ISSN 1816-0301 (print)
УДК 004.912

Поступила в редакцию 18.10.2017
Received 18.10.2017

С. Ф. Липницкий

Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь

МОДЕЛИРОВАНИЕ АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ И КРАТКИХ СООБЩЕНИЙ НА ОСНОВЕ ВЕРБАЛЬНЫХ АССОЦИАЦИЙ

Аннотация. Предлагается математическая модель процессов анализа текстовых документов и кратких сообщений на основе вербальных ассоциаций, т. е. семантических связей между словами и словосочетаниями, соответствующих ассоциативным отношениям между обозначаемыми ими сущностями. Целью анализа является создание для каждого анализируемого текста кортежа вербально-ассоциативных сетей всех предложений текста, а также построение вербально-ассоциативной сети самого текста. Вербально-ассоциативная сеть предложения – это граф, вершинами которого являются все коммуникативные фрагменты предложения (устойчивые словосочетания), а ребра соответствуют вербально-ассоциативным связям между ними. В случае текста в целом вершинами его вербально-ассоциативной сети являются все коммуникативные фрагменты текста.

Ключевые слова: алгоритм, вербально-ассоциативная сеть, дискурсивная сочетаемость, информативность, коммуникативный фрагмент, математическая модель, текст

Для цитирования. Липницкий, С. Ф. Моделирование анализа текстовых документов и кратких сообщений на основе вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2018. – Т. 15, № 1. – С. 70–80.

S. F. Lipnitsky

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus

MODELING OF ANALYSIS OF TEXT DOCUMENTS AND BRIEF COMMUNICATIONS BASED ON VERBAL ASSOCIATIONS

Abstract. A mathematical model is proposed for the analysis of text documents and short messages based on verbal associations, that is, the semantic links between words and word-combinations corresponding to associative relations between the entities they designate. The aim of the analysis is to create for each analyzed text a tuple of verbal-associative networks of all sentences of the text, as well as constructing a verbal-associative network of the text itself. The verbal-associative supply network is a graph whose vertices are all communicative fragments of the sentence (stable word combinations), and the edges correspond to verbal-associative connections between them. In the case of the text as a whole, the vertices of its verbal-associative network are all communicative fragments of the text.

Keywords: algorithm, communicative fragment, discourse compatibility, informative, mathematical model, text, verbal-associative network

For citation. Lipnitsky S. F. Modeling of analysis of text documents and brief communications based on verbal associations. *Informatics*, 2018, vol. 15, no. 1, pp. 70–80 (in Russian).

Введение. При решении задач компьютерной лингвистики (например, автоматического перевода, информационного поиска, автоматического аннотирования и реферирования) актуальной является проблема анализа естественно-языковых текстов. Для ее решения необходимы модели и алгоритмы распознавания структуры входных цепочек, вычисления их информативности, построения баз знаний и двуязычных словарей.

В настоящей статье решаются задачи создания вербально-ассоциативных сетей текстовых документов и кратких сообщений на основе вербальных ассоциаций. Вербально-ассоциативная сеть представляет собой граф, вершинами которого являются устойчивые словосочетания, а ребра соответствуют вербально-ассоциативным связям между ними.

Вербально-ассоциативные сети формируются при анализе текстов. Процесс анализа текста реализуется в три этапа. На первом этапе каждое предложение входного текста разбивается на коммуникативные фрагменты. При этом используется сформированный словарь таких фрагментов. Если соответствующее словосочетание в словаре отсутствует, то словарь оперативно пополняется новым коммуникативным фрагментом путем вычисления информативности вербально-ассоциативных связей между словами, а также между словами и словосочетаниями (субтекстами). На втором этапе создается кортеж вербально-ассоциативных сетей всех предложений входного текста. На третьем этапе формируется вербально-ассоциативная сеть текста.

На каждом из этапов вычисляются значения информативности словоформ текста и коммуникативных фрагментов, а также оценивается сила (информативность) их вербально-ассоциативной связи. Краткое сообщение при вычислении информативности словоформ и вербальных ассоциаций заменяется релевантным тематическим корпусом текстов.

Информативность слов, предложений и субтекстов. Определим формально основные понятия, связанные с анализом текстовых документов и кратких сообщений.

Входной язык. В информационных системах различают, как правило, три языка: входной, внутренний и выходной. Для определения понятия входного языка введем в рассмотрение формальную порождающую грамматику $G = \langle V, N, I, R\rangle$, где V – непустое множество терминальных элементов (слов), $N = \{I, '\}$ – множество нетерминальных элементов, I – начальный символ, а R – схема грамматики, т. е. множество правил вывода вида $\alpha \to \beta$ (α и β – различные непустые цепочки в словаре $V \cup N$). Схему R грамматики G определим следующим образом:

– для любого слова $a \in V$ существуют правила вывода $I \to a'$ и $a' \to a$;

– все остальные правила вывода имеют вид $a' \to a'b'$ или $a' \to b'a'$, где $a, b \in V$.

Для удобства в состав нетерминальных символов введен символ «'» (штрих).

Язык $L(G)$, порождаемый грамматикой G, назовем входным языком, а его цепочки – предложениями входного языка.

Пример входного языка. Пусть $G = \langle V, N, I, R\rangle$, где $V = \{$*быстрыми, интеллектуальные, информационные, развиваются, темпами, технологии*$\}$; $N = \{I, '\}$; $R = \{I \to$ *быстрыми'*, $I \to$ *интеллектуальные'*, $I \to$ *информационные'*, $I \to$ *развиваются'*, $I \to$ *темпами'*, $I \to$ *технологии'*, *быстрыми'* $\to$ *быстрыми*, *интеллектуальные'* $\to$ *интеллектуальные*, *информационные'* $\to$ *информационные*, *развиваются'* $\to$ *развиваются*, *темпами'* $\to$ *темпами*, *технологии'* $\to$ *технологии*, *технологии'* $\to$ *технологии' развиваются'*, *технологии'* $\to$ *информационные' технологии'*, *технологии'* $\to$ *интеллектуальные' технологии'*, *развиваются'* $\to$ *развиваются' темпами'*, *темпами'* $\to$ *быстрыми' темпами'*$\}$.

Грамматика G порождает, в частности, следующие цепочки:

– *интеллектуальные информационные технологии;*

– *темпами;*

– *информационные технологии развиваются;*

– *интеллектуальные информационные технологии;*

– *интеллектуальные информационные технологии развиваются быстрыми темпами.*

Информативность слов. Пусть $D \in L(G)$ – произвольный текстовый документ, порождаемый грамматикой G. Информативность I_D^a слова a из текста D определена в статье [1] как вероятность того, что слово a имеется в данном текстовом документе при условии, что оно содержится в полном корпусе текстов. (Полный корпус текстов – это объединение всех тематических корпусов, т. е. совокупностей текстов по конкретным тематикам.)

Пусть имеются тематические корпусы текстов Ct_i ($i = \overline{1, n}$; $n \geq 1$) и полный корпус текстов $Cf = \bigcup_{i=1}^{n} Ct_i$. Рассмотрим совокупность событий (в вероятностном смысле):

S_D – некоторая словоформа a извлечена случайным образом из текстового документа D ($D \in Cf$);

H_D – появление документа D;

S_{Cf} – словоформа a содержится в полном корпусе текстов Cf.

Тогда $I_D^a = P(S_D / S_{Cf})$, где $P(S_D / S_{Cf})$ – условная вероятность того, что словоформа a извлечена из текста D при условии, что она уже извлечена из полного корпуса текстов Cf. Эта вероятность вычисляется следующим образом:

$$P(S_D / S_{Cf}) = \frac{P(S_D \cdot S_{Cf})}{P(S_{Cf})} = \frac{P(S_D) \cdot P(S_{Cf} / S_D)}{P(S_{Cf})}.$$

Учитывая, что $P(S_{Cf} / S_D) = 1$, и воспользовавшись формулой полной вероятности, получим

$$P(S_D / S_{Cf}) = \frac{P(S_D / H_D)}{P(S_{Cf})} \cdot P(H_D).$$

При достаточно больших объемах полного корпуса текстов Ct и текстового документа D можно считать, что

$$P(S_D / H_D) \approx \frac{n_D}{N_D},\ P(S_{Cf}) \approx \frac{n_{Cf}}{N_{Cf}},\ P(H_D) \approx \frac{N_D}{N_{Cf}},$$

где n_D, n_{Cf} – частоты встречаемости (с учетом словоизменения и синонимии) словоформы a в тексте D и полном корпусе текстов Cf; N_D, N_{Cf} – число вхождений всех словоформ в D и Cf соответственно. Тогда будем считать, что формула для вычисления информативности I_D^a словоформы a в тексте D имеет вид

$$I_D^a = \frac{n_D}{n_{Cf}}. \quad (1)$$

Информация о частотах словоформ в корпусах текстов, а также о парадигматике и синонимии слов хранится в специальных лингвистических словарях:

– частотном словаре словоформ

$$Dic_a = \{\langle a,\ n_{Cf}^a,\ n_{Ct_1}^a,\ n_{Ct_2}^a, \ldots,\ n_{Ct_n}^a \rangle |\ a \in W_{Cf}\},$$

в котором каждой словоформе приписаны частоты ее встречаемости n_{Cf}^a, $n_{Ct_1}^a$, $n_{Ct_2}^a$, …, $n_{Ct_n}^a$ во всех корпусах текстов (W_{Cf} – множество всех словоформ полного корпуса текстов Cf);

– словаре словоизменительных парадигм

$$Dic_{par} = \{(a, Par_a) | a \in W_{Cf}, a \in Par_a\},$$

состоящем из пар ⟨*словоформа*, *парадигма*⟩. В позиции парадигмы Par_a представлены все словоизменения словоформы a;

– словаре синонимичных словоформ

$$Dic_{syn} = \{(a, Syn_a) | a \in W_{Cf}, a \in Syn_a\},$$

включающем в себя пары ⟨*словоформа*, *синонимичные словоформы*⟩, в которых каждой словоформе a соответствует множество ее синонимов Syn_a.

Используя лингвистические словари, формулу (1) перепишем в виде

$$I_D^a = \frac{n_D^a + n_D^{Par_a} + n_D^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}. \quad (2)$$

В формуле (2) $n_D^{Par_a}$ – это число вхождений всех словоформ текста D, являющихся словоизменениями словоформы a, т. е.

$$n_D^{Par_a} = \sum_{b \in Par_a,\, b \neq a} n_D^b .$$

Параметр $n_D^{Syn_a}$ означает количество синонимов словоформы *a* в тексте *D*:

$$n_D^{Syn_a} = \sum_{c \in Syn_a,\, c \neq a} n_D^c .$$

Аналогичный смысл имеют параметры $N_{Cf}^{Par_a}$ и $N_{Cf}^{Syn_a}$:

$$N_{Cf}^{Par_a} = \sum_{b \in Par_a,\, b \neq a} n_{Cf}^b , \quad N_{Cf}^{Syn_a} = \sum_{c \in Syn_a,\, c \neq a} n_{Cf}^c .$$

Если $Q \in L(G)$ – краткое сообщение, то формула (2) примет вид

$$I_Q^a = \frac{n_{Ct}^a + n_{Ct}^{Par_a} + n_{Ct}^{Syn_a}}{n_{Cf}^a + N_{Cf}^{Par_a} + N_{Cf}^{Syn_a}}, \tag{3}$$

где *Ct* – тематический корпус текстов, релевантный сообщению *Q*.

Информативность предложений. Пусть *T* – текстовый документ или краткое сообщение, $W_{Cf} = \langle a_1, a_2, ..., a_l \rangle$ – кортеж всех слов полного корпуса текстов *Cf*, *E* – *l*-мерное евклидово пространство, а π – произвольное предложение (или словосочетание) текста *T*. Представим это предложение вектором в пространстве *E*: $\mathbf{\Pi} = (I_\pi^{a_1}, I_\pi^{a_2}, ..., I_\pi^{a_l})$, где $I_\pi^{a_1}, I_\pi^{a_2}, ..., I_\pi^{a_l}$ – значения информативности слов предложения π. При этом компонента вектора $\mathbf{\Pi}$ равна нулю, если соответствующего слова нет в предложении π. С учетом рассмотренных обозначений нормализованную информативность I_T^π предложения π можно интерпретировать как проекцию вектора $\mathbf{e} = (1, 1, ..., 1)$ на направление вектора $\mathbf{\Pi}$, т. е. скалярное произведение векторов $\mathbf{\Pi}$ и $\mathbf{e}$:

$$I_T^\pi = \frac{\sum_{i=1}^{l} I_\pi^{a_i}}{\sqrt{\sum_{i=1}^{l} (I_\pi^{a_i})^2}} . \tag{4}$$

При реализации алгоритма вычисления информативности предложений удобно пользоваться следующей формулой, полученной из выражения (4):

$$I_T^\pi = \frac{I_1 + I_2 + ...}{\sqrt{I_1^{\,2} + I_2^{\,2} + ...}}, \tag{5}$$

где I_1, I_2, ... – значения информативности всех слов предложения π.

Информативность субтекстов. Субтекст – это «компонент связного текста, развивающий одну из его основных тем» [2]. Обозначим через $Sub = \pi_1 \pi_2 ...$ произвольный субтекст текста *T*. Представим субтекст *Sub* в виде вектора аналогично вектору $\mathbf{\Pi}$: $\mathbf{Sub} = (I_T^{\pi_1}, I_T^{\pi_2}, ..., I_T^{\pi_j})$. Тогда информативность субтекста *Sub* имеет вид

$$I_T^{Sub} = \frac{\sum_{i=1}^{m} I_T^{\pi_i}}{\sqrt{\sum_{i=1}^{m} (I_T^{\pi_i})^2}} . \tag{6}$$

По аналогии с формулой (5) выражение (6) представим как

$$I_T^{Sub} = \frac{I_T^1 + I_T^2 + \ldots}{\sqrt{(I_T^1)^2 + (I_T^2)^2 + \ldots}}, \quad (7)$$

где I_T^1, I_T^2, ... – значения информативности всех предложений текста *T*.

Информативность вербальных ассоциаций между словами, предложениями и субтекстами. Приведем формулы для вычисления информативности вербально-ассоциативной связи между словами и словосочетаниями.

Вербальные ассоциации. Под вербальными ассоциациями в лингвистике и психолингвистике понимают семантические связи между словами, которые соответствуют ассоциативным отношениям между обозначаемыми ими сущностями в реальном мире. Для моделирования вербальных ассоциаций определим понятие отношения вербально-ассоциативной связи слов.

Рассмотрим полный корпус текстов *Cf*. Обозначим через *W* множество всех слов корпуса текстов *Cf*. Тогда отношение толерантности Θ (рефлексивное и симметричное бинарное отношение) на множестве *W* назовем отношением вербально-ассоциативной связи слов в полном корпусе текстов *Cf*, если любая упорядоченная пара слов (a, b) из множества *W* является элементом отношения Θ тогда и только тогда, когда слова *a* и *b* из этой пары содержатся хотя бы в одном предложении корпуса *Cf*. Если пара (a, b) любых слов из множества *W* является элементом отношения Θ, т. е. $(a, b) \in W$, то (a, b) будем называть вербально-ассоциативной парой слов.

Пусть имеется текстовый документ *D*. Обозначим через W_D множество всех слов текста *D*. Рассмотрим сужение Θ_D отношения Θ на множество W_D, т. е. $\Theta_D = \Theta \cap (W_D \times W_D)$. Отношение Θ_D назовем отношением вербально-ассоциативной связи слов в текстовом документе *D*.

Информативность вербальных ассоциаций между словами. Понятие информативности вербально-ассоциативной связи между словами *a* и *b* в текстовом документе *D* определим по аналогии с понятием информативности слова:

$$I_D^{ab} = \frac{n_D^{ab} + n_D^{Par_{ab}} + n_D^{Syn_{ab}}}{n_{Cf}^{ab} + N_{Cf}^{Par_{ab}} + N_{Cf}^{Syn_{ab}}}. \quad (8)$$

При вычислении значений вербальных ассоциаций между словами будем использовать частотный словарь вербально-ассоциативных пар слов

$$Dic_{ab} = \{\langle (a, b),\ n_{Cf}^{ab},\ n_{Ct_1}^{ab},\ n_{Ct_2}^{ab},\ \ldots,\ n_{Ct_n}^{ab} \rangle \mid a, b \in W_{Cf},\ n_{Cf}^{ab} \neq 0,\ n_{Ct_i}^{ab} \neq 0,\ i = \overline{1, n}\},$$

где n_{Cf}^{ab}, $n_{Ct_i}^{ab}$ – абсолютные частоты совместной встречаемости слов *a* и *b* в одном и том же предложении полного *Cf* и *i*-го тематического корпуса текстов Ct_i ($i = \overline{1, n}$).

Параметр $n_D^{Par_{ab}}$ в формуле (8) указывает на число вхождений всех пар словоформ, являющихся словоизменениями соответственно слов *a* и (или) *b* и встречающихся в одном и том же предложении текста *D*:

$$n_D^{Par_{ab}} = \sum_{\substack{c \in Par_a,\ d \in Par_b,\\ c \neq a \text{ и/или } d \neq b\\ c, d \in \rho,\ \rho \in D}} n_D^{cd}.$$

Подобное выражение можно записать и для параметра $n_D^{Syn_{ab}}$:

$$n_D^{Syn_{ab}} = \sum_{\substack{c \in Syn_a,\ d \in Syn_b,\\ c \neq a \text{ и/или } d \neq b\\ c,d \in \rho,\ \rho \in D}} n_D^{df}\ .$$

Для параметров $N_{Cf}^{Par_{ab}}$ и $N_{Cf}^{Syn_{ab}}$ верны аналогичные выражения, отличающиеся тем, что в каждом из них индекс D заменяется на Cf.

Формулу (8) для краткого сообщения $Q \in L(G)$ запишем в виде

$$I_Q^{ab} = \frac{n_{Ct}^{ab} + n_{Ct}^{Par_{ab}} + n_{Ct}^{Syn_{ab}}}{n_{Cf}^{ab} + N_{Cf}^{Par_{ab}} + N_{Cf}^{Syn_{ab}}}, \tag{9}$$

где Ct – по-прежнему тематический корпус текстов, релевантный сообщению Q.

Информативность вербальных ассоциаций между предложениями. Рассмотрим произвольный текстовый документ или краткое сообщение T. Пусть π и ρ – произвольные предложения или словосочетания текста T, а I_T^{ab} – информативность вербально-ассоциативной связи между его словами a и b. Тогда информативность $I_T^{\pi\rho}$ вербально-ассоциативной связи этих предложений будем вычислять по аналогии с вычислением информативности предложений (см. выражение (5)) по формуле

$$I_T^{\pi\rho} = \frac{\sum\limits_{a \in \pi,\ b \in \rho} I_T^{ab}}{\sqrt{\sum\limits_{a \in \pi,\ b \in \rho} (I_T^{ab})^2}}. \tag{10}$$

В частности, формула вычисления информативности между предложением π и словом b текста T имеет вид

$$I_T^{\pi b} = \frac{\sum\limits_{a \in \pi} I_T^{ab}}{\sqrt{\sum\limits_{a \in \pi} (I_T^{ab})^2}}. \tag{11}$$

Информативность вербальных ассоциаций между субтекстами. Обозначим через Sub_1 и Sub_2 субтексты текста T. Пусть по-прежнему $I_T^{\pi\rho}$ – информативность вербально-ассоциативной связи между любыми предложениями π и ρ текста T. Тогда для вычисления информативности вербально-ассоциативной связи между субтекстами Sub_1 и Sub_2 построим аналог формулы (6):

$$I_Q^{Sub_1, Sub_2} = \frac{\sum\limits_{\pi \in Sub_1,\ \rho \in Sub_2} I_Q^{\pi\rho}}{\sqrt{\sum\limits_{\pi \in Sub_1,\ \rho \in Sub_2} (I_Q^{\pi\rho})^2}}. \tag{12}$$

Разбиение предложений текста на коммуникативные фрагменты. В отличие от устоявшейся традиции, в силу которой синтез текста рассматривается как процесс последовательной генерации морфем, лексем, синтаксических фраз и, наконец, предложений, в монографии [3] показано, что основой использования языка человеком является его языковая память. Другими словами, предложения при синтезе строятся из готовых хранящихся в памяти компонентов, названных коммуникативными фрагментами. Поэтому анализу текста будет предшествовать разбиение всех его предложений на такие фрагменты.

Формализация понятия коммуникативного фрагмента. Пусть $\pi = a_1 a_2 \ldots a_n$ – произвольное предложение (или подцепочка некоторого предложения) входного языка $L(G)$ из тематического корпуса текстов Ct. Определим формально понятие коммуникативного фрагмента:

1. Если $n = 1$, то слово a_1 цепочки π назовем коммуникативным фрагментом.

2. Если $n \geq 2$ и $I_{Ct}^{a_1 a_2} < I_{Ct}^{00}$ (I_{Ct}^{00} – пороговое значение информативности), то слово a_1 будем называть коммуникативным фрагментом. Информативность вербально-ассоциативной связи между словами a_1 и a_2 вычисляется по формуле (8) (или (9) в случае краткого сообщения).

3. Если при $n \geq 2$ выполняется последовательность неравенств $I_{Ct}^{a_1 a_2} \geq I_{Ct}^{00}$, $I_{Ct}^{(a_1 a_2) a_3} \geq I_{Ct}^{00}$, ..., $I_{Ct}^{(a_1 a_2 \ldots a_{n-1}) a_n} \geq I_{Ct}^{00}$, то цепочку $a_1 a_2 \ldots a_n$ назовем коммуникативным фрагментом. Значения информативности $I_{Ct}^{(a_1 a_2) a_3}$, ..., $I_{Ct}^{(a_1 a_2 \ldots a_{n-1}) a_n}$ вычисляются по формуле (11).

4. Пусть $2 \leq m < n$. Подцепочку $a_1 a_2 \ldots a_m$ цепочки π назовем коммуникативным фрагментом, если справедлива последовательность неравенств $I_{Ct}^{a_1 a_2} \geq I_{Ct}^{00}$, $I_{Ct}^{(a_1 a_2) a_3} \geq I_{Ct}^{00}$, ..., $I_{Ct}^{(a_1 a_2 \ldots a_{m-1}) a_m} \geq I_{Ct}^{00}$, а $I_{Ct}^{(a_1 a_2 \ldots a_m) a_{m+1}} < I_{Ct}^{00}$.

Словарь коммуникативных фрагментов. В каждой записи словаря представлен коммуникативный фрагмент CF на входном языке и фрагменты-эквиваленты CF_i на других языках:

$$Dic_{CF} = \{(CF, CF_1, CF_2, ..., CF_n) \mid CF \in L(G), CF_i \in L(G_i), i = \overline{1, n}, n \geq 1\}, \quad (13)$$

где G_i – формальные порождающие грамматики для различных языков.

Алгоритмы разбиения предложений текста на коммуникативные фрагменты. Рассмотрим произвольное предложение $\pi = a_1 a_2 \ldots$ входного языка. В соответствии с приведенным выше формальным определением процесс разбиения предложения π на коммуникативные фрагменты реализуется следующим образом.

Проверяем, является ли коммуникативным фрагментом слово a_1. Если не является, то рассматриваем словосочетания $a_1 a_2$, $a_1 a_2 a_3$ и т. д. Если некоторый коммуникативный фрагмент $a_1 a_2 \ldots a_l$ в цепочке π выделен, то этот фрагмент исключается из фразы π; последовательно всем словам оставшейся подцепочки приписываются новые индексы, начиная с 1, и снова повторяется описанная процедура.

Рассмотрим алгоритмы выявления первого коммуникативного фрагмента в цепочках вида $a_1 a_2 \ldots$.

А л г о р и т м 1. В соответствии с алгоритмом 1 ищется первый коммуникативный фрагмент в цепочке из входного текста с помощью словаря Dic_{CF}. На входе алгоритма – цепочка $\pi = a_1 a_2 \ldots a_n$, на выходе – множество CF, единственным элементом которого является коммуникативный фрагмент $a_1 a_2 \ldots a_l$ ($l \geq 1$, $l \leq n$) или $CF = \varnothing$.

1. $i := 1$, $CF := \varnothing$.

2. Искать слово a_1 в словаре Dic_{CF}. Если слово a_1 найдено, то перейти к п. 3, иначе перейти к п. 4.

3. $CF := \{a_1\}$. Конец алгоритма (выделен коммуникативный фрагмент a_1).

4. $i := i + 1$. Если $i \leq n$, то перейти к п. 5, иначе перейти к п. 7.

5. Искать словосочетание $a_1 \ldots a_i$ в словаре Dic_{CF}. Если словосочетание $a_1 \ldots a_i$ найдено, то перейти к п. 6, иначе перейти к п. 4.

6. $CF := \{a_1 \ldots a_i\}$. Конец алгоритма (выделен коммуникативный фрагмент $a_1 \ldots a_i$).

7. Конец алгоритма (использовать алгоритм 2 для выделения первого коммуникативного фрагмента).

А л г о р и т м 2. С использованием алгоритма 2 в цепочке входного текста выявляется первый коммуникативный фрагмент путем вычисления информативности вербальных ассоциаций по формуле (8) (или (9), если входной текст является кратким сообщением). На входе алгоритма – словосочетание $\pi = a_1 a_2 \ldots a_n$, на выходе – коммуникативный фрагмент $a_1 a_2 \ldots a_l$ ($l \geq 1$, $l \leq n$).

1. $CF := \varnothing$.

2. Если $n = 1$, то $CF := \{a_1\}$; конец алгоритма (выделен коммуникативный фрагмент a_1). Если $n > 1$, то перейти к п. 3.

3. Если $n \geq 2$ и $I_{Ct}^{a_1 a_2} < I_{Ct}^{00}$, то $CF := \{a_1\}$; конец алгоритма (выделен коммуникативный фрагмент a_1). Если $n \geq 2$, а $I_{Ct}^{a_1 a_2} \geq I_{Ct}^{00}$, то перейти к п. 4.

4. Если $n \geq 2$ и выполняются неравенства $I_{Ct}^{a_1a_2} \geq I_{Ct}^{00}$, $I_{Ct}^{(a_1a_2)a_3} \geq I_{Ct}^{00}$, ... , $I_{Ct}^{(a_1a_2\ldots a_{n-1})a_n} \geq I_{Ct}^{00}$, то $CF := \{a_1a_2 \ldots a_n\}$; конец алгоритма (выделен коммуникативный фрагмент $a_1a_2 \ldots a_n$). Если $n \geq 2$, а хотя бы одно из неравенств $I_{Ct}^{a_1a_2} \geq I_{Ct}^{00}$, $I_{Ct}^{(a_1a_2)a_3} \geq I_{Ct}^{00}$, ... , $I_{Ct}^{(a_1a_2\ldots a_{n-1})a_n} \geq I_{Ct}^{00}$ не выполнено, то перейти к п. 5.

5. Если $3 \leq m \leq n$ и справедливы неравенства $I_{Ct}^{a_1a_2} \geq I_{Ct}^{00}$, $I_{Ct}^{(a_1a_2)a_3} \geq I_{Ct}^{00}$, ... , $I_{Ct}^{(a_1a_2\ldots a_{m-2})a_{m-1}} \geq I_{Ct}^{00}$, а $I_{Ct}^{(a_1a_2\ldots a_{m-1})a_m} < I_{Ct}^{00}$, то $CF := \{a_1a_2 \ldots a_{m\text{-}1}\}$; конец алгоритма (выделен коммуникативный фрагмент $a_1a_2 \ldots a_{m\text{-}1}$).

А л г о р и т м 3 является обобщающим (управляющим). Согласно этому алгоритму все предложения входного текста разбиваются на коммуникативные фрагменты. При работе алгоритма 3 используются алгоритмы 1 и 2. Одновременно пополняется словарь коммуникативных фрагментов. На входе алгоритма – текстовый документ или краткое сообщение T, на выходе – кортеж T_{CF}, все предложения которого разбиты на коммуникативные фрагменты.

1. $T_{CF} := \varnothing$.

2. Выбрать очередное предложение $\pi = a_1a_2 \ldots a_n$ из текста T, перейти к п. 3. Если же предложения исчерпаны, то конец алгоритма (во всех предложениях входного текста выделены коммуникативные фрагменты).

3. Выполнить алгоритм 1. Если $CF = \varnothing$, то перейти к п. 4. Если $CF \neq \varnothing$, то перейти к п. 5.

4. Выполнить алгоритм 2. Поместить коммуникативный фрагмент из множества CF в кортеж T_{CF} и в позицию CF словаря Dic_{CF}, исключить этот фрагмент из предложения π, $CF := \varnothing$. Если все слова исключены из предложении π, то перейти к п. 2; в противном случае перейти к п. 6.

5. Поместить коммуникативный фрагмент из множества CF в кортеж T_{CF} и исключить этот фрагмент из предложения π, $CF := \varnothing$. Если все слова исключены из предложении π, то перейти к п. 2; в противном случае перейти к п. 6.

6. Перенумеровать все слова оставшейся подцепочки предложения π, начиная с 1. Перейти к п. 3.

Создание вербально-ассоциативной сети предложения. Будем различать вербально-ассоциативные сети отдельных анализируемых предложений и сети текстов.

Отношение вербально-ассоциативной связи коммуникативных фрагментов. Рассмотрим тематические корпусы текстов Ct_{ij} ($i = \overline{1, m}$, $j = \overline{1, n_i}$) и полные корпусы $Cf_i = \bigcup_{j=1}^{n_i} Ct_{ij}$. Полный корпус текстов Cf_i соответствует i-му входному языку $L(G_i)$, а тематический корпус Ct_{ij} – j-й предметной области для языка $L(G_i)$.

Обозначим через $CoCf_i$ множество всех коммуникативных фрагментов полного корпуса текстов Cf_i. Тогда отношение толерантности Θ_i (рефлексивное и симметричное бинарное отношение) на множестве $CoCf_i$ назовем отношением вербально-ассоциативной связи коммуникативных фрагментов в полном корпусе текстов Cf_i, если любая упорядоченная пара коммуникативных фрагментов (f, g) из множества $CoCf_i$ является элементом отношения Θ_i тогда и только тогда, когда фрагменты f и g из этой пары содержатся хотя бы в одном предложении корпуса Cf_i.

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

Определим на множестве $CoCf_i$ всех коммуникативных фрагментов в полном корпусе текстов Cf_i антирефлексивное бинарное отношение Λ_i, такое, что для любых фрагментов $f, g \in Ft$ соотношение $(f, g) \in \Lambda_i$ выполняется тогда и только тогда, когда в некотором тексте $T \in CoCf_i$ существует предложение π, в котором коммуникативный фрагмент f непосредственно

предшествует фрагменту g. Отношение Λ_i будем называть отношением дискурсивной сочетаемости коммуникативных фрагментов в полном корпусе текстов Cf_i.

Определение вербально-ассоциативной сети предложения. Пусть $T \in L(G)$ – текстовый документ или краткое сообщение, π – произвольное предложение текста T, а $Co\pi$ – множество всех коммуникативных фрагментов предложения π. Рассмотрим сужение Θ_π отношения Θ_i на множество $Co\pi$, т. е. $\Theta_\pi = \Theta_i \cap (Co\pi \times Co\pi)$. Отношение Θ_π назовем отношением вербально-ассоциативной связи коммуникативных фрагментов в предложении π.

Построим сужение Λ_π отношения Λ_i на множество $Co\pi$, т. е. $\Lambda_\pi = \Lambda_i \cap (Co\pi \times Co\pi)$. Отношение Λ_π назовем отношением дискурсивной сочетаемости коммуникативных фрагментов в предложении π.

Рассмотрим граф отношения Θ_π. Пометим каждую вершину f этого графа значением ее информативности I_T^f (с учетом синонимии и словоизменения), а каждое ребро (f, g) – значением информативности I_T^{fg} вербально-ассоциативной связи фрагментов f и g (также учитывая синонимию и словоизменения). Информативность I_T^f вычисляется по формуле (5), а информативность I_T^{fg} – по формуле (11). Пусть (f, g) – произвольное ребро графа. Если $(f, g) \in \Lambda_\pi$, то для всех таких пар ребро (f, g) заменим дугой, направленной от f к g. Обозначим полученный смешанный граф через Net_π.

Граф Net_π назовем вербально-ассоциативной сетью предложения π. В информационной системе граф Net_π представим в виде

$$Net_\pi = \{\langle (f, I_T^f); (g, I_T^g); (I_T^{fg}, Arc)\rangle \mid f \in \pi, g \in \pi\}, \qquad (14)$$

где $Arc = 1$, если $(f, g) \in \Lambda_\pi$; $Arc = -1$, если $(g, f) \in \Lambda_\pi$, и $Arc = 0$, если $(f, g) \notin \Lambda_\pi$ и $(g, f) \notin \Lambda_\pi$.

Алгоритм формирования вербально-ассоциативной сети предложения. Пусть по-прежнему $T \in L(G)$ – текстовый документ или краткое сообщение, а $\pi = a_1 a_2 \ldots a_n$ – произвольное предложение текста T. Вербально-ассоциативную сеть предложения π представим в виде множества кортежей Net_π.

А л г о р и т м 4. При работе алгоритма 4 используются алгоритмы 1–3. Согласно алгоритму 4 в предложении выявляются коммуникативные фрагменты, для каждого фрагмента вычисляется его информативность, а также определяется сила вербально-ассоциативной связи между фрагментами. На входе алгоритма – предложение $\pi = a_1 a_2 \ldots a_n$, на выходе – вербально-ассоциативная сеть предложения π в виде множества кортежей Net_π.

1. Выполнить алгоритм 3. Представить предложение π в виде цепочки коммуникативных фрагментов: $\pi = f_1 f_2 \ldots f_n$. (Обозначим через $Co\pi = \{h_1, h_2, \ldots, h_m\}$ множество всех коммуникативных фрагментов предложения π.)

2. $Net_\pi := \varnothing$, $i := 0$, $j := 1$.

3. $i := i + 1$. Если $i = m$, то перейти к п. 7, в противном случае перейти к п. 4.

4. $j := j + 1$. Если $j := m + 1$, то $j := 1$, перейти к п. 3. В противном случае перейти к п. 5.

5. Вычислить информативность $I_T^{h_i}$ и информативность $I_T^{h_j}$ коммуникативных фрагментов h_i и h_j соответственно по формуле (5). Вычислить информативность $I_T^{h_i h_j}$ вербально-ассоциативной связи между фрагментами h_i и h_j по формуле (10). Если $(h_i, h_j) \in \Lambda_T$, то $Arc := 1$; если $(h_j, h_i) \in \Lambda_T$, то $Arc := -1$; если же $(h_i, h_j) \notin \Lambda_T$ и $(h_j, h_i) \notin \Lambda_T$, то $Arc := 0$. Перейти к п. 6.

6. Сформировать очередной кортеж $\langle (f, I_T^f); (g, I_T^g)); (I_T^{fg}, Arc)\rangle$. Поместить h_i в позицию f, $I_T^{h_i}$ – в позицию I_T^f, h_j – в позицию g, $I_T^{h_j}$ – в позицию I_T^g, значение Arc – в позицию Arc. Поместить сформированный кортеж $\langle (h_i, I_T^{h_i}); (h_j, I_T^{h_j})); (I_T^{h_i h_j}, Arc)\rangle$ в множество Net_π. Перейти к п. 4.

7. Конец алгоритма. Вербально-ассоциативная сеть Net_π сформирована.

Создание вербально-ассоциативной сети текста. При формировании кортежа вербально-ассоциативных сетей всех предложений входного текста алгоритм 4 используется для каждого предложения.

Рассмотрим процесс создания вербально-ассоциативной сети текста в целом.

Определение вербально-ассоциативной сети текста. Обозначим через CoT множество всех коммуникативных фрагментов текста T. Рассмотрим сужение Θ_T отношения Θ_i на множество CoT, т. е. $\Theta_T = \Theta_i \cap (CoT \times CoT)$. Отношение Θ_T назовем отношением вербально-ассоциативной связи коммуникативных фрагментов в тексте T.

Построим также сужение Λ_T отношения Λ_i на множество CoT, т. е. $\Lambda_T = \Lambda_i \cap (CoT \times CoT)$. Отношение Λ_T назовем отношением дискурсивной сочетаемости коммуникативных фрагментов в тексте T.

Рассмотрим граф отношения Θ_T. Пометим каждую его вершину значением информативности соответствующего фрагмента, а каждое ребро (f, g) – значением информативности вербально-ассоциативной связи между инцидентными ему вершинами f и g. Если $(f, g) \in \Lambda_T$, то ребро (f, g) заменяем дугой, направленной от f к g. Обозначим полученный смешанный граф через Net_T. Граф Net_T будем называть вербально-ассоциативной сетью текста T.

При практической реализации информационной системы граф Net_T целесообразно представить в виде

$$Net_T = \{\langle (f, I_T^f); (g, I_T^g); (I_T^{fg}, Arc)\rangle \mid f \in T, g \in T\}, \qquad (15)$$

где $Arc = 1$, если $(f, g) \in \Lambda_T$; $Arc = -1$, если $(g, f) \in \Lambda_T$, и $Arc = 0$, если $(f, g) \notin \Lambda_T$ и $(g, f) \notin \Lambda_T$.

Алгоритм формирования вербально-ассоциативной сети текста.

А л г о р и т м 5. В соответствии с алгоритмом 5 в тексте T выявляются коммуникативные фрагменты, для каждого фрагмента вычисляется его информативность, а также определяется сила вербально-ассоциативной связи между фрагментами. На входе алгоритма – текст T, на выходе – вербально-ассоциативная сеть текста T в виде множества Net_T.

1. Выполнить алгоритм 3 для каждого предложения текста T. Представить все предложения текста T в виде цепочек коммуникативных фрагментов. (Обозначим через $CoT = \{t_1, t_2, \ldots, t_l\}$ множество всех коммуникативных фрагментов текста T.)

2. $Net_T := \varnothing$, $i := 0$, $j := 1$.

3. $i := i + 1$. Если $i = l$, то перейти к п. 7, в противном случае перейти к п. 4.

4. $j := j + 1$. Если $j := l + 1$, то $j := 1$, перейти к п. 3. В противном случае перейти к п. 5.

5. Вычислить информативность $I_T^{t_i}$ и информативность $I_T^{t_j}$ коммуникативных фрагментов t_i и t_j соответственно по формуле (5). Вычислить информативность $I_T^{t_i t_j}$ вербально-ассоциативной связи между фрагментами t_i и t_j по формуле (10). Если $(t_i, t_j) \in \Lambda_T$, то $Arc := 1$; если $(t_j, t_i) \in \Lambda_T$, то $Arc := -1$; если же $(t_i, t_j) \notin \Lambda_T$ и $(t_j, t_i) \notin \Lambda_T$, то $Arc := 0$. Перейти к п. 6.

6. Сформировать очередной кортеж $\langle (f, I_T^f); (g, I_T^g); (I_T^{fg}, Arc)\rangle$. Поместить t_i в позицию f, $I_T^{t_i}$ – в позицию I_T^f, t_j – в позицию g, $I_T^{t_j}$ – в позицию I_T^g, значение Arc – в позицию Arc. Поместить сформированный кортеж $\langle (t_i, I_T^{t_i}); (t_j, I_T^{t_j}); (I_T^{t_i t_j}, Arc)\rangle$ в множество Net_T. Перейти к п. 4.

7. Конец алгоритма. Вербально-ассоциативная сеть Net_T сформирована.

Заключение. Предложенная в статье модель процессов анализа текстовых документов и кратких сообщений может быть использована при индексировании, поиске и реферировании текстовой информации в Интернете, корпоративных сетях и в локальных базах данных. При соответствующем подборе тематики и языка представления корпусов текстов возможны поиск и реферирование документов на различных входных языках. Реализация этой функции сводится к формированию корпусов текстов и созданию словарей базы знаний (без коррекции программного обеспечения системы). При наличии персональных тематических корпусов текстов обеспечивается адаптация процессов поиска и реферирования к информационным потребностям соответствующих индивидуальных и корпоративных пользователей.

Список использованных источников

1. Липницкий, С. Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С. Ф Липницкий // Информатика. – 2011. – № 4(32). – С. 21–28.

2. Вейзе, А. А. Чтение, реферирование и аннотирование иностранного текста / А. А. Вейзе. – М.: Высшая школа, 1985. – 127 с.

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

Информация об авторе

Липницкий Станислав Феликсович – доктор технических наук, главный научный сотрудник, Объединенный институт проблем информатики НАН Беларуси (ул. Сурганова, 6, Минск, Республика Беларусь). E-mail: lipn@newman.bas-net.by

Information about the author

Stanislav F. Lipnitsky – D. Sc. (Engineering), Chief Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganova Str., Minsk, Republic of Belarus). E-mail: lipn@newman.bas-net.by