ИНФОРМАТИКА

2007 **апрель-июнь** **№ 2**

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 681.322

С.Ф. Липницкий

МОДЕЛИ ЗНАНИЙ О ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ РЕШЕНИЯ ЗАДАЧ ПОИСКА И ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ

Предлагаются модели знаний о предметной области в системе интеллектуальной обработки текстовой информации. Модели могут быть использованы для решения задач документального и фактографического поиска, классификации текстовых документов, автоматического реферирования и аннотирования текстов, машинного перевода.

Введение

Эффективность функционирования информационных систем существенным образом зависит от их интеллектуальности, т. е. способности работать не только с данными, но и знаниями об объектах и явлениях предметной области [1–4].

В данной статье предложены две модели знаний для решения задач поиска и обработки текстовых документов. Первая модель ориентирована на использование для целей анализа /синтеза текстов только одного языка (без языка-посредника). Во второй модели предполагается, что тексты поступают в информационную систему на входном языке, обрабатываются на внутреннем и интерпретируются на выходном.

Предложенные модели знаний отличаются универсальностью, т. е. независимостью от конкретного естественного языка. Адаптация системы к входному языку реализуется путем создания соответствующей базы знаний (комплекса словарей) без изменения программного обеспечения.

1. Языки системы поиска и обработки текстовой информации

В системе поиска и обработки текстовой информации будем различать три языка: входной, внутренний и выходной. Для их определения в работе [5] введена формальная порождающая грамматика $G = \langle V, N, I, R\rangle$, 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 грамматика G названа *штрих-грамматикой*.

1.1. Входной язык

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

Пример фрагмента входного языка. Пусть в грамматике $G_{\text{вх.}} = \langle V_{\text{вх.}}, N, I, R_{\text{вх.}}\rangle$:

$V_{\text{вх.}}$ = {быстрыми, интеллектуальные, информационные, развиваются, темпами, технологии};

$N = \{I, '\}$;

$R_{\text{вх.}}$ = {$I \to$ быстрыми′, $I \to$ интеллектуальные′, $I \to$ информационные′, $I \to$ развиваются′, $I \to$ темпами′, $I \to$ технологии′, быстрыми′ $\to$ быстрыми, интеллектуальные′ $\to$ интеллектуальные, информационные′ $\to$ информацион-

ные, развиваются′ → развиваются, темпами′ → темпами, технологии′ → технологии, технологии′ → технологии′ развиваются′, технологии′ → информационные′ технологии′, технологии′ → интеллектуальные′ технологии′, развиваются′ → развиваются′ темпами′, темпами′ → быстрыми′ темпами′}.

Грамматика $G_{вх.}$ порождает, в частности, следующие цепочки: интеллектуальные информационные технологии; темпами; технологии развиваются; информационные технологии развиваются; интеллектуальные информационные технологии развиваются быстрыми темпами.

1.2. Внутренний язык

Обозначим через *Wo* некоторое непустое подмножество лексем входного словаря $V_{вх.}$. (Под лексемой в лингвистике понимают «слово в совокупности всех его словоизменительных форм» [6, с. 251].) 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 $a \in Wo$, $p \in Si$. Множество $V_{вн.}$ будем называть *внутренним словарем*, а его элементы – *понятиями*.

Примеры понятий: технологии ⟨что⟩, развиваться ⟨что делают⟩, темпы ⟨как⟩, банк ⟨где⟩, лук-оружие ⟨что⟩.

Пусть имеется штрих-грамматика $G_{вн.} = \langle V_{вн.}, N, I, R_{вн.} \rangle$ 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 штрих-грамматикой $G_{вых.} = \langle V_{вых.}, N, I, R_{вых.} \rangle$. Схема $R_{вых.}$ этой грамматики формируется по аналогии со схемой $R_{вх.}$ грамматики $G_{вх.}$.

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

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

2. Моделирование ситуативных связей между понятиями предметной области

Известно, что понимание текста человеком связано со знанием языка, с одной стороны, и распознаванием ситуативного контекста – с другой. При отсутствии ситуативных знаний восприятие текста возможно только на лингвистическом уровне. В связи с этим промоделируем знания о предметной области в виде ситуативных связей между информативными понятиями предметной области (определение информативности приведено в работе [7]).

2.1. Текст. Корпуса текстов

Текстом будем называть любое непустое подмножество цепочек языка $L(G)$, если на этом подмножестве определено отношение линейного порядка. Цепочки текста назовем *предложениями*.

Под корпусом текстов будем понимать текст, полученный в результате объединения («склейки») различных текстов. Будем различать тематические корпуса и полный корпус текстов.

Пусть имеется некоторое непустое множество текстов входного языка $L(G_{вх.})$ (набор текстов по конкретной тематике). Сформируем текст *Th*, 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 Th_i (i – номер корпуса). Объединение $Fu = \bigcup_{i=1}^{n} Th_i$ всех тематических корпусов назовем *полным корпусом* текстов.

На практике тематический корпус представляет собой набор текстов по некоторой конкретной тематике, а полный корпус – это объединение всех тематических корпусов. Если в

полном корпусе текстов представлен только один тематический, то полный корпус дополняется корпусом текстов с общеупотребительной лексикой, т. е. полный корпус текстов должен содержать, как минимум, два тематических.

2.2. Прагматически полные синтагматические структуры

Прагматически полная синтагматическая структура (ПП-структура) – это информативная в некотором тематическом разделе предметной области (т. е. хотя бы в одном тематическом корпусе текстов) синтагматическая структура, выражаемая устойчивым словосочетанием.

Примеры ПП-структур: информационная технология; множество; гидрофизический институт; информационная система; синтаксический анализ предложения.

Формализуем понятие ПП-структуры. Рассмотрим некоторое предложение $\pi = a_1 a_2 ... a_{i-1} a_i a_{i+1} ... a_n$ входного языка $L(G_{вх.})$, где a_1, a_2,..., a_{i-1}, a_i, a_{i+1},..., a_n – вхождения слов в это предложение. Пусть a_i – информативное слово предложения π. Последовательно присоединяя к слову a_i слева и справа другие слова предложения π, сформируем множество Ch_0 всех его двухсловных, трехсловных (и т. д.) подцепочек, синтаксическими графами которых являются ордеревья (синтаксические деревья [5]). Слово a_i также включим в множество Ch_0. Поставим в соответствие каждой из выбранных подцепочек α вероятность $P(\alpha)$ ее появления в полном корпусе текстов *Fu*. Выберем пороговое значение p_0 этой вероятности и удалим из множества Ch_0 все цепочки, вероятность появления которых в корпусе *Fu* меньше p_0. Обозначим через Ch_1 множество всех оставшихся в Ch_0 однословных цепочек, через Ch_2 – двухсловных цепочек и т. д. Обозначим, наконец, через Ch_j ($j \geq 1$) непустое множество из совокупности $\{Ch_1, Ch_2, \ldots\}$ с наибольшим индексом и введем следующее определение.

Все подцепочки цепочки π из множества Ch_j будем называть *прагматически полными синтагматическими структурами*.

2.3. Ситуативно-синтагматическая сеть

Построим модель базы знаний информационной системы в виде графа, вершинами которого являются ПП-структуры, а ребрами – ситуативные связи между ними, которые формализуем в виде ситуативного отношения на множестве синтагматических структур.

Обозначим через *Str* множество всех ПП-структур полного корпуса текстов *Fu*. Тогда отношение толерантности Θ (рефлексивное и симметричное бинарное отношение) на множестве *Str* назовем *ситуативным отношением* в полном корпусе текстов *Fu*, если любая упорядоченная пара ПП-структур (μ, ν) из множества *Str* является элементом отношения Θ тогда и только тогда, когда вероятность совместной встречаемости ПП-структур μ и ν в корпусе текстов *Fu* не меньше некоторого порогового значения (*уровня* ситуативной связи).

Под совместной встречаемостью двух ПП-структур здесь понимается наличие этих структур (или их синонимов) в одном и том же предложении корпуса *Fu*. Граф $S_{сит.}$ ситуативного отношения будем называть *ситуативно-синтагматической сетью*. Ситуативно-синтагматическая сеть как база знаний о предметной области ориентирована главным образом на решение задачи семантического сжатия информации (рис. 1).

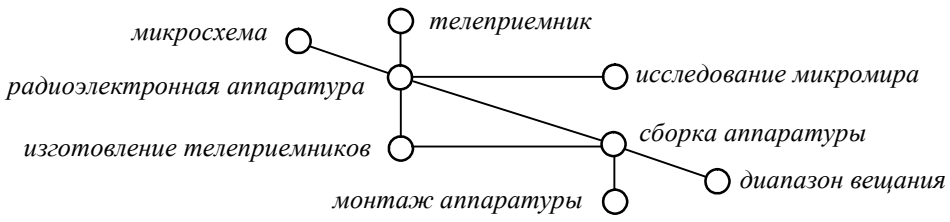

Рис. 1. Фрагмент ситуативно-синтагматической сети

При реализации информационной системы ситуативное отношение Θ представляется ситуативно-синтагматическим словарем (табл. 1). Каждая запись ситуативно-синтагматического словаря

содержит пару ПП-структур и ее частоту в предложениях полного корпуса текстов. Пары ПП-структур типа «радиоэлектронная аппаратура – изготовление телеприемников» и «изготовление телеприемников – радиоэлектронная аппаратура» считаем двумя вхождениями одной и той же пары ПП-структур, т. е. частота в третьем столбце табл. 1 в таких случаях увеличивается на два.

Таблица 1

Фрагмент ситуативно-синтагматического словаря

ПП-структура	**ПП-структура**	**Частота в *Fu***
•••		
радиоэлектронная аппаратура	*исследование микромира*	02312
сборка аппаратуры	*диапазон вещания*	01561
•••		

2.4. Маршрут и граф информативности. Семантический след текста

Пусть имеется текст (т. е. кортеж предложений) *Te*. Вычислим информативность синтагматических структур всех предложений текста *Te*. Исключим из *Te* все неинформативные предложения, т. е. предложения, не содержащие информативные структуры. В результате получим кортеж предложений (в порядке их следования в *Te*) $Te_{\text{инф.}} = \langle \pi_1, \pi_2, \ldots, \pi_n \rangle$. Кортеж $Te_{\text{инф.}}$ будем называть *маршрутом информативности* текста *Te*.

Построим орграф $Gr_{\text{инф.}}$, считая все предложения маршрута информативности $Te_{\text{инф.}}$ его вершинами. Всякую пару вершин π_i, π_j ($i < j$, $1 \le i \le n-1$, $2 \le j \le n$) соединим дугой (π_i, π_j) тогда и только тогда, когда в ситуативно-синтагматической сети $S_{\text{сит.}}$ существует хотя бы одна пара вершин (подцепочек предложений π_i и π_j соответственно), соединенных ребром, которое указывает на существование ситуативной связи между этими подцепочками.

Орграф $Gr_{\text{инф.}}$, на множестве вершин которого определен линейный порядок, соответствующий порядку предложений в маршруте информативности $Te_{\text{инф.}}$, будем называть *графом информативности* текста *Te*.

Маршрут информативности $Te_{\text{инф.}}$ является основой для семантического сжатия текста *Te* (т. е. для построения его реферата, аннотации). Для регулирования объема маршрута информативности и выявления в нем монотематических фрагментов построим семантический след текста, понятие которого определим следующим образом.

Семантическим следом Tr текста *Te* будем называть подграф графа информативности $Gr_{\text{инф.}}$, вершинами которого являются все вершины орграфа $Gr_{\text{инф.}}$ с числом дуг, инцидентных им, не меньше некоторого n_0 (рис. 2).

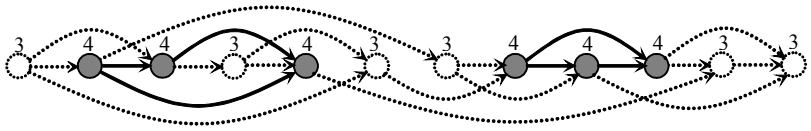

Рис. 2. Пример семантического следа текста в графе информативности

На рис. 2 каждая вершина графа информативности $Gr_{\text{инф.}}$ помечена числом, обозначающим количество инцидентных ему дуг (в данном случае $n_0 = 4$). Вершины и дуги орграфа $Gr_{\text{инф.}}$, не вошедшие в состав семантического следа *Tr*, изображены пунктирными линиями. Связные подграфы семантического следа соответствуют двум монотематическим фрагментам текста.

3. Моделирование семантических структур

Ситуативно-синтагматическую сеть $S_{\text{сит.}}$ как базу знаний о предметной области целесообразно использовать при прямом анализе и синтезе текстов, т. е. без использования внутреннего языка. Для реализации варианта анализа/синтеза текстовых документов по схеме «входной язык – внутренний язык – выходной язык» построим модель базы знаний в виде сети семантических структур.

3.1. Соотнесенное синтаксическое дерево предложения

Синтаксические связи между словами предложения языка $L(G)$ изображаются в виде синтаксического дерева, вершинами которого являются вхождения слов в предложение, а дуги соответствуют синтаксическим связям между ними [5] (рис. 3).

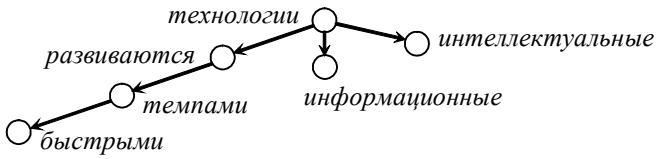

Рис. 3. Пример синтаксического дерева предложения «Интеллектуальные информационные технологии развиваются быстрыми темпами»

Соотнесенное неориентированное дерево, очевидно, существующее для синтаксического дерева каждого предложения $\pi \in L(G)$, будем называть *соотнесенным* синтаксическим деревом предложения π (рис. 4).

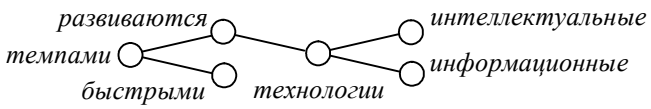

Рис. 4. Пример соотнесенного синтаксического дерева предложения

3.2. Семантическое дерево предложения внутреннего языка

Если предложение π входного языка $L(G_{\text{вх.}})$ синонимично некоторому предложению ρ внутреннего языка $L(G_{\text{вн.}})$, т. е. $(\pi, \rho) \in \Lambda$ [5], то соотнесенное синтаксическое дерево предложения ρ будем называть *семантическим деревом* предложения π (рис. 5).

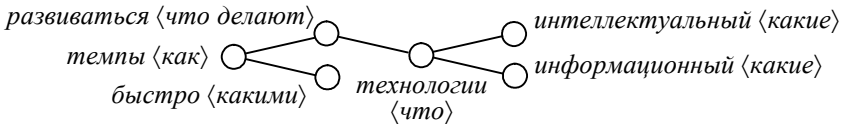

Рис. 5. Пример семантического дерева предложения

3.3. Синтаксический базис входного языка

Синтаксические деревья языка $L(G_{\text{вх.}})$ являются основой для построения семантического дерева входного предложения путем замены некоторых поддеревьев его синтаксического дерева семантическими деревьями синтагматических структур языка $L(G_{\text{вн.}})$, синонимичных соответствующим этим поддеревьям предложениям языка $L(G_{\text{вх.}})$. Эти поддеревья будем рассматривать как синтаксический базис, понятие которого введем следующим образом.

Множество $B_{\text{вх.}}$ синтаксических деревьев предложений языка $L(G_{\text{вх.}})$ назовем *синтаксическим базисом* входного языка, если синтаксическое дерево любого предложения $\alpha \in L(G_{\text{вх.}})$ может быть получено как объединение некоторого подмножества $B^{\alpha}_{\text{âõ.}}$ ордеревьев базиса $B_{\text{вх.}}$. Множество ордеревьев $B^{\alpha}_{\text{âõ.}}$ будем называть синтаксическим базисом предложения α (рис. 6).

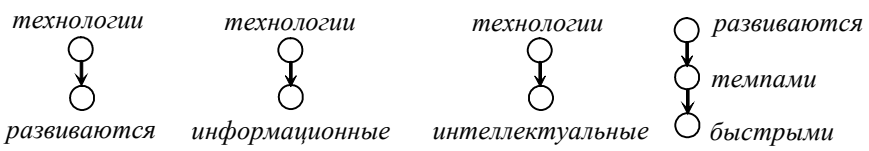

Рис. 6. Пример синтаксического базиса предложения «Интеллектуальные информационные технологии развиваются быстрыми темпами»

3.4. Семантический базис внутреннего языка

Семантическое дерево входного предложения α можно получить заменой всех ордеревьев синтаксического базиса $B^{\alpha}_{âõ.}$ семантическими деревьями некоторых синонимичных предложений внутреннего языка, используя следующее понятие семантического базиса языка $L(G_{вн.})$.

Обозначим через D некоторое непустое множество семантических деревьев предложений языка $L(G_{вн.})$. (На практике это преимущественно цепи длины 1.) Множество $B_{вн.} \in 2^D$ (2^D – множество всех подмножеств множества D) назовем *семантическим базисом* внутреннего языка, если существует сюръективное отображение $\Xi : B_{вх.} \to B_{вн.}$, такое, что для любого синтаксического дерева $a \in B_{вх.}$ предложения α и для любого семантического дерева $b \in \Xi(\alpha)$ предложения β выполняется соотношение $(\alpha, \beta) \in \Lambda$, т. е. цепочки α и β являются синонимами. Образ $\Xi(B^{\alpha}_{âõ.})$ синтаксического базиса любого входного предложения α будем называть *семантическим базисом* этого предложения и обозначать его через $B^{\alpha}_{âí.}$.

Двуязычный словарь «Входной–внутренний». Синтаксический и семантический базисы используются при анализе входных предложений, т. е. при их переводе с входного языка $L(G_{вх.})$ на внутренний язык $L(G_{вн.})$. Для реализации процессов анализа в информационной системе отображение $\Xi : B_{вх.} \to B_{вн.}$ можно представить в виде двуязычного словаря «Входной–внутренний» (табл. 2). Стрелками в табл. 2 изображены дуги синтаксического дерева, а знаком тире – ребра семантического дерева.

Таблица 2

Фрагмент двуязычного словаря «Входной–внутренний»

Синтаксическое дерево	**Семантическое дерево**
•••	
лежал → в → банке	*лежать ⟨что делал⟩ – банка ⟨где⟩*
	лежать ⟨что делал⟩ – банк ⟨где⟩
лук → лежал	*лук-растение ⟨что⟩ – лежать ⟨что делал⟩*
	лук-оружие ⟨что⟩ – лежать ⟨что делал⟩
•••	

3.5. Синтаксическое покрытие выходного языка

Задача синтеза выходных предложений на языке $L(G_{вых.})$ является обратной задаче семантического анализа, т. е. перевода предложений с языка $L(G_{вх.})$ на язык $L(G_{вн.})$. При интерпретации предложений языка $L(G_{вн.})$ на языке $L(G_{вых.})$ используется словарь, каждая статья которого содержит семантическое дерево некоторого фрагмента внутреннего предложения и синонимичное этому фрагменту синтаксическое дерево синтагматической структуры выходного языка. Для моделирования процесса синтеза введем понятие синтаксического покрытия внутреннего языка.

Пусть $L(G_{вых.})/\Lambda$ – фактор-множество множества $L(G_{вых.})$ по эквивалентности (отношению синонимии) Λ [7]. Поставим в соответствие каждому смежному классу из фактор-множества $L(G_{вых.})/\Lambda$, который содержит синонимичные синтагматические структуры, множество D синтаксических деревьев всех синтагматических структур этого класса. Обозначим совокупность всех таких множеств синтаксических деревьев через D.

Совокупность всех множеств синтаксических деревьев типа D будем называть *синтаксическим покрытием* выходного языка $L(G_{вых.})$, если существует сюръективное отображение $\Sigma : B_{вх.} \to D$, такое, что для любого семантического дерева $a \in B_{вх.}$ цепочки $\alpha \in L(G_{вн.})$ и для любого синтаксического дерева $b \in \Sigma(a)$ синтагматической структуры β выполняется соотношение $(\alpha, \beta) \in \Lambda$. Образ $D_\alpha = \Sigma\ (B^{\alpha}_{âõ.})$ семантического базиса $B^{\alpha}_{âõ.}$ любого предложения $\alpha \in L(G_{вн.})$ назовем *синтаксическим покрытием предложения* α.

При практической реализации данной математической модели в информационной системе сюръективное отображение $\Sigma : B_{вх.} \to D$ может быть представлено упомянутым выше двуязычным словарем «Внутренний–выходной» (табл. 3).

Таблица 3

Фрагмент двуязычного словаря «Внутренний–выходной»

Семантическое дерево	**Синтаксическое дерево**
•••	
лежать ⟨что делал⟩ – банк ⟨где⟩	*лежал → в → банке*
	находился → в → банке
	положен → на → хранение → в → банк
лежать ⟨что делал⟩ – банка ⟨где⟩	*лежал → в → банке*
	хранился → в → банке
лук-оружие ⟨что⟩ – лежать ⟨что делал⟩	*лук → лежал*
	лук → находился
	лук → положен → на → хранение
лук-растение ⟨что⟩ – лежать ⟨что делал⟩	*лук → лежал*
	лук → хранился
•••	

3.6. Сеть семантических структур

Для реализации варианта анализа/синтеза текстовых документов по схеме «входной язык – внутренний язык – выходной язык» построим модель базы знаний в виде сети семантических структур.

3.6.1. Семантический элемент. Пусть $ap \in V_{\text{вн.}}$ – произвольное информативное хотя бы в одном тематическом корпусе текстов понятие внутреннего языка. (Понятие считаем информативным, если его семантическое дерево является образом синтаксического дерева некоторой информативной синтагматической структуры языка $L(G_{\text{вх.}})$ при отображении $\Sigma : B_{\text{вх.}} \to D$.) Зафиксируем некоторое непустое подмножество слов Va входного словаря $V_{\text{вх.}}$; назовем эти слова *семантическими переменными*. Рассмотрим множество цепочек $b_i r_i \in V_{\text{вн.}} \cup Va$ ($i = \overline{1, n}$), где b_i – семантическая переменная ($b_i \in Va$) или лексема ($b_i \in Wo$), а r_i – семантический признак ($r_i \in Si$).

Семантическое дерево El_{ap} цепочки $\alpha = apb_1r_1b_2r_2\ldots b_nr_n$, такой, что для всех $i = \overline{1, n}$ выполняются соотношения $(ap, b_ir_i) \in \Omega_{L(G)} \cup \Omega^{-1}_{L(G)}$, будем называть *семантическим элементом* понятия ap (определение отношения $\Omega_{L(G)}$ приведено в работе [5]). Если при некотором значении i ($1 \le i \le n$) $b_i \in Va$, цепочку b_ir_i назовем *слотом* семантического элемента El_{ap} (рис. 7).

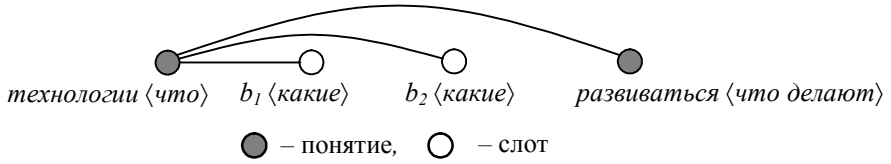

Рис. 7. Пример семантического элемента

Множество всех семантических элементов целесообразно хранить в специальном словаре (табл. 4).

Таблица 4

Фрагмент словаря семантических элементов

Понятие	**Понятие или слот**
•••	
развиваться ⟨что делают⟩	*быстро ⟨как⟩*
	b_2 ⟨с какой целью⟩
	b_3 ⟨в каком направлении⟩
технологии ⟨что⟩	*информационный ⟨какие⟩*
	интеллектуальный ⟨какие⟩
	развиваться ⟨что делают⟩
•••	

3.6.2. Семантическая сцена. Пусть G_0, G_1 – произвольные графы, множествами вершин которого являются любые непустые подмножества множества $V_{\text{вн.}} \cup Va$. Если хотя бы одна вершина *br* графа G_0 является слотом, то будем говорить, что этот слот *заполнен понятием* $cr \in V_{\text{вн.}}$ при условии, что в графе G_0 вершина *br* заменена вершиной *cr*. Если одной из вершин графа G_1 является понятие *cr*, то скажем, что слот *br* графа G_0 *заполнен графом* G_1; при этом слот *br* графа G_0 должен быть заполнен понятием *cr* и полученный в результате граф объединен (в смысле операции объединения графов) с графом G_1. Используя эти договоренности о «заполнении слотов», определим понятие семантической сцены.

Семантической сценой (рис. 8) будем называть граф, полученный:

– из семантического элемента любого понятия в результате заполнения хотя бы одного его слота элементом семантического базиса $B_{\text{вн.}}$;

– из семантического элемента любого понятия или семантической сцены в результате заполнения хотя бы одного их слота элементом семантического базиса $B_{\text{вн.}}$ или семантической сценой.

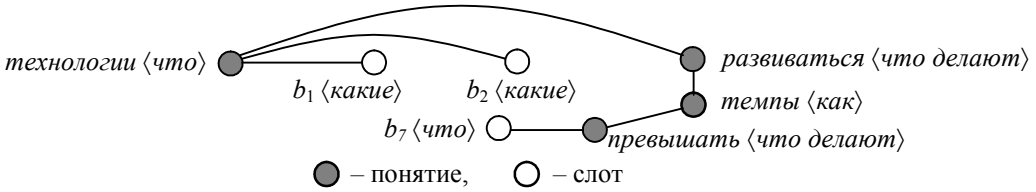

Рис. 8. Пример семантической сцены

В программной реализации информационной системы совокупность семантических сцен хранится в виде словаря, аналогичного словарю семантических элементов (см. табл. 4). В первом столбце словаря семантических сцен представлены семантический элемент или семантическая сцена, а во втором – элемент или сцена, заполняющие слоты графа из первого столбца.

3.6.3. Семантический эпизод. Определим понятие семантического эпизода как семантического элемента или семантической сцены, всем или некоторым понятиям которых парадигматически подчинены совокупности других семантических элементов или сцен.

Семантическим эпизодом (рис. 9) назовем семантический элемент или семантическую сцену, в которых имеется слот *br*, такой, что выполняется соотношение $(br, \beta_i) \in \Delta$ [7], где β_i ($i = \overline{1, l}$; $l \geq 1$) – понятие семантического элемента или семантической сцены. Эти последние семантический элемент или семантическую сцену будем называть *семантическими ожиданиями* слота *br*.

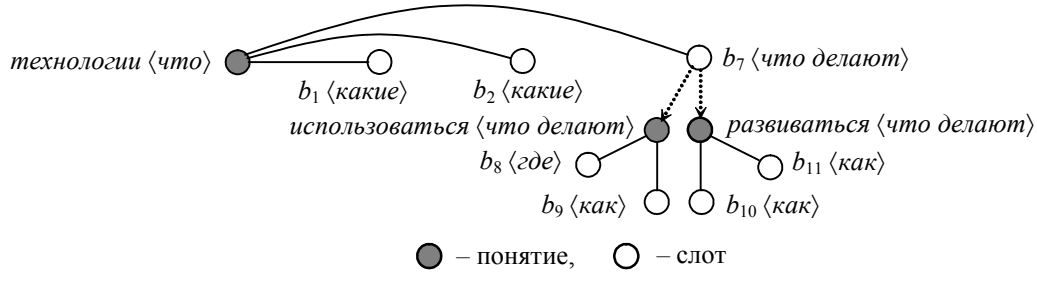

Рис. 9. Пример семантического эпизода (пунктирной стрелкой изображено парадигматическое подчинение Δ)

В информационной системе семантические эпизоды можно хранить аналогично хранению семантических сцен. Парадигматически подчиненные семантические элементы или сцены хранятся в специальном парадигматическом словаре. В его первом столбце представлены семантические элементы и сцены со слотами, а во втором – их семантические ожидания.

3.6.4. Семантический сценарий. Определим, наконец, понятие семантического сценария как совокупности семантических элементов, сцен и эпизодов, моделирующей динамику предметной области.

Обозначим через E некоторое непустое множество семантических элементов, сцен и эпизодов (*семантических структур*). Определим на E отношение строгого порядка $\prec_E$ (транзитивное и антирефлексивное бинарное отношение). *Семантическим сценарием* назовем строго упорядоченное множество $\langle E, \prec_E \rangle$ семантических элементов, сцен и эпизодов.

3.6.5. Определение сети семантических структур. Рассмотрим объединение всех зафиксированных в базе знаний информационной системы семантических элементов и семантических сцен, а также всех семантических элементов и сцен, входящих в состав семантических эпизодов. Дополним это объединение всеми дугами орграфа отношения $\prec_E$ и дугами, соответствующими отношению Δ парадигматического подчинения на множестве всех понятий этих семантических элементов и сцен, и обозначим полученный смешанный граф через $S_{\text{сем.}}$. Смешанный граф $S_{\text{сем.}}$ назовем *сетью семантических структур*.

Заключение

База знаний информационной системы, основанная на ситуативно-синтагматической сети, может быть использована при решении следующих задач:

– *автоматическое индексирование текстовой информации*. Выявление в тексте информативных ПП-структур реализуется путем сопоставления их частотных характеристик в полном корпусе текстов и в индексируемом документе (или в релевантном ему тематическом корпусе текстов в случае небольшого объема индексируемого текста). При этом возможно устранение лексической омонимии на основе использования ситуативных связей между ПП-структурами;

– *классификация текстовых документов как с использованием заранее заданных классов* (*категоризация*)*, так и при их отсутствии* (*кластеризация*). Разбиение предметной области на тематические категории осуществляется путем разбиения на классы множества всех ПП-структур. Эта задача тесным образом связана с задачей разбиения полнотекстового документа на монотематические фрагменты;

– *автоматическое реферирование и аннотирование текстовых документов*. Реферирование основывается на выявлении в тексте информативных предложений. Для реализации аннотирования (пересказа краткого содержания) в информационной системе следует предусмотреть достаточное для предметной области количество типов ситуативных связей между ПП-структурами;

– *информационный поиск*. При документальном поиске в качестве поискового предписания используется совокупность ПП-структур, полученная в результате индексирования запроса на естественном языке. Фактографическому поиску в полнотекстовом документе предшествует его разбиение на монотематические фрагменты.

Создание ситуативно-синтагматической сети в конкретной информационной системе сводится к построению ситуативно-синтагматического словаря (см. табл. 1) путем программной реализации алгоритма подсчета частоты совместной встречаемости пар ПП-структур в предложениях полного корпуса текстов.

Решение перечисленных выше задач возможно также на основе базы знаний в виде сети семантических структур, однако ее использование не представляется целесообразным для этих задач в связи с большим объемом «ручной» работы на этапе создания базы знаний. В человеко-машинном режиме создаются все словари базы знаний (словари семантических элементов, семантических сцен и семантических эпизодов). «Вручную» в этих словарях формируются списки понятий внутреннего языка и отношение строгого порядка при построении семантических сценариев, а также устанавливается соответствие между понятиями и слотами, слотами и семантическими ожиданиями.

Наиболее эффективным является применение сети семантических структур для реализации *машинного перевода* с одного естественного языка на другой с представлением промежу-

точных результатов на языке-посреднике (внутреннем языке информационной системы). При анализе входного текста и при синтезе выходного используются двуязычные словари «Входной–внутренний» (см. табл. 2) и «Внутренний–выходной» (см. табл. 3). Сеть семантических структур применяется главным образом с целью устранения лексической и синтаксической омонимии, а также для анализа неграмматичных предложений.

Список литературы

1. Осипов, Г.С. Приобретение знаний интеллектуальными системами: основы теории и технологии / Г.С. Осипов. – М.: Наука, 1997. – 112 с.

2. Демьянков, В.З. Интерпретация, понимание и лингвистические аспекты их моделирования на ЭВМ / В.З. Демьянков. – М.: Изд-во Моск. ун-та, 1989. – 172 с.

3. Хан, У. Системы автоматического реферирования / У. Хан, И. Мани // Открытые системы [Электронный ресурс]. – 2000. – № 12. – Режим доступа: http://www.osp.ru/os/2000/12/178370. – Дата доступа: 24.04.2007.

4. Технологии извлечения знаний из текста / Н. Ильин [и др.] // Открытые системы [Электронный ресурс]. – 2006. – № 6. – Режим доступа: http://www.i-teco.ru/ article104.html. – Дата доступа: 24.04.2007.

5. Липницкий, С.Ф. Математическая модель синтаксического анализа текста в информационно-аналитической системе / С.Ф. Липницкий // Информатика. – 2004. – № 1. – С. 28–36.

6. Реформатский, А.А. Введение в языковедение: учебник для вузов / Под ред. В.А. Виноградова / А.А. Реформатский. – М.: Аспект Пресс, 2002. –536 с.

7. Липницкий С.Ф. Семантический анализ текста на основе ситуативно-синтагматической сети / С.Ф. Липницкий // Информатика. – 2005. – № 2 (6). – С. 102–110.

Поступила 23.03.07

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

S.F. Lipnitsky

OBJECT DOMAIN KNOWLEDGE MODELS FOR INFORMATION SEARCH AND TEXT PROCESSING

Object domain knowledge models in a system of intellectual text processing are offered. These models could be used for document information retrieval and data search, classification of text documents, automatic abstracting and machine translation.