*ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ И ТЕКСТА*

*УДК 004.912*

**С.Ф. Липницкий**

# МОДЕЛИРОВАНИЕ СМЫСЛОВОГО СОДЕРЖАНИЯ ТЕКСТА НА ОСНОВЕ СЛИЯНИЯ КОММУНИКАТИВНЫХ ФРАГМЕНТОВ

*Предлагается математическая модель процесса пересказа содержания тестовых сообщений на основе слияния коммуникативных фрагментов. Формально определяются понятия таких фрагментов, а также вербально-ассоциативных сетей в качестве моделей знаний о предметной области и пересказываемых текстах. Приводится описание алгоритма пересказа содержания текста.*

## Введение

Проблема пересказа содержания текстовых документов возникает при решении задач аналитико-синтетической обработки информации (например, при индексировании текстов, их аннотировании и реферировании). Традиционно синтез текста рассматривается как процесс последовательной генерации морфем, лексем, синтаксических фраз и, наконец, предложений, т. е. предполагается, что в иерархии этих языковых элементов каждый следующий складывается из предыдущих по известным синтаксическим правилам. Однако в монографии [1] показано, что основой использования языка человеком является его языковая память. Согласно этой концепции предложения при синтезе строятся из готовых хранящихся в памяти компонентов, названных коммуникативными фрагментами. Такие фрагменты не образуются по синтаксическим правилам, а извлекаются из памяти целиком.

Сложность моделирования и алгоритмизации процесса пересказа текста на основе слияния коммуникативных фрагментов заключается в их динамичности. Динамичность связана с изменчивостью границ фрагментов в предложениях, а также с лексическим и синтаксическим многообразием их представления.

В данной статье предлагается формальная модель, позволяющая алгоритмизировать процесс интерпретации содержания текстовых документов путем слияния коммуникативных фрагментов.

## 1. Формализация понятия коммуникативного фрагмента

В работе [1] под коммуникативными фрагментами понимаются «отрезки речи различной длины», которые человек использует при синтезе предложений. Эти отрезки «хранятся в памяти говорящего в качестве стационарных частиц его языкового опыта». С целью алгоритмизации синтеза текстов на основе слияния коммуникативных фрагментов, а также создания системы лингвистических словарей, используемых при синтезе, формализуем понятие коммуникативного фрагмента.

### *1.1. Определение коммуникативного фрагмента*

Рассмотрим произвольное предложение $\pi = a_1\, a_2 \ldots a_n$ из тематического корпуса текстов $Ct$ [2]. Подцепочку $f = a_1 a_2 \ldots a_m$ цепочки $\pi$ назовем коммуникативным фрагментом, если значения информативности вербально-ассоциативной связи между словами этой подцепочки удовлетворяют следующим двум условиям:

– для любых индексов $i, j$, таких, что $i \geq 1$, $j \leq m$, $i < j$, для значений информативности $I_{Ct}^{a_i a_j}$ вербально-ассоциативной связи между словами $a_i$ и $a_j$ выполняется соотношение $I_{Ct}^{a_i a_j} \geq I_{Ct}^{00}$, где $I_{Ct}^{00}$ – пороговое значение информативности;

– существует хотя бы одно слово $a_r \in \{a_1, a_2, ..., a_m\}$, такое, что справедливо соотношение $I_{Ct}^{a_r a_{m+1}} < I_{Ct}^{00}$.

Информативность $I_{Ct}^{ab}$ вербально-ассоциативной связи между словами *a* и *b* в тематическом корпусе текстов *Ct* вычисляется по формуле

$$I_{Ct}^{ab} = \frac{n_{Ct}^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_{Ct}^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} (n_{Ct}^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_{Ct}^{pq})}{n_{Cf}^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_{Cf}^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} (n_{Cf}^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_{Cf}^{pq})}, \quad (1)$$

где $n_{Ct}^{ab}$, $n_{Ct}^{cd}$, $n_{Ct}^{rs}$ и $n_{Ct}^{pq}$ – абсолютные частоты совместной встречаемости слов *a* и *b*, *c* и *d*, *r* и *s*, а также *p* и *q* соответственно в одном и том же предложении тематического корпуса текстов $Ct_i$; $n_{Cf}^{ab}$, $n_{Cf}^{cd}$, $n_{Cf}^{rs}$ и $n_{Cf}^{pq}$ – абсолютные частоты их совместного появления в предложениях полного корпуса текстов *Cf* [2]; $Par_a$ – множество всех словоизменений слова *a*; $Syn_a$ – множество всех его синонимов. Для хранения указанных словоизменений и синонимов используются следующие лингвистические словари:

– словарь словоизменительных парадигм $Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf}, a \in Par_a\}$, где $Par_a$ – множество всех словоизменений слова *a*, $W_{Cf}$ – множество всех словоформ из корпуса текстов *Cf*. В словаре $Dic_{par}$ для каждой словоформы представлены все ее словоизменения;

– словарь синонимичных словоформ, состоящий из совокупностей синонимичных слов: $Dic_{syn}^{a} = \{(a, Syn_a) \mid a \in W_{Cf}, a \in Syn_a\}$, где $Syn_a$ – множество всех синонимов слова *a*.

Оба словаря формируются «вручную» экспертом-лингвистом информационной системы.

Для хранения информации о частотах совместной встречаемости слов в одном и том же предложении полного корпуса текстов *Cf* и всех тематических корпусов будем использовать словарь вербально-ассоциативных пар слов $Dic_{ab} = \{\langle (a, b),\ n_{Cf}^{ab},\ n_{Ct_1}^{ab},\ n_{Ct_2}^{ab},\ \ldots,\ n_{Ct_n}^{ab} \rangle \mid a, b \in W_{Cf},\ n_{Cf}^{ab} \neq 0,\ n_{Ct_i}^{ab} \neq 0,\ i = \overline{1, n}\}$.

Словарь $Dic_{ab}$ создается программно следующим образом. Формируется множество $W_{Cf}^{ab}$ всех пар слов (*a*, *b*), таких, что слова *a* и *b* из каждой пары содержатся хотя бы в одном предложении полного корпуса текстов *Cf*. Для каждой пары слов вычисляется частота $n_{Ct}^{ab}$ ее появления в корпусе *Cf*, а также определяются частоты $n_{Ct_i}^{ab}$ ($i = \overline{1, n}$) появления пары слов (*a*, *b*) в тематических корпусах текстов $Ct_i$. На основе этих данных формируются кортежи вида $\langle (a, b), n_{Cf}^{ab}, n_{Ct_1}^{ab}, n_{Ct_2}^{ab}, \ldots, n_{Ct_n}^{ab} \rangle$, являющиеся компонентами создаваемого словаря $Dic_{ab}$.

Обозначим через *F* множество всех коммуникативных фрагментов в полном корпусе текстов *Cf*. Рассмотрим совокупность коммуникативных фрагментов $Dic_f = \{\langle f, I_{Ct_1}^{f}, I_{Ct_2}^{f}, \ldots, I_{Ct_n}^{f} \rangle \mid f \in F\}$, где $I_{Ct_i}^{f}$ ($i = \overline{1, n}$) – значение информативности коммуникативного фрагмента *f* в тематическом корпусе текстов $Ct_i$:

$$I_{Ct_i}^{f} = \frac{\sum\limits_{a \in f} I_{Ct_i}^{a}}{\sqrt{\sum\limits_{a \in f} (I_{Ct_i}^{a})^2}}. \quad (2)$$

Множество $Dic_f$ – это словарь коммуникативных фрагментов. В формуле (2) информативность $I_{Ct_i}^{a}$ слова *a* в тематическом корпусе текстов $Ct_i$ вычисляется по формуле, аналогичной выражению (1) из статьи [2]:

$$I^a_{Ct_i} = \frac{n^a_{Ct_i} + \sum\limits_{b\in Par_a,\, b\neq a} n^b_{Ct_i} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^c_{Ct_i} + \sum\limits_{d\in Par_c,\, d\neq c} n^d_{Ct_i})}{n^a_{Cf} + \sum\limits_{b\in Par_a,\, b\neq a} n^b_{Cf} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^c_{Cf} + \sum\limits_{d\in Par_c,\, d\neq c} n^d_{Cf})}. \qquad (3)$$

Для вычисления значений информативности $I^a_{Ct_i}$ используется словарь словоформ $Dic_a = \{\langle a,\ n^a_{Cf}\ ,\ n^a_{Ct_1}\ ,\ n^a_{Ct_2}\ ,\ \ldots\ ,\ n^a_{Ct_n}\rangle |\ a \in W_{Cf}\}$, где $n^a_{Cf}$ и $n^a_{Ct_i}$ – абсолютные частоты появления словоформы *a* соответственно в полном и *i*-м тематическом корпусах текстов.

Формируется словарь $Dic_a$ программно аналогично словарю $Dic_{ab}$ путем последовательного вычисления частот $n^a_{Cf}$ и $n^a_{Ct_i}$ ($i = \overline{1, n}$) появления словоформ в полном и тематических корпусах текстов.

Рассмотрим процесс создания словаря $Dic_f$ коммуникативных фрагментов. Он формируется в два этапа. На первом этапе каждое предложение $\pi = a_1 a_2 \ldots a_l$ ($l \geq 2$) входного текста разбивается на коммуникативные фрагменты в полном соответствии с их формальным определением. Вначале формируется множество всех пар слов вида $(a_r, a_j)$ из совокупности слов $\{a_1, a_2, \ldots, a_j\}$, причем в каждой паре $r < j$. Для каждой такой пары проверяется справедливость неравенства $I^{a_r a_j}_{Ct} < I^{00}_{Ct}$, где $I^{00}_{Ct}$ – пороговое значение информативности. Если неравенство выполняется при некотором значении $j$, то согласно упомянутому выше определению получаем коммуникативный фрагмент. Далее процедура повторяется аналогичным образом. В результате работы алгоритма получаем предложение в виде цепочки коммуникативных фрагментов $\pi = f_1 f_2 \ldots$ .

На втором этапе формируется список лексикографически упорядоченных коммуникативных фрагментов и для каждого фрагмента вычисляется множество значений информативности $\{ I^f_{Ct_1}\ ,\ I^f_{Ct_2}\ ,\ \ldots\ ,\ I^f_{Ct_n}\ \}$ по формулам (2) и (3).

### *1.2. Отношение контаминации*

Как отмечается в работе [1], различные коммуникативные фрагменты *контаминируются*, т. е. объединяются, перемещаются друг в друга, образуя новые языковые единицы. Для формализации понятия контаминации введем в рассмотрение следующее отношение.

Определим на множестве *F* строгий порядок (транзитивное и антирефлексивное отношение) $\prec_F$, такой, что для любых коммуникативных фрагментов (непустых цепочек) $f_1, f_2 \in F$ соотношение $f_1 \prec_F f_2$ выполняется тогда и только тогда, когда фрагмент $f_1$ получается из фрагмента $f_2$ путем исключения из него одного или более слов. Строгий порядок $\prec_F$ назовем *отношением контаминации* на множестве *F*.

Например, пусть $f_1$ – это коммуникативный фрагмент «национальный университет», а $f_2$ – фрагмент «национальный технический университет». В данном случае фрагмент $f_1$ получен из фрагмента $f_2$ исключением из него слова «технический».

### *1.3. Классы контаминированных коммуникативных фрагментов*

Пусть $\{ K^F_i \mid i = \overline{1, l}\ \}$ – множество подмножеств множества *F*, такое, что $F = \bigcup\limits_{i=1}^{l} K^F_i$ . Множества $K^F_i$ будем называть классами контаминированных коммуникативных фрагментов, если все они (т. е. при любом $i = \overline{1, l}$) являются максимально совершенными (по терминологии из монографии [3], определение 4.4). Это означает, что при синтезе пересказа содержания текста возможна взаимозаменяемость коммуникативных фрагментов внутри одного класса. Более того, согласно теореме 4.3 Хаусдорфа [3] для любого коммуникативного фрагмента из множества *F* существует свой класс, элементом которого данный фрагмент является.

## 2. Моделирование знаний о предметной области

Эффективность систем аналитико-синтетической обработки текстовой информации существенным образом зависит от их интеллектуальности, т. е. способности работать не только с данными, но и знаниями об объектах и явлениях предметной области. При автоматизации процесса пересказа текста необходимые знания накапливаются в базе знаний о предметной области. Построим модель представления таких знаний.

### *2.1. Вербально-ассоциативное отношение коммуникативных фрагментов предметной области*

Обозначим через *Ft* множество всех коммуникативных фрагментов произвольного тематического корпуса текстов $Ct \in Cf$. Определим на множестве *Ft* отношение толерантности $\Delta$ (рефлексивное и симметричное бинарное отношение), такое, что пара $(f, g)$ любых фрагментов из множества *Ft* является элементом отношения $\Delta$, т. е. $(f, g) \in \Delta$ тогда и только тогда, когда фрагменты *f* и *g* из этой пары содержатся хотя бы в одном предложении корпуса *Ct*. Отношение $\Delta$ будем называть *вербально-ассоциативным отношением коммуникативных фрагментов предметной области*, определяемой тематическим корпусом текстов *Ct*.

Вербально-ассоциативное отношение $\Delta$ – это отношение вербально-ассоциативной связи коммуникативных фрагментов в тематическом корпусе текстов *Ct*.

Пусть *f* и *g* – произвольные коммуникативные фрагменты предметной области, определяемой тематическим корпусом текстов *Ct*. Для вычисления силы вербально-ассоциативной связи между фрагментами *f* и *g* будем использовать следующую формулу из статьи [4]:

$$I_{Ct}^{fg} = \frac{\sum\limits_{a \in f,\, b \in g} I_{Ct}^{ab}}{\sqrt{\sum\limits_{a \in f,\, b \in g} (I_{Ct}^{ab})^2}}. \qquad (4)$$

В выражении (5) информативность $I_{Ct}^{ab}$ вычисляется по формуле (1).

### *2.2. Дискурсивная сочетаемость коммуникативных фрагментов*

Текст как связную последовательность предложений, обладающую семантическим единством, в лингвистике отождествляют с понятием дискурса [5]. Для получения «хороших» предложений при их синтезе из коммуникативных фрагментов будем использовать отношение дискурсивной сочетаемости таких фрагментов. Понятие этого отношения введем следующим образом.

Определим на множестве *Ft* всех коммуникативных фрагментов в тематическом корпусе текстов *Ct* антирефлексивное бинарное отношение $\Lambda$, такое, что для любых фрагментов $f, g \in Ft$ соотношение $(f, g) \in \Lambda$ выполняется тогда и только тогда, когда в некотором тексте $T \in Ct$ существует предложение $\pi$, в котором коммуникативный фрагмент *f* непосредственно предшествует фрагменту *g*. Отношение $\Lambda$ будем называть *отношением дискурсивной сочетаемости коммуникативных фрагментов* в тематическом корпусе текстов *Ct*.

Дискурсивно-сочетаемыми являются, например, следующие пары коммуникативных фрагментов: «используется при» и «капитальном строительстве», «отложен ежегодный визит» и «председателя объединения».

Упорядоченные пары $(f, g) \in \Lambda$ будем хранить в специальном списке – словаре дискурсивно-сочетаемых коммуникативных фрагментов:

$$Dic_{fg} = \{(f, g) \mid f \in Ft, g \in Ft, (f, g) \in \Lambda\}. \qquad (5)$$

Основой для формирования словаря $Dic_{fg}$ являются несовпадающие предложения полного корпуса текстов, представленные в виде цепочек, которые состоят из коммуникативных фрагментов.

### *2.3. Вербально-ассоциативная сеть предметной области*

Рассмотрим граф вербально-ассоциативного отношения Δ. Пометим каждую вершину *f* этого графа значением информативности $I_{Ct}^{f}$ коммуникативного фрагмента (с учетом синонимии и словоизменения), а каждое ребро (*f, g*) – значением информативности $I_{Ct}^{fg}$ вербально-ассоциативной связи фрагментов *f* и *g* (также учитывая синонимию и словоизменения). Пусть ( *f, g*) – произвольное ребро этого графа. Если ( *f, g*) ∈ Λ, то для всех таких пар ( *f, g*) вершины *f* и *g* соединим дугой, направленной от *f* к *g*. Обозначим полученный смешанный граф через $Net_{Ct}$.

Граф $Net_{Ct}$ назовем *вербально-ассоциативной сетью предметной области*, определяемой тематическим корпусом текстов *Ct*.

Сеть $Net_{Ct}$ является моделью поискового образа тематического корпуса текстов *Ct*. Построим этот поисковый образ в виде множества коммуникативных фрагментов, вербально-ассоциативных пар таких фрагментов, в которой фрагментам приписаны значения их информативности, а парам – значения информативности вербально-ассоциативных связей между их фрагментами:

$$\text{ПО}_{Ct} = \{(f, I_{Ct}^{f}), \ldots, ((g,h), I_{Ct}^{gh}), \ldots, ((\overrightarrow{p,q}), I_{Ct}^{pq}), \ldots \mid I_{Ct}^{f} > I_{Ct}^{0}; I_{Ct}^{gh}, I_{Ct}^{pq} > I_{Ct}^{00}\}, \quad (6)$$

где $I_{Ct}^{0}$, $I_{Ct}^{00}$ – пороговые значения информативности коммуникативного фрагмента и вербально-ассоциативной связи этих фрагментов соответственно, стрелка над парой фрагментов (*p, q*) означает, что (*p, q*) ∈ Λ.

### *2.4. Базовые и связующие коммуникативные фрагменты*

В зависимости от информативности будем различать базовые и связующие коммуникативные фрагменты предметной области, определяемой тематическим корпусом текстов *Ct*. Обозначим через $I_{Ct}^{0}$ пороговое значение информативности коммуникативного фрагмента. Тогда коммуникативный фрагмент *f* будем называть *базовым*, если значение его информативности $I_{Ct}^{f}$ удовлетворяет неравенству $I_{Ct}^{f} \geq I_{Ct}^{0}$. Если же $I_{Ct}^{f} < I_{Ct}^{0}$, то фрагмент *f* назовем *связующим*. Связующим, например, является коммуникативный фрагмент «предлагается новый подход к решению проблемы», а базовым – фрагмент «принятия решений в условиях неопределенности».

Обозначим через $Ft_{\text{баз.}}$ множество всех базовых коммуникативных фрагментов, а через $Ft_{\text{св.}}$ – множество всех связующих. Тогда множество всех коммуникативных фрагментов предметной области – это объединение множеств базовых и связующих фрагментов, т. е. $Ft = Ft_{\text{баз.}} \cup Ft_{\text{св.}}$.

## 3. Моделирование знаний о пересказываемом тексте

Основой для моделирования знаний о пересказываемом тексте является вербально-ассоциативная сеть предметной области, тематический корпус которой релевантен этому тексту. С использованием сети предметной области строится вербально-ассоциативная сеть пересказываемого текста.

### *3.1. Вербально-ассоциативная сеть пересказываемого текста*

Пусть *Q* ∈ *Cf* – некоторый текст, а *Ct* – релевантный ему тематический корпус текстов. Вычислим информативность каждого предложения π текста *Q* по формуле, являющейся аналогом выражения (1):

$$I_{Q}^{\pi} = I_{Ct}^{\pi} = \frac{\sum\limits_{a \in \pi} I_{Ct}^{a}}{\sqrt{\sum\limits_{a \in \pi} (I_{Ct}^{a})^{2}}}. \quad (7)$$

Обозначим через $I_Q^0$ некоторое пороговое значение информативности предложений текста $Q$. Если $I_Q^\pi \geq I_Q^0$, то предложение $\pi$ будем считать информативным. Исключив из текста все неинформативные предложения, т. е. такие, для которых $I_Q^\pi < I_Q^0$, получим кортеж информативных предложений $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$. Каждое предложение $\pi_i$ ($i = \overline{1, m}$) представим в виде цепочки коммуникативных фрагментов $\pi_i = f_1 f_2 \ldots$. Обозначим через $\Delta_T$ сужение вербально-ассоциативного отношения $\Delta$ на множество $F_T$ всех коммуникативных фрагментов кортежа предложений $T$, т. е. $\Delta_T = \Delta \cap (F_T \times F_T)$, а через $\Lambda_T$ – сужение отношения дискурсивной сочетаемости коммуникативных фрагментов $\Lambda$ на это же множество. Тогда вербально-ассоциативную сеть пересказываемого текста $Q$ построим следующим образом.

Рассмотрим вербально-ассоциативную сеть предметной области *Net*. Исключим из сети *Net* все ребра $(f, g)$, такие, что $(f, g) \notin \Delta_T$, и все дуги $(r, s)$, для которых $(r, s) \notin \Lambda_T$. Исключим также из сети *Net* инцидентные исключенным ребрам и дугам вершины. Полученный граф назовем *вербально-ассоциативной сетью пересказываемого текста Q*.

### *3.2. Поиск релевантного тематического корпуса текстов*

Текст $Q \in Cf$ – это запрос на поиск релевантного ему тематического корпуса текстов. Исключим из всех поисковых образов тематических корпусов текстов значения информативности коммуникативных фрагментов и вербально-ассоциативных пар таких фрагментов, т. е. преобразуем выражение (6) к виду

$$\text{ПО}_{Ct_i} = \{f, \ldots, (g, h), \ldots, (\overrightarrow{p, q}), \ldots \mid f, g, h, p, q \in Ct_i\}. \quad (8)$$

Аналогично представим поисковое предписание, т. е. поисковый образ текста $Q$:

$$\text{ПП}_Q = \{f, \ldots, (g, h), \ldots, (\overrightarrow{p, q}), \ldots \mid f, g, h, p, q \in Q\}. \quad (9)$$

Реализуем процесс поиска тематического корпуса текстов, используя в качестве критерия выдачи косинус угла между векторами поискового предписания и поискового образа документа в евклидовом пространстве $E$. Этот критерий применяется в большинстве известных информационных систем.

Обозначим через $W$ множество всех различных коммуникативных фрагментов, вербально-ассоциативных пар фрагментов и упорядоченных пар фрагментов вида $(\overrightarrow{p, q})$, входящих в поисковые образы всех тематических корпусов текстов. Пусть их количество равно $n$. Лексикографически упорядочим все элементы множества $W$, т. е. представим $W$ в виде кортежа $W = \langle w_1, w_2, \ldots, w_n \rangle$. Для каждого тематического корпуса текстов $Ct \in Cf$ построим вектор его поискового образа в пространстве $E$: $\mathbf{F}_{Ct} = (p_1, p_2, \ldots, p_n)$, где $p_i = 1$, если элемент $w_i$ входит в этот поисковый образ, в противном случае $p_i = 0$. Аналогично представим вектор поискового предписания, построенного для запроса $Q$: $\mathbf{F}_Q = (q_1, q_2, \ldots, q_n)$. Тогда для вычисления меры близости между векторами $\mathbf{F}_{Ct}$ и $\mathbf{F}_Q$ воспользуемся критерием выдачи

$$\cos\varphi = \frac{\mathbf{F}_{Ct}\mathbf{F}_Q}{|\mathbf{F}_{Ct}||\mathbf{F}_Q|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}}\sqrt{\sum_{i=1}^{n} q_i^{\,2}}}. \quad (10)$$

При реализации информационной системы критерий (10) целесообразно преобразовать следующим образом. Пусть $l$ – количество совпавших элементов поискового образа $\text{ПО}_{Ct}$ и поискового предписания $\text{ПП}_Q$, $n_{Ct}$ – количество элементов в множестве $\text{ПО}_{Ct}$, а $n_Q$ – их количество в множестве $\text{ПП}_Q$. Тогда критерий (10) примет вид

$$\cos\varphi = \frac{l}{\sqrt{n_{Ct} n_Q}}. \qquad (11)$$

Результатом поиска будет тематический корпус текстов *Ct*, для которого значение $\cos\varphi$ является наибольшим из всех значений, таких, что $\cos\varphi \geq \eta_0$.

### *3.3. Фрагментно-слотовый шаблон предложения*

Пусть имеется предложение $\pi = f_1 f_2 \ldots f_l$. Цепочку, полученную из предложения $\pi$ заменой его базовых коммуникативных фрагментов слотами («пустыми» фрагментами), будем называть *фрагментно-слотовым шаблоном предложения* $\pi$.

Фрагментно-слотовые шаблоны предложений создаются в автоматизированном режиме: сначала на основе специально подготовленных текстов программно формируется совокупность фрагментно-слотовых шаблонов, а затем они корректируются экспертом-лингвистом информационной системы. При синтезе предложения слоты заменяются коммуникативными фрагментами. При этом может учитываться отношение контаминации.

### *3.4. Фрагментно-слотовый шаблон текста*

Пусть имеется текст в виде кортежа предложений $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$, такой, что каждому предложению $\pi_i$ ($i = \overline{1, m}$) соответствует его фрагментно-слотовый шаблон $H_i$. Рассмотрим кортеж фрагментно-слотовых шаблонов предложений $SH_T = \langle H_1, H_2, \ldots, H_m \rangle$. В качестве характеристики связности фрагментно-слотовых шаблонов предложений текста $T$ определим на множестве $SH_T$ антирефлексивное бинарное отношение $\Omega_T$, элементами которого являются пары соседних фрагментно-слотовых шаблонов предложений из множества $Sh_T$, т. е. $\Omega_T = \{(H_i, H_{i+1}) | i = \overline{1, m-1}\}$. Отношение $\Omega_T$ назовем *фрагментно-слотовым шаблоном текста T*.

Вершины-слоты соседних фрагментно-слотовых шаблонов предложений текста могут быть помечены символом «′» или символом «″». Это означает, что после заполнения первого слота коммуникативным фрагментом $f'$ второй слот должен быть заполнен коммуникативным фрагментом $f''$, таким, что выполняется соотношение $f'' \prec_F f'$, т. е. $f'$ и $f''$ принадлежат некоторому классу контаминированных коммуникативных фрагментов.

### *3.5. Фрагментно-слотовый шаблон предметной области*

Для формирования фрагментно-слотового шаблона предметной области необходимо предварительно подготовить множество $\{T_i \mid i = \overline{1, r}\}$ некоторых «хороших» текстов. Обозначим через $\Omega_{T_i}$ фрагментно-слотовый шаблон текста $T_i$. Тогда объединение множеств $\Omega_{Ct} = \bigcup_{i=1}^{r} \Omega_{T_i}$ назовем *фрагментно-слотовым шаблоном предметной области*, определяемой тематическим корпусом текстов *Ct*.

## 4. Описание алгоритма пересказа содержания текста

Пусть $T \in Cf$ – некоторый текст, $Ct$ – релевантный ему тематический корпус текстов, а $Q = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$ – кортеж информативных предложений этого текста. Алгоритм пересказа содержания текста $T$ работает следующим образом.

Из вербально-ассоциативной сети предметной области $Net_{Ct}$ исключим все вершины, которые соответствуют базовым коммуникативным фрагментам, отсутствующим в предложениях кортежа $Q$. Удалим также из сети $Net_{Ct}$ инцидентные исключенным вершинам ребра и дуги. Полученный граф обозначим через $Net_T^+$.

Из фрагментно-слотового шаблона предметной области $\Omega_{Ct}$ сформируем множество начальных фрагментно-слотовых шаблонов предложений $Form_1 = \{H_1, H_2, \ldots\}$. Построим для каждого шаблона из множества $Form_1$ его вербально-ассоциативную сеть $Net_{H_1}$. Сформируем множество всех орцепей $Req^{12}_{H_1}$ длиной 1 и 2 графа $Net_{H_1}$. Элементами множества $Req^{12}_{H_1}$ являются орцепи вида $f_1s_1$, $s_2g_2$, $f_3s_3g_3$, где $f_1s_1$ – конечные фрагменты предложения, а $s_2g_2$ – начальные; $s_1$, $s_2$ и $s_3$ – слоты; $f_3$ и $g_3$ – коммуникативные фрагменты, не являющиеся начальными и конечными в фрагментно-слотовом шаблоне $H_1$. Орцепи из множества $Req^{12}_{H_1}$ являются запросами на поиск релевантных орцепей в графе $Net^{+}_{T}$ с целью заполнения найденных слотов коммуникативными фрагментами. Орцепи графов $Net^{+}_{T}$ и $Net_{H_1}$ считаем совпавшими, если совпали соответствующие коммуникативные фрагменты. Например, совпавшими являются орцепи *fsg* и *frg*, где *r* – коммуникативный фрагмент, заполняющий слот *s*. Если все слоты фрагментно-слотового шаблона заполнены, то полученное в результате предложение включаем в некоторое множество *Sen*. Далее описанную процедуру повторяем для всех остальных шаблонов из множества $Form_1$.

В сформированном множестве *Sen* ищем релевантное графу $Net^{+}_{T}$ предложение $\pi_l$, полученное из шаблона $H_l$. Оно является началом формируемого пересказа текста.

Далее создаем множество $Form_2$, состоящее из всех фрагментно-слотовых шаблонов предложений *H*, таких, что $(H_l, H) \in \Omega_T$. Процессы заполнения слотами шаблонов из множеств $Form_2$, $Form_3$ и т. д. аналогичны такой процедуре для шаблонов из множества $Form_1$.

П р и м е р . Рассмотрим следующий текстовый фрагмент из статьи [5]:

*В ходе интерпретации воссоздается мысленный мир, в котором, по презумпции интерпретатора, автор конструировал дискурс и в котором описывается реальное или нереальное положение дел. При этом анализ дискурса предполагает наличие языкового инструментария, при котором исследователь обращается не только к собственным лингвистическим знаниям, но также и общему фоновому знанию о реальном мире, поскольку в процессах понимания и порождения речи взаимодействуют все базы данных, хранящиеся в когнитивном аппарате человека. В основном анализу подвергаются не отдельные слова, а более крупные объединения (предложения или даже целые тексты), так как известно, что трансляция смысла ведется с помощью именно текстов. Именно поэтому текст стал объектом исследования отдельного направления языкознания, лингвистики текста, которое стремится выйти за рамки предложения. Дискурс может членироваться на высказывания, в то время как существуют другие объединения, которые складываются из последовательных предложений, например текст.*

На начальном этапе генерации пересказа данного текста формируется множество фрагментно-слотовых шаблонов предложений $Form_1$: $Form_1$ = {⟨обсуждается проблема / ◊ ⟩, ⟨рассматриваются вопросы / ◊ / путем использования ◊ ⟩, ⟨речь идет о / ◊ ⟩, в работе / приведены результаты / ⟩,...}, где символом «/» обозначены разделители между коммуникативными фрагментами, а символом «◊» – слоты.

После построения для всех фрагментно-слотовых шаблонов их вербально-ассоциативных сетей и поиска коммуникативных фрагментов для заполнения слотов выбранного шаблона получим начальное предложение пересказа исходного текста в виде цепочки коммуникативных фрагментов: *Рассматриваются вопросы / конструирования дискурса / путем использования / лингвистических знаний / о реальном мире.*

На последующих этапах синтеза выходного текста процесс поиска фрагментно-слотовых шаблонов предложений и заполнения их слотов повторяется аналогичным образом. В результате получим следующие предложения сформированного пересказа текста: *Для изучения дискурса / возникло направление / в языкознании / – / лингвистика текста. При конструировании / анализируются фрагменты / данного дискурса. Элементами дискурса / являются высказывания.*

**Заключение**

Модель пересказа содержания текста на основе слияния коммуникативных фрагментов может быть использована при решении следующих задач:

– индексирование текстовых документов и запросов на поиск информации. В индексируемом тексте выявляются коммуникативные фрагменты. Поисковый образ проиндексированного текста или запроса на поиск информации состоит из информативных фрагментов, каждому соответствует значение информативности;

– автоматическое реферирование и аннотирование текста. Одним из этапов при реализации этих процессов является построение шаблона входного текста. На основе этого шаблона генерируется выходной текст.

**Список литературы**

1. Гаспаров, Б.М. Язык, память, образ. Лингвистика языкового существования / Б.М. Гаспаров. – М. : Новое литературное обозрение, 1996. – 352 с.

2. Липницкий, С.Ф. Индексирование текстовой информации на основе моделирования вербальных ассоциаций / С.Ф Липницкий // Информатика. – 2012. – № 3. – С. 94–102.

3. Шрейдер, Ю.А. Равенство. Сходство. Порядок / Ю.А. Шрейдер. – М. : Наука, 1971. – 256 с.

4. Липницкий, С.Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С.Ф Липницкий // Информатика. – 2011. – № 4. – С. 21–28.

5. Темнова, Е.В. Современные подходы к изучению дискурса / Е.В. Темнова // Язык, сознание, коммуникация : сб. статей. – М. : МАКС Пресс, 2004. – Вып. 26. – С. 24–32.

**Поступила 20.05.2016**

*Объединенный институт проблем*
*информатики НАН Беларуси,*
*Минск, Сурганова, 6*
*e-mail: lipn@newman.bas-net.by*

**S.F. Lipnitsky**

## MODELING THE TEXT CONTENT BY MERGING COMMUNICATIVE FRAGMENTS

Mathematical model of the process of retelling the content of the test messages by the convergence of communicative fragments is proposed. Formally defined concepts such fragments, as well as the verbal-associative networks as models of domain knowledge and retell the texts. Algorithm description of the content of the text presented.