на графематическом уровне анализа текст сегментируется на слова и предложения, на морфологическом уровне выявляется вся морфологическая информация о словах $\{B_j\}_1 = \{m_j\}$, а на синтаксическом – информация 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. Структуры синтаксического уровня укладываются в рамки словаря шаблонов минимальных структурных схем предложения и словаря валентностей глаголов [5].

В случае реализации полной лингвистической обработки для каждого простого предложения можно построить расширенную предикатную структуру, которая после небольших преобразований сводится тоже к звездочке $d = \cup_j \langle c_i r_k c_j \rangle$, где c_i – предикат, c_j – его актаны. 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

Формирование ассоциативной сети целого текста. Статистический анализ текста сводится к выявлению частоты встречаемости p_i слов в тексте, и к выявлению частоты попарной встречаемости p_{ij} слов в смысловых фрагментах текста. Попарная встречаемость характеризует смысловую сочетаемость слов в языке [12].

В простых случаях статистического анализа текста, для того чтобы анализ был более устойчивым, а полученные результаты – более интерпретируемыми, слово-формы 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.

Ассоциативная (однородная семантическая) сеть N как совокупность несимметричных пар понятий (корневых основ) $\langle c_i c_j \rangle$, где c_i и c_j – понятия (корневые основы), связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста, например, в предложении) $\langle c_i c_j \rangle = B_i \in \{B_i\}_4$:

$$N = \cup_i \langle c_i c_j \rangle. \quad (8)$$

При этом пары корневых основ связываются через одинаковые корневые основы: $\langle c_1 c_2 \rangle * \langle c_2 c_3 \rangle$, где (*) означает присоединение справа. В результате получается цепочка $\langle c_1 c_2 c_3 \rangle$, 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 $d = \langle c_i \langle c_j \rangle \rangle = \cup_j \langle c_i c_j \rangle$ (где c_i – главное слово, $\langle c_j \rangle$ – множество его семантических признаков), то можно сказать, что сеть может быть построена и объединением всех звездочек:

$$N = \cup_i \langle c_i \langle c_j \rangle \rangle. \quad (9)$$

Переранжирование понятий. Элементы семантической (ассоциативной) сети $N = \cup_i \langle c_i \langle c_j \rangle \rangle$ и их связи имеют числовые характеристики, отражающие их относительный вес в данной предметной области – семантический вес. Для более точной оценки семантических весов понятий используются веса всех связанных с ними понятий, т.е. веса целого «семантического сгущения». В результате итеративной процедуры переранжирования на каждой итерации понятия, связанные с понятиями, имеющими большой вес, свой вес увеличивают. Другие его равномерно теряют:

$$w_i(t+1) = \left(\sum_{i \neq j} w_i(t) w_{ij} \right) \sigma(\bar{E}), \quad (10)$$

здесь $w_i(0) = p_i$, $w_{ij} = p_{ij}/p_j$ и $\sigma(\bar{E}) = 1/(1 + e^{-k\bar{E}})$ – 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 смысловый вес – характеризует степень их важности в тексте.

4 Объединение подходов.

Семантический и прагматический анализ целого текста

Наконец, мы подошли к объединению подходов.

Семантический анализ целого текста. Если выявить расширенную предикатную структуру предложения с использованием полного лингвистического анализа предложения, привести ее к виду звездочки, а потом из этих звездочек построить семантическую сеть, и переранжировать ее вершины, мы получим сеть, в которой ассоциативные связи заменены на соответствующие типы связей. При этом в отличие от звездочки с простыми ассоциативными связями, в звездочке, построенной из расширенной предикатной структуры, вместо пар понятий используются тройки $\langle c_i r_i c_j \rangle$, где между парой понятий имеется связь, размеченная одним из типов отношений.

И в этом случае строится частотный портрет текста, то есть выявляются частоты p_i встречаемости понятий, 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.

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 понятий. Далее, мы можем удалить предложения текста, вес которых превышает заданный порог. Мы получим квазиреферат текста. Связность текста может быть нарушена, но предложения, в нем содержащиеся, будут нести основной смысл текста.

Формирование цепочек звездочек. Отдельные предложения квазиреферата, и соответствующие им расширенные предикатные структуры, описывают отдельные фрагменты ситуации. Расширенной предикатной структуре соответствует, после описанных выше преобразований, звездочка $d = \cup_j \langle c_i r_i c_j \rangle$. Тогда цепочка из расширенных предикатных структур содержит смысл квазиреферата:

$$D = (d_i \mid i = \overline{1, N}) \quad (11)$$

где N – число предложений в квазиреферате.

Модель предметной области как множество классов цепочек звездочек. Подберем корпус текстов таким образом, чтобы он описывал некоторую предметную область. В этом случае квазирефераты текстов корпуса включают в себя 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 :

$$M = \cup_i D_i. \quad (12)$$

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. Тогда множество цепочек звездочек – представителей классов – будет минимальным описанием модели предметной области.

Классификация текстов. При наличии множества моделей предметных областей $\{M_j\}$, 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. Чтобы интерпретация цепочки была более понятной, из расширенных предикатных структур оставим только их наиболее важные части: (субъект-предикат-главный объект).

В качестве примера формирования расширенной предикатной структуры предложения возьмем предложение из учебника Т. И. Трофимовой «Курс физики», Москва, «Высшая школа», 2001: *«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)».

Далее, после удаления из текста учебника не несущих смысла слов, и стемминга оставшихся слов, подсчитываем частоту появления в тексте корневых основ оставшихся после удаления слов. После формирования семантической сети, частоты встречаемости корневых основ пересчитываются в их смысловые веса, что позволяет вычислить смысловые веса предложений.

Если удалить из текста предложения, имеющие смысловый вес менее заданного порогового значения, то останется квазиреферат текста, фрагмент которого представлен ниже.

«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. Инерциальной системой отсчета является такая система отсчета, относительно которой материальная точка, свободная от внешних воздействий, либо покоится, либо движется равномерно и прямолинейно».

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). Остальные члены расширенных предикатных структур опущены.

Выводы

В работе изложен подход, объединяющий статистические и лингвистические методы анализа текстов, а семантическая и прагматическая обработка текста с помощью предложенного подхода демонстрируется на конкретных примерах. Совместное использование быстрых статистических алгоритмов обработки текста, и лингвистических алгоритмов и баз знаний в виде словарей валентности позволяют получить семантические представления целого текста с точностью, присущей лингвистическому подходу. Предложенное в работе понимание прагматики текста в общем случае не является общепринятым. Дело в том, что 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. Тем не менее, такое представление оказывается достаточно конструктивным для реализации реальных механизмов автоматической обработки текстовой информации.

Список литературы

1. Леонтьева Н. Н. Автоматическое понимание текстов. Системы, модели, ресурсы / Н. Н. Леонтьева. – М. : «Academia». – 2006. – 304 с.
2. Харламов А. А. Нейросетевая технология представления и обработки информации (естественное представление знаний) / Харламов А. А. – М. : «Радиотехника». – 2006. – 89 с.
3. Kharlamov A. A. Networks constructed of neuroid elements capable of temporal summation of signals / A. A. Kharlamov, V. V. Raevsky // Neural Information Processing: Research and Development, Jagath C. Rajapakse and Lipo Wang, Editors, Springer-Verlag. – 2004 – P. 56-76.
4. Виноградова О. С. Гиппокамп и память / Виноградова О. С. – М. : Наука, 1975. – 239 с.
5. Дорохина Г. В. Автоматическое выделение синтаксически связанных слов простого распространенного неосложненного предложения / Г. В. Дорохина, Д. С. Гнисько // Сучасна інформаційна Україна: інформатика, економіка, філософія: матеріали доповідей конференції, 12 – 13 травня 2011 року. – Донецьк, 2011. – Т. 1. – С. 34-38.
6. Харламов А. А. Перестройка модели мира, формируемой на материале анализа текстовой информации с использованием искусственных нейронных сетей, в условиях динамики внешней среды / А. А. Харламов, В. В. Раевский // Речевые технологии – 2008. – № 3. – С. 27-35.
7. Kharlamov A. A. Semantical Text Analysis on the Basis of Detecting of Key Predicate Structures / A. A. Kharlamov, T. V. Yermolenko // Proceedings of the 14-th International Conference «Speech and Computer SPECOM'2011» – М. – 2011. – P. 383-388.
8. Hopfield J. J. Neural networks and physical systems with emergent collective computational abilities // Proc. Natl. Acad. Sci. 79. – 1982. – P. 2554-2558.
9. Радченко А. Н. Моделирование основных механизмов мозга / А. Н. Радченко– Л. : «Наука» – 1969. – 211 с.
10. Харламов А. А. Предсинтаксический анализ русско-английских текстов / А. А. Харламов, Т. В. Ермоленко, Г. В. Дорохина, А. О. Журавлёв // Программная инженерия – 2013. – № 10 – С. 37-47.

11. Харламов А. А. Разработка компонента синтаксического анализа предложений русского языка для интеллектуальной системы обработки естественно-языкового текста / А. А. Харламов, Т. В. Ермоленко // Программная инженерия – 2013. – № 7 – С. 37-47.
12. Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость / Рахилина Е. В. – М. : Русские словари – 2000. – 416 с.

References

1. Leont'eva N.N. Avtomaticheskoe ponimanie tekstov. Sistemy, modeli, resursy – М.: «Academia». – 2006. – 304 s.
2. Harlamov A.A. Nejrosetevaja tehnologija predstavlenija i obrabotki informacii (estestvennoe predstavlenie znanij). – М.: «Radiotekhnika». – 2006. – 89 s.
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 – 2004 – ISBN 3-540-21123-3. – P. 56-76.
4. Vinogradova O. S. Gippokamp i pamjat'. – М.: «Nauka» – 1975 – 239 s.
5. Dorohina G. V. Avtomaticheskoe vydelenie sintaksicheski svjazannyh slov prostogo rasprostranennogo neoslozhennogo predlozhenija / G.V. Dorohina, D. S. Gnit'ko // «Suchasna informacija Ukraïna: informatika, ekonomika, filosofija»: materiali dopovidej konferencii, 12 13 travnja 2011 roku, Donec'k, 2011. – Т. 1. – С. 34-38.
6. Harlamov A.A. Perestrojka modeli mira, formiruemoj na materiale analiza tekstovoj informacii s ispol'zovanijem iskustvennyh nejronnyh setej, v uslovijah dinamiki vneshnej sredy / A.A. Harlamov, V.V. Raevskij // Rechevye tehnologii – 2008. – №3 – С. 27-35.
7. Kharlamov A.A., Yermolenko 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. – P. 383-388.
8. Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. 79 – 1982. – P. 2554-2558.
9. Radchenko A.N. Modelirovanie osnovnyh mehanizmov mozga. – L.: «Nauka» – 1969. – 211 s.
10. Harlamov A.A. Predsintaksicheskij analiz russko-anglijskih tekstov / A.A. Harlamov, T.V. Ermolenko, G.V. Dorohina, A.O. Zhuravljov // Programmaja inzhenerija – 2013. – № 10 – С. 37-47.
11. Harlamov A. A. Razrabotka komponenta sintaksicheskogo analiza predlozhenij russkogo jazyka dlja intellektual'noj sistemy obrabotki estestvenno-jazykovogo teksta / A.A. Harlamov, T.V. Ermolenko // Programmaja inzhenerija – 2013. – № 7 – С. 37-47.
12. Rahilina E.V. Kognitivnyj analiz predmetnyh imen: semantika i sochetaemost'. – М.: Russkie slovari – 2000. – 416 s.

RESUME

A. A. Kharlamov, T. V. Yermolenko

Text Analysis: Linguistics, Semantics, Pragmatics in the Cognitive Approach

Background: Currently, two basic approaches prevail in the automatic semantic analysis of texts: linguistic and statistical ones. The first one provides a very detailed analysis of the meaning of the text sentences, and the second one makes it possible to create a semantic representation of the whole text.

Materials and methods: Integration of the two approaches – linguistic and statistical ones – is represented in the paper. One can realize semantic and pragmatic analysis due to the integration. That allows to identification predicate structure chains for sentences of a text essential for representation of the text meaning.

Results: Predicate structures of sentences from T. I. Trofimova's textbook "Physics course", Moscow, "High School", 2001 are used as an example of an extended predicate structure of the text.

Conclusion: Combined application of fast statistical algorithms of text processing, as well as linguistic algorithms and knowledge bases in the form of dictionaries of valencies make it possible to obtain semantic representations of the whole text with accuracy typical for the linguistic approach.

Статья поступила в редакцию 06.07.2015.