ПРИКЛАДНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.912

Л.В. Степура

АВТОМАТИЧЕСКОЕ РЕФЕРИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ МОДЕЛИРОВАНИЯ СИТУАТИВНЫХ СВЯЗЕЙ МЕЖДУ ПОНЯТИЯМИ ПРЕДМЕТНОЙ ОБЛАСТИ

Рассматривается модель процессов реферирования текстовой информации на основе формализации информационных языков средствами специальной порождающей грамматики, а также понятий информативности слов и ситуативных связей между ними. В рамках реализации модели предлагается метод синтеза связных рефератов текстовых документов путем выявления информативных предложений, построения их контекста и генерации кортежей синтаксических деревьев.

Введение

Стремительный рост объемов данных, в том числе текстовой информации, ведет к спросу на системы интеллектуального анализа текста и к бурному развитию такого инструментария. Значимое место в списке систем обработки текстовой информации занимают программы автоматического реферирования и аннотирования, позволяющие многократно упрощать и ускорять процедуры обработки больших массивов текстов.

Реферат играет важную роль в системе сведений о текстовом документе: он дополняет его библиографическое описание и содержит ряд характеристик, которые дают первичное представление о содержании публикации. Известны три основных подхода к автоматическому реферированию текстовых документов: создание рефератов путем выделения и последующей «склейки» информативных фрагментов текста с использованием статистических оценок их информативности; синтез рефератов на основе лингвистической обработки документов; гибридные методы, в которых используются как статистические характеристики текста, так и результаты лингвистического анализа.

В статье предлагается метод автоматического реферирования текстовой информации на основе моделирования главных структурных и функциональных компонентов информационной системы. Рассматривается модель представления знаний о предметной области в виде ситуативной сети, т. е. графа, вершинами которого являются информативные слова, а ребрами – ситуативные связи между ними. Моделируются процессы синтаксического анализа текстовых документов, разбиения их на предложения и синтеза связных рефератов. Для целей реферирования используется база знаний, включающая обобщенное представление выходного текста в виде упорядоченного множества синтаксических шаблонов предложений, а также словари информативных словоформ и устойчивых словосочетаний.

В существующих системах поиска и аналитической обработки текстовых документов используются главным образом технологии, ориентированные на исследование структуры и статистических характеристик самих документов без привлечения дополнительной информации [1, 2]. В данной статье эти задачи решаются с использованием тематических корпусов текстов и сформированных на их основе лингвистических словарей [3, 4].

1. Информационные языки системы реферирования

При реферировании текстовой информации используются три информационных языка – входной, внутренний и выходной. Для их определения рассмотрим формальную порождающую грамматику $G = \langle V, N, I, R \rangle$, где V – непустое множество терминальных элементов (слов), $N = \{I, '\}$ – множество нетерминальных, I – начальный символ, а R – схема грамматики, т. е.

множество правил вывода вида $\alpha \to \beta$ (α и β – различные непустые цепочки в словаре $V \cup N$). Схема R грамматики G формируется по следующим правилам:

– 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). В связи с этим грамматику G будем называть *штрихграмматикой*.

1.1. Входной язык

Пусть $V_{вх}$ – словарь некоторого естественного языка, который будем называть *входным словарем*, а его элементы – *словами* входного языка. По аналогии со схемой R штрихграмматики G построим совокупность правил вывода $R_{вх}$. Тогда язык $L(G_{вх})$, порождаемый штрихграмматикой $G_{вх} = \langle V_{вх}, N, I, R_{вх} \rangle$, будем называть *входным языком*.

1.2. Внутренний язык

Обозначим через W некоторое непустое подмножество лексем входного словаря $V_{вх}$ (под лексемой в лингвистике понимают «слово в совокупности всех его словоизменительных форм»). Зафиксируем также некоторое непустое подмножество Si элементов словаря $V_{вх}$ (назовем их *семантическими признаками*). Рассмотрим множество $V_{вн}$ цепочек вида ap языка $L(G_{вх})$, где $a \in W$, $p \in Si$. Множество $V_{вн}$ будем называть *внутренним словарем*, а его элементы – *понятиями*.

Пусть имеется штрихграмматика $G_{вн} = \langle V_{вн}, N, I, R_{вн} \rangle$ и язык $L(G_{вн})$, порождаемый этой грамматикой. Язык $L(G_{вн})$ будем называть *внутренним языком* системы, а словарь $V_{вн}$ – *внутренним словарем*. Схема $R_{вн}$ грамматики $G_{вн}$ аналогична схеме $R_{вх}$ грамматики $G_{вх}$.

1.3. Выходной язык

Пусть $V_{вых}$ – некоторое непустое множество терминальных элементов (назовем его *выходным словарем*). Тогда наряду с входным $L(G_{вх})$ и внутренним $L(G_{вн})$ языками информационной системы будем рассматривать *выходной язык* $L(G_{вых})$ как язык, порождаемый штрихграмматикой $G_{вых} = \langle V_{вых}, N, I, R_{вых} \rangle$. Схема $R_{вых}$ этой грамматики формируется по аналогии со схемой $R_{вх}$ грамматики $G_{вх}$.

В конкретной реализации системы реферирования выходной язык может совпадать с входным. Возможны также случаи, когда входных и/или выходных языков несколько.

При рассмотрении положений, касающихся всех трех рассмотренных языков, индексы «вх», «вн» и «вых» будем опускать.

2. Ситуативные связи между понятиями предметной области

С целью интеллектуализации системы реферирования будем использовать модель знаний о предметной области в виде ситуативной сети, т. е. графа, вершинами которого являются информативные слова и словосочетания предметной области, а ребрами – ситуативные связи между ними.

2.1. Корпусы текстов

В ко́рпусной лингвистике под корпусом текстов понимают совокупность текстов, накопленных и размеченных по определенным принципам в зависимости от назначения. В случае отсутствия разметки эти совокупности иногда называют корпусами текстов первого порядка. Будем различать тематические и полные корпусы текстов.

Любое непустое подмножество T входного языка $L(G_{вх})$ будем называть *текстом*, если на этом подмножестве определена редукция $\prec^r = \prec \setminus \prec^2$ линейного порядка $\prec$ (транзитивного и антисимметричного бинарного отношения на множестве T, которое связанно на T, т. е. для любых $a, b \in T$ или $a \prec b$, или $b \prec a$, или $a = b$). Цепочки текста T назовем *предложениями*.

Пусть имеется некоторое непустое множество текстов (совокупность текстов по конкретной тематике). Сформируем текст Ct, объединив все множества предложений каждого из этих текстов, и назовем его тематическим корпусом текстов. Поскольку в информационной системе

представлено, как правило, несколько таких корпусов, будем обозначать их *Ctj* (*j* – номер корпуса). Объединение $Cf_i = \bigcup_{j=1}^{n_i} Ct_j$ всех тематических корпусов назовем полным корпусом текстов (*i* – номер полного корпуса текстов).

2.2. Ситуативное отношение

Рассмотрим тематические корпусы текстов Ct_j и полные корпусы $Cf_i = \bigcup_{j=1}^{n_i} Ct_j$ ($i = \overline{1, m}$; $j = \overline{1, n_i}$). Полный корпус текстов Cf_i соответствует *i*-му входному языку (например, английскому), а тематический корпус Ct_j – *j*-й предметной области для *i*-го языка (например, предметной области Mathematics, представленной текстами на английском языке).

Обозначим через W_i множество всех слов полного корпуса текстов Cf_i. Тогда отношение толерантности Θ_i (рефлексивное и симметричное бинарное отношение) на множестве W_i назовем *ситуативным отношением* в полном корпусе текстов Cf_i, если любая упорядоченная пара слов (*a*, *b*) из множества W_i является элементом отношения Θ_i тогда и только тогда, когда слова *a* и *b* из этой пары содержатся хотя бы в одном предложении корпуса Cf_i.

Обозначим через W_j множество всех слов тематического корпуса текстов Ct_j. Рассмотрим сужение Θ_j отношения Θ_i на множество W_j, т. е. $\Theta_j = \Theta_i \cap (W_j \times W_j)$. Отношение Θ_j назовем *ситуативным отношением* в тематическом корпусе текстов Ct_j.

2.3. Информативность слов и ситуативных связей между словами

Информативность $I^a_{Ct_j}$ слова *a* из тематического корпуса текстов Ct_j – это вероятность того, что слово *a* имеется в данном корпусе при условии, что оно содержится в полном корпусе текстов. При достаточно больших объемах тематического и полного корпусов текстов формула для вычисления информативности слова имеет вид [3]

$$I^a_{Ct_j} = n^a_{Ct_j} / n^a_{Cf_i}, \qquad (1)$$

где $n^a_{Ct_j}$, $n^a_{Cf_i}$ – абсолютные частоты встречаемости слова *a* (с учетом синонимии и словоизменения) в тематическом Ct_j и полном Cf_i корпусах текстов.

Понятие информативности ситуативной связи между словами определим по аналогии с понятием информативности слова.

Пусть имеются слова *a, b* входного языка. Рассмотрим следующую совокупность событий (в теоретико-вероятностном смысле):

$S^{ab}_{Ct_j}$ – извлечение случайным образом слов *a* и *b* из одного и того же предложения тематического корпуса текстов Ct_j;

$S^{ab}_{Cf_i}$ – извлечение случайным образом слов *a* и *b* из одного и того же предложения полного корпуса текстов Cf_i;

H_{Ct_j} – появление тематического корпуса текстов Ct_j.

Обозначим через $P(S^{ab}_{Ct_j} / S^{ab}_{Cf_i})$ вероятность того, что слова *a* и *b* извлечены из одного и того же предложения множества C_{ij} при условии, что они уже извлечены из одного и того же предложения полного корпуса текстов Cf_i. Эта условная вероятность вычисляется следующим образом:

$$P(S^{ab}_{Ct_j} / S^{ab}_{Cf_i}) = \frac{P(S^{ab}_{Ct_j} \cdot S^{ab}_{Cf_i})}{P(S^{ab}_{Cf_i})} = \frac{P(S^{ab}_{Ct_j}) \cdot P(S^{ab}_{Cf_i} / S^{ab}_{Ct_j})}{P(S^{ab}_{Cf_i})}.$$

Вероятность $P(S^{ab}_{Ct_j} / S^{ab}_{Cf_i})$ будем называть *информативностью ситуативной связи между словами a и b* в тематическом корпусе текстов Ct_j (или предметной области, определяемой корпусом Ct_j).

По аналогии с формулой (1) информативность $I^{ab}_{Ct_j}$ можно представить в виде

$$I^{ab}_{Ct_j} = n^{ab}_{Ct_j} / n^{ab}_{Cf_i}, \tag{2}$$

где $n^{ab}_{Ct_j}$, $n^{ab}_{Cf_i}$ – абсолютные частоты совместной встречаемости слов a и b (с учетом синонимии и словоизменения) в одном и том же предложении тематического Ct_j и полного Cf_i корпусов текстов.

Информативность I_π предложения (словосочетания) $\pi \in L(G_{\text{вх}})$ определяется по формуле [5]

$$I_\pi = \frac{I_a + I_b + \ldots}{\sqrt{I_a{}^2 + I_b{}^2 + \ldots}}, \tag{3}$$

где $I_a, I_b, \ldots$ – показатели информативности слов $a, b, \ldots$ предложения или словосочетания π.

2.4. Информативность ситуативных связей между предложениями и фрагментами текста

Пусть π и ρ – произвольные предложения или словосочетания некоторого текста T, а I^{ab}_T – информативность ситуативной связи между его словами. Тогда информативность $I^{\pi\rho}_T$ ситуативной связи между этими предложениями будем вычислять по аналогии с вычислением информативности предложений по формуле

$$I^{\pi\rho}_T = \frac{\sum\limits_{a\in\pi,\, b\in\rho} I^{ab}_T}{\sqrt{\sum\limits_{a\in\pi,\, b\in\rho} (I^{ab}_T)^2}}. \tag{4}$$

Обозначим через Sub_1 и Sub_2 фрагменты, или субтексты, текста T. Пусть по-прежнему $I^{\pi\rho}_T$ – информативность ситуативной связи между любыми предложениями π и ρ текста T. Тогда для вычисления информативности ситуативной связи между фрагментами Sub_1 и Sub_2 построим аналог предыдущей формулы:

$$I^{Sub_1, Sub_2}_T = \frac{\sum\limits_{\pi\in Sub_1,\, \rho\in Sub_2} I^{\pi\rho}_T}{\sqrt{\sum\limits_{\pi\in Sub_1,\, \rho\in Sub_2} (I^{\pi\rho}_T)^2}}. \tag{5}$$

2.5. Определение ситуативной сети предметной области

Пусть S_j – граф ситуативного отношения Θ_j в корпусе Ct_j. Пометим каждую вершину a графа S_j значением информативности $I^{a}_{Ct_j}$ этого слова (с учетом синонимии и словоизменения), а каждое ребро (a, b) – значением информативности $I^{ab}_{Ct_j}$ ситуативной связи слов a и b (также учитывая синонимию и словоизменения). Обозначим полученный граф через Net_j.

Граф Net_j назовем *ситуативной сетью предметной области*, определяемой тематическим корпусом текстов Ct_j.

3. Синтаксический анализ реферируемого текста

Предлагается метод синтаксического анализа и разбиения текста на предложения, основанный на моделировании процесса распознавания синтагм (пар синтаксически связанных слов) средствами рассмотренной выше штрихграмматики, которая обеспечивает универсальность метода для различных проективных естественных языков. Конечная цель синтаксического анализа текста при автоматическом реферировании – построение для него кортежа синтаксических деревьев и разбиение таким образом его на предложения.

3.1. Основное свойство маргинальных синтагм

Содержательно под маргинальной будем понимать синтагму, определяющий член которой не имеет в реферируемом тексте синтаксически зависимых членов, т. е. не является определяемым ни для каких других слов текста. Определим формально понятие маргинальной синтагмы.

Пусть $\alpha a \beta b \gamma$ (или $\alpha b \beta a \gamma$) – произвольная цепочка языка $L(G_{\text{вх}})$, где $\alpha, \beta, \gamma \in V^*$ (V^*– множество всех цепочек в словаре $V_{\text{вх}}$ грамматики $G_{\text{вх}}$), ab (или ba) – синтагма этой цепочки с определяемым членом a и определяющим b.

Синтагму ab (или ba) назовем *маргинальной синтагмой* цепочки $\alpha a \beta b \gamma$ (или $\alpha b \beta a \gamma$), если для любого вхождения слова c ($c \neq b$) в $\alpha a \beta b \gamma$ (или в $\alpha b \beta a \gamma$) цепочки bc и cb не являются синтагмами цепочки $\alpha a \beta b \gamma$ (или $\alpha b \beta a \gamma$). Слово b синтагм ab и ba будем называть *маргинальным словом* синтагм αb и ba.

Пусть ab – синтагма некоторой цепочки языка $L(G)$. Тогда будем говорить, что синтаксическая связь направлена от слова a к слову b, если $(a, b) \in \Omega_\pi$. Если же $(b, a) \in \Omega_\pi$, то у такой связи противоположное направление. Для краткости направление синтаксической связи между словами будем обозначать стрелкой с началом над определяемым членом синтагмы и концом над определяющим (например, $\alpha \overrightarrow{a \beta b} \gamma$, $\alpha \overleftarrow{a \beta b} \gamma$).

Следующее утверждение является теоретической основой для построения алгоритма синтаксического анализа текста при автоматическом реферировании.

У т в е р ж д е н и е 1. *Если* $\rho \in L(G_{\text{вх}})$, *а* ab (*или* ba) – *маргинальная синтагма цепочки* ρ, *причем в схеме* $R_{\text{вх}}$ *грамматики* $G_{\text{вх}}$ *имеется правило вывода* $a' \to a'b'$ (*или* $a' \to b'a'$), *то цепочка* σ, *полученная из* ρ *удалением определяющего члена* b *синтагмы* ab (*или* ba), *является цепочкой языка* $L(G_{\text{вх}})$.

Д о к а з а т е л ь с т в о . Поскольку ab – синтагма цепочки ρ, то для цепочки ρ существует вывод $W = (I, \alpha, \beta, \ldots, \gamma, \mu a' \nu, \mu a'b' \nu, \ldots, \mu ab \nu, \ldots, \rho)$ в грамматике G, где $\alpha, \beta, \gamma, \mu, \nu \in V^*$. Так как ab – маргинальная синтагма цепочки ρ, то в силу определения маргинальной синтагмы для любого слова c цепочки ρ цепочка bc не является синтагмой, т. е. при выводе цепочки ρ не используются правила типа $b' \to b'c'$, а цепочка $\mu a'b' \nu$ в выводе W получена из цепочки $\mu a' \nu$ применением правила вывода $a' \to a'b'$. Если цепочку $\mu a'b' \nu$ исключить из вывода W, то получим вывод цепочки σ из начального символа I. Аналогично рассматривается случай, когда синтагмой цепочки ρ является цепочка ba. ■

3.2. Поиск маргинальных синтагм в реферируемом тексте

Согласно утверждению 1 процесс синтаксического анализа текста и разбиения его на предложения может быть реализован следующим образом:

– в исходном тексте ищутся маргинальные синтагмы;

– строятся синтаксические деревья найденных синтагм;

– исключаются определяющие члены найденных синтагм. В результате их исключения в полученном тексте могут появиться новые маргинальные синтагмы;

– далее процесс повторяется до тех пор, когда поиск новых маргинальных синтагм окажется безрезультатным.

Докажем ряд утверждений, которые будут способствовать нахождению в тексте маргинальных синтагм.

Пусть по-прежнему $\pi = a_1 a_2 \ldots a_n$. Исследуем условия существования маргинальных синтагм цепочки π в синтагматических структурах следующих четырех типов:

1) $\overleftarrow{a_1 a_2}$, $\overrightarrow{a_{n-1} a_n}$;

2) $\overrightarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$, $\overleftarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$ $(i = \overline{1, n-2})$; $\overrightarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$, $\overleftarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$ $(i = \overline{1, j-1},\ j = \overline{2, n-2})$;

3) $\overrightarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$ $(i = \overline{1, n-2})$; $\overrightarrow{\overrightarrow{a_i a_{i+1}} a_{i+2} \ldots a_j}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}} a_{i+2} \ldots a_j}$ $(i = \overline{1, j-3},\ j = \overline{4, n})$;

4) $\overrightarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}}}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}}}$ $(i = \overline{1, n-3})$; $\overrightarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}} \ldots a_k}$, $\overleftarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}} \ldots a_k}$ $(i = \overline{1, j-1},\ j = \overline{2, k-4},\ k = \overline{6, n})$.

Утверждение 2. *Синтагмы* $\overleftarrow{a_1 a_2}$, $\overrightarrow{a_{n-1} a_n}$ *являются маргинальными синтагмами цепочки* π.

Доказательство. Пусть от противного цепочка $a_{n-1} a_n$ не является маргинальной синтагмой, т. е. в цепочке π имеется слово c, которое служит определяемым членом синтагмы ca_n. Тогда в схеме R грамматики G должны существовать правила, обеспечивающие вывод цепочки $c' \ldots a'_{n-1} a'_n$ из цепочки $a'_{n-1} a'_n$, что противоречит определению штрихграмматики G. Аналогично доказывается, что синтагма $\overleftarrow{a_1 a_2}$ является маргинальной. ■

По аналогии с доказательством утверждения 2 доказываются утверждения 3–5.

Утверждение 3. *Синтагматические структуры* $\overrightarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$, $\overleftarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$ *цепочки* π *содержат маргинальную синтагму* $\overleftarrow{a_{i+1} a_{i+2}}$.

Синтагматические структуры $\overrightarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$, $\overleftarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$ *цепочки* π *содержат маргинальную синтагму* $\overrightarrow{a_j a_{j+1}}$.

Утверждение 4. *Синтагматические структуры* $\overrightarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$ *цепочки* π *содержат маргинальную синтагму* $\overrightarrow{a_i a_{i+1}}$.

Синтагматические структуры $\overrightarrow{\overrightarrow{a_i a_{i+1}} a_{i+2} \ldots a_j}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}} a_{i+2} \ldots a_j}$ *цепочки* π *содержат маргинальную синтагму* $\overrightarrow{a_i a_{i+1}}$.

Утверждение 5. *Синтагматические структуры* $\overrightarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}}}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}}}$ *цепочки* π *содержат маргинальные синтагмы* $\overrightarrow{a_i a_{i+1}}$, $\overleftarrow{a_{i+2} a_{i+3}}$.

Синтагматические структуры $\overrightarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}} \ldots a_k}$, $\overleftarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}} \ldots a_k}$ *цепочки* π *содержат маргинальные синтагмы* $\overrightarrow{a_j a_{j+1}}$, $\overleftarrow{a_{j+2} a_{j+3}}$.

Утверждения 3–5 позволяют найти маргинальные синтагмы в анализируемом тексте. Процесс их поиска реализуется следующим образом. В цепочке $\pi = a_1 a_2 \ldots a_n$ ищутся маргинальные синтагмы типа $\overleftarrow{a_1 a_2}$ и $\overrightarrow{a_{n-1} a_n}$ и исключаются из π их определяющие члены. Процесс поиска таких синтагм и исключения определяющих членов повторяется до тех пор, пока синтагмы указанного типа будут присутствовать в цепочке π. Далее аналогичная процедура повторяется для синтагматических структур типа $\overrightarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$, $\overleftarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$ и $\overrightarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$, $\overleftarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$, затем для структур $\overrightarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$, $\overleftarrow{a_i \ldots \overrightarrow{a_j a_{j+1}} a_{j+2}}$ и т. д.

Пример 1. Рассмотрим процесс синтаксического анализа текста и разбиения его на предложения на примере цепочки «Данный проект реализуется поэтапно первый этап запланирован на этот год окончание работ в следующем году».

На первом шаге анализа в исходной цепочке выявляются неразделенные синтагмы:

Данный проект $\overleftarrow{\text{реализуется поэтапно}}$ *первый этап* $\overleftarrow{\text{запланирован на этот}}$ *год*

$\overleftarrow{\text{окончание работ в следующем}}$ *году*

На втором шаге из полученной цепочки исключаются определяющие члены маргинальных синтагм:

проект реализуется этап запланирован год окончание работ в году

Процесс последовательного выявления маргинальных синтагм и исключения из них определяющих членов продолжается аналогичным образом на третьем, четвертом, пятом и шестом шагах синтаксического анализа:

проект реализуется этап запланирован год окончание в году

проект реализуется этап запланирован год окончание в году

проект реализуется этап запланирован год окончание в

проект реализуется этап запланирован год окончание

В результате получим три синтаксических дерева анализируемого текста (рис. 1).

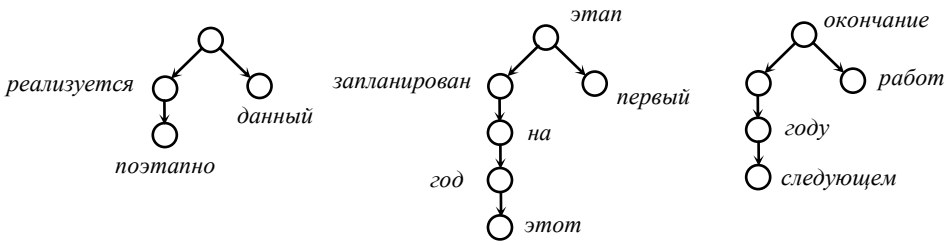

Рис. 1. Синтаксические деревья текста

Полученным синтаксическим деревьям соответствует разбиение исходного текста на три предложения: данный проект реализуется поэтапно, первый этап запланирован на этот год, окончание работ в следующем году.

4. Выявление контекста информативных предложений

Понятие контекста предложения тесным образом связано с понятием сверхфразового единства, т. е. кортежа предложений единой тематической направленности. При построении в тексте сверхфразовых единств вначале будем использовать ситуативные связи между словами текста, а затем между его предложениями. Ситуативные связи между предложениями текста представим в виде графа ситуативных связей. Введем предварительно понятие графа информативности текста.

4.1. Граф информативности текста

Пусть имеется текст (т. е. кортеж предложений) $T = \langle \pi_1, \pi_2, \ldots \rangle$. Вычислим информативность всех предложений текста T и исключим из T неинформативные предложения, т. е. все предложения π, информативность I_π которых меньше некоторого I_0. В результате получим кортеж предложений $T_{\text{инф}} = \langle \pi_{i_1}, \pi_{i_2}, \ldots \rangle$, который будем называть *маршрутом информативности* текста T. Соединив последовательно вершины графа текста G_T (т. е. графа редукции линейного порядка на множестве всех предложений текста T), соответствующие информативным предложениям, получим орграф $G_{\text{инф.}}$, который будем называть *графом информативности* текста T (рис. 2). Вершины и дуги графа текста, не вошедшие в состав графа информативности $G_{\text{инф}}$, изображены пунктирными линиями.

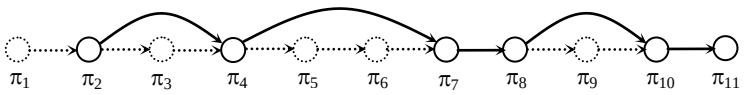

Рис. 2. Пример графа информативности текста

4.2. Граф ситуативных связей между предложениями текста

Пусть T^+ – множество всех информативных, а T^- – всех неинформативных предложений текста T ($T = T^+ \cup T^-$). Определим на паре множеств T^+, T^- симметричное отношение Ξ, такое, что для любых предложений $\pi \in T^+$ и $\rho \in T^-$ $(\pi, \rho) \in \Xi$ тогда и только тогда, когда информативность $I_{\pi\rho}$ ситуативной связи между предложениями π и ρ не меньше некоторого значения. Граф отношения Ξ назовем *графом ситуативных связей* между предложениями текста T (рис. 3). Информативные предложения изображены на рис. 3 сплошными линиями, а неинформативные – пунктирными. Пунктирными стрелками представлены дуги графа текста, а скобками объединены возможные кандидаты в сверхфразовые единства.

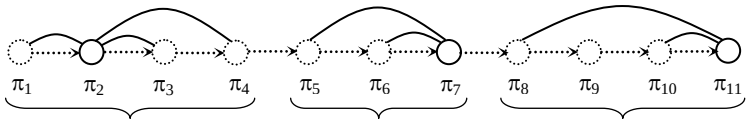

Рис. 3. Пример графа ситуативных связей между предложениями текста

Процесс разбиения текста на сверхфразовые единства осуществляется в два этапа. На первом этапе в тексте выявляются информативные предложения, т. е. строится маршрут информативности $T_{\text{инф.}}$. На втором этапе выявляются ситуативные связи между предложениями текста, на основе которых формируются сверхфразовые единства.

5. Синтез связного реферата

Для целей синтеза связного реферата будем использовать специальную базу знаний, включающую обобщенное представление выходного текста в виде упорядоченного множества синтаксических шаблонов предложений, а также словари информативных словоформ и устойчивых словосочетаний. Задачу синтеза будем решать в два этапа: на первом этапе сформируем кортеж синтаксических деревьев, используя синтаксический шаблон предметной области, а на втором синтезируем предложения текста.

5.1. Синтаксические шаблоны

Пусть имеется текст $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$ в виде кортежа предложений $\pi_1, \pi_2, \ldots, \pi_m$. Обозначим через D_π синтаксическое дерево любого предложения π из текста T. Ордерево Dr_π, полученное из синтаксического дерева D_π заменой всех его поддеревьев, которые являются синтаксическими деревьями прагматически полных синтагматических структур (ПП-структур), слотами («пустыми» вершинами), будем называть *синтаксическим шаблоном предложения* π (рис. 4). Под ПП-структурой понимается информативная в некотором тематическом разделе предметной области (т. е. хотя бы в одном тематическом корпусе текстов) синтагматическая структура, выражаемая устойчивым словосочетанием (например, «информационные технологии», «входной язык», «радиоаппаратура»).

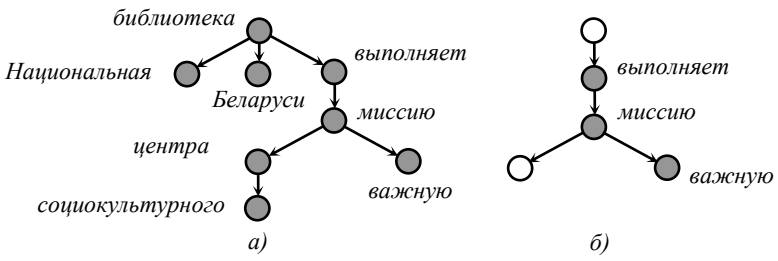

Рис. 4. Анализ предложения «Национальная библиотека Беларуси выполняет важную миссию социокультурного центра»: *а)* синтаксическое дерево; *б)* синтаксический шаблон

При синтезе синтаксического дерева предложения слоты заменяются синтаксическими деревьями синтагматических структур из графа ситуативных связей между предложениями текста.

Обозначим через D_i ($i = \overline{1, m}$) синтаксический шаблон предложения π_i. Тогда кортеж $Sh = \langle D_1, D_2, \ldots, D_m \rangle$ синтаксических шаблонов всех предложений текста $T = \langle \pi_1, \pi_2, \ldots, \pi_m \rangle$ назовем *синтаксическим шаблоном* этого текста.

5.2. Отношение дискурсивной сочетаемости

Синтаксический шаблон текста строится на основе отношения дискурсивной сочетаемости Δ, которое определим следующим образом.

Пусть имеется множество $\{Sh_i \mid i = \overline{1, n}\}$ синтаксических шаблонов некоторых текстов. Рассмотрим объединение множеств синтаксических шаблонов $Sh = \bigcup_{i=1}^{n} Sh_i$. Определим на множестве Sh антирефлексивное бинарное отношение Δ, такое, что для любых синтаксических шаблонов $D_r, D_s \in Sh$ некоторых предложений π_r и π_s соотношение $(D_r, D_s) \in \Delta$ справедливо тогда и только тогда, когда существует синтаксический шаблон текста Sh_j ($1 \le j \le n$), элементами которого являются синтаксические шаблоны D_r и D_s предложений π_r и π_s соответственно; в синтаксическом шаблоне текста Sh_j синтаксический шаблон предложения π_r непосредственно предшествует синтаксическому шаблону предложения π_s. Отношение Δ назовем *отношением дискурсивной сочетаемости* синтаксических шаблонов предложений.

Определим на множестве всех пар отношения Δ (т. е. на множестве Δ) отношение строгого порядка $\prec$ (антирефлексивное и транзитивное бинарное отношение) следующим образом: будем считать, что для любых пар синтаксических шаблонов предложений (D_i, D_j) и (D_k, D_l) отношения Δ соотношение $(D_i, D_j) \prec (D_k, D_l)$ справедливо тогда и только тогда, когда $D_j = D_k$, т. е. D_j и D_k – один и тот же синтаксический шаблон.

Множество Sh с определенным на нем отношением дискурсивной сочетаемости Δ и строгим порядком $\prec$, заданным на множестве Δ, назовем *синтаксическим шаблоном предметной области*.

Используя строгий порядок $\prec$, синтаксический шаблон реферата можно построить в два этапа: сначала в виде ориентированного маршрута в графе, вершинами которого являются упорядоченные пары синтаксических шаблонов предложений, а дуги соответствуют отношению $\prec$, затем в виде орцепи, где вершины (синтаксические шаблоны предложений) соединены дугами, определяющими порядок использования этих шаблонов при синтезе текста. Выбор каждого очередного элемента множества Δ реализуется путем сравнения синтаксического шаблона и синтаксического дерева реферата.

5.3. Синтез кортежа синтаксических деревьев

Процедура построения кортежа синтаксических деревьев синтезируемого текста реализуется в два этапа.

На первом этапе ищется минимальный (в смысле строгого порядка $\prec$) элемент множества Δ, т. е. пара синтаксических шаблонов предложений (D_1, D_2). Шаблон D_1 должен удовлетворять следующему условию: в графе ситуативных связей между предложениями текста должны существовать ПП-структуры для заполнения слотов шаблона D_1. Далее в множестве Δ ищутся пары синтаксических шаблонов предложений (D_2, D_3), (D_3, D_4), … . Шаблоны предложений $D_2, D_3, \ldots$ также должны удовлетворять упомянутому выше условию. После заполнения слотов всех найденных шаблонов (кроме поименованных) ПП-структурами из графа ситуативных связей между предложениями текста и ситуативной сети получим требуемый кортеж синтаксических деревьев с незаполненными поименованными слотами.

На втором этапе заполняются поименованные слоты сформированного кортежа синтаксических деревьев.

5.4. Упорядоченное синтаксическое дерево

Определим предварительно понятия «расстояние между словами предложения», «отношение семантической близости» и «упорядочивающие отображения».

Расстоянием $R(a_i, a_j)$ между словами a_i и a_j цепочки $a_1a_2\ldots a_i\ldots a_j\ldots a_n$ назовем модуль разности j и i, т. е. $R(a_i, a_j) = |j - i|$.

Пусть a – произвольное слово некоторого предложения ($a \in V$, V – словарь), а L – множество синтаксически корректных синтагматических структур из полного корпуса текстов. (Факт синтаксической корректности устанавливает эксперт-лингвист.) Определим на множестве $\Omega \cap (\{a\} \times V)$ бинарное отношение $\geq_a$, являющееся объединением эквивалентности $=_a$ и строгого порядка $>_a$, следующим образом. Будем считать, что для любых слов $b, c \in V$, таких, что $(a, b) \in \Omega$ и $(a, c) \in \Omega$, выполняется соотношение $(a, b) >_a (a, c)$, если в множестве L существует синтагматическая структура из слов a, b и c, такая, что $R(a, b) > R(a, c)$, и нет структуры из этих же слов, где выполняется неравенство противоположного знака. Считаем также, что $(a, b) =_a (a, c)$, если существует синтаксически корректная синтагматическая структура, где $R(a, b) = R(a, c)$, или найдутся две таких структуры, в которых соответственно $R(a, b) < R(a, c)$ и $R(a, b) > R(a, c)$. Отношение $\geq_a$ назовем *отношением семантической близости*. Если $(a, b) >_a (a, c)$, то будем говорить, что слова a и b *семантически связаны сильнее,* чем слова a и c. Если же $(a, b) =_a (a, c)$, то скажем, что слова b и c *семантически равнозначны относительно слова* a.

Для всех троек слов a, b, c типа рассмотренных выше построим совокупность отображений $\Phi_a : \Omega \cap (\{a\} \times V) \to \{1, 2, \ldots\}$, таких, что $\Phi_a(a, b) > \Phi_a(a, c)$, если $(a, b) >_a (a, c)$, а если $(a, b) =_a (a, c)$, то $\Phi_a(a, b) = \Phi_a(a, c)$. Такие отображения Φ_a назовем *упорядочивающими.*

На практике в качестве совокупности L используется полный корпус текстов, а образы всех дуг, исходящих из вершины a синтаксического дерева предложения, при упорядочивающем отображении Φ_a можно рассматривать как числовые метки на этих дугах.

Синтаксическое дерево любого предложения назовем *упорядоченным*, если все его дуги помечены натуральными числами, являющимися образами этих дуг при отображениях Φ_a.

Процесс построения упорядоченного синтаксического дерева реализуется следующим образом.

Ищется произвольная висячая вершина синтаксического дерева b_1, являющаяся конечной вершиной орцепи максимальной длины, и смежная ей вершина a.

Ищутся все дуги $(a, b_1), (a, b_2), \ldots$, исходящие из вершины a.

Выявляются натуральные числа, которые являются образами всех найденных дуг при отображениях Φ_a, и помечаются ими эти дуги.

Условно исключаются из синтаксического дерева все дуги, исходящие из вершины a, и их конечные вершины. Если в синтаксическом дереве после такого исключения имеются дуги, то процесс начинается сначала, иначе алгоритм заканчивает работу.

Пример 2. Рассмотрим процедуру синтеза предложения, упорядоченное синтаксическое дерево которого изображено на рис. 5.

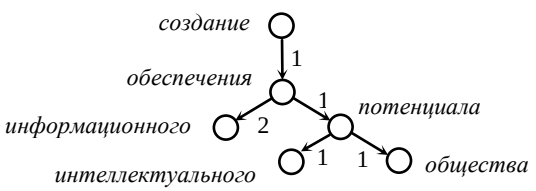

Рис. 5. Пример упорядоченного синтаксического дерева

Процедура синтеза осуществляется поэтапно (рис. 6):
1) строится синтагма «создание обеспечения»;
2) строится синтагма «информационного обеспечения» и «обеспечения потенциала»;
3) строится синтагма «потенциала общества»;
4) синтезируется синтагма «интеллектуального потенциала».

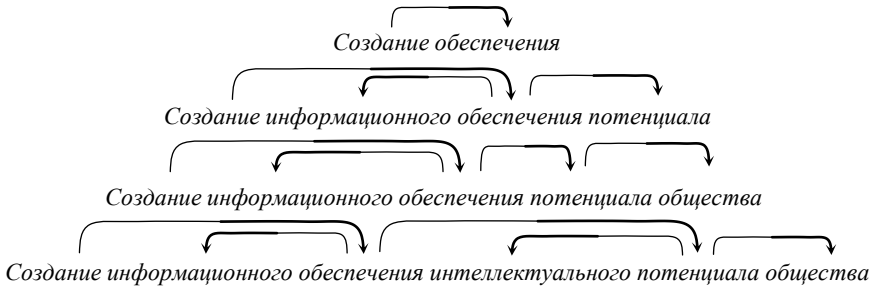

Рис. 6. Этапы синтеза предложения

В результате получаем предложение «Создание информационного обеспечения интеллектуального потенциала общества».

Заключение

Предложенная модель представления знаний о предметной области в виде ситуативной сети отличается универсальностью, т. е. независимостью от конкретного естественного языка. Адаптация системы реферирования к входному языку реализуется путем наполнения соответствующей базы знаний без изменения программного обеспечения.

Разработанный в статье метод автоматического реферирования текстовой информации обеспечивает качественное формирование рефератов с применением синтаксического анализа и разбиением текста на предложения, с выявлением контекста информативных предложений и синтеза связного реферата.

Список литературы

1. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы / С. Тактаев // [Электронный ресурс]. – Режим доступа : http:// www.searchengines.ru/articles/004603.html. – Дата доступа : 30.01.2013.

2. Тарасов, С.Д. Современные методы автоматического реферирования / С.Д. Тарасов // Научно-технические ведомости СПбГПУ. – СПб. : СПбГПУ, 2010. – № 6. – С. 59–74.

3. Кравцов, А.А. Индексирование и реферирование текста на основе ситуативно-синтагматической сети / А.А. Кравцов, С.Ф. Липницкий, Л.В. Степура // Искусственный интеллект. Интеллектуальные системы (ИИ-2009) : материалы X Междунар. науч.-техн. конф. – Таганрог : Изд-во ТТИ ЮФУ, 2009. – С. 277–279.

4. Липницкий, С.Ф. Модели знаний о предметной области для решения задач поиска и обработки текстовой информации / С.Ф. Липницкий // Информатика. – 2007. – № 2. – С. 25–34.

5. Липницкий, С.Ф. Алгоритмы создания гипертекста на основе ситуативно-синтагматической сети / С.Ф Липницкий, Л.В. Степура // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2010. – № 3. – С. 90–95.

Поступила 12.02.13

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

L.V. Stepura

AUTOMATIC ABSTRACTING OF TEXTUAL INFORMATION BASED ON SITUATIONAL LINKS MODELING OF APPLICATION DOMAIN CONCEPTS

A model of abstracting textual information, based on the formalization of information languages by a special generative grammar and the concepts of word self-descriptiveness and contextual links between words, is considered. A method of abstracts synthesis for textual documents by identifying informative sentences, constructing their context and generating chains of syntax trees is suggested.