ИНФОРМАТИКА

2005 **апрель-июнь** **№ 2**

УДК 681.322

С.Ф. Липницкий

СЕМАНТИЧЕСКИЙ АНАЛИЗ ТЕКСТА НА ОСНОВЕ СИТУАТИВНО-СИНТАГМАТИЧЕСКОЙ СЕТИ

Предложена модель базы знаний информационной системы в виде ситуативно-синтагматической сети, элементами которой являются синтаксические деревья информативных синтагматических структур, а также семантические и ситуативные связи между ними. Средствами модели разработаны алгоритмы семантического анализа, который сводится к построению маршрута информативности текстового документа и его семантического следа в сети.

Введение

Повышение эффективности процессов поиска и семантической обработки текстовых документов достигается за счет интеллектуализации информационных систем на основе технологий представления и накопления знаний с целью «компьютерного понимания» текстов и их фрагментов [1, 2]. Как отмечается в [3], такое понимание связано с моделированием право- и левополушарной мыслительной деятельности человека [4]. При этом модель правополушарного мышления обеспечивает восприятие основного содержания документа, а левополушарный логический (семантико-синтаксический) анализ текста используется для более детального его осмысления.

Для реализации семантического анализа текстовых документов в данной статье построена математическая модель базы знаний, главным компонентом которой является ситуативно-синтагматическая сеть, состоящая из информативных синтагматических структур и ситуативных связей между ними.

1. Ситуативно-синтагматическая сеть

Понимание текста человеком связано со знанием языка, с одной стороны, и распознаванием ситуативного контекста – с другой [2]. При отсутствии ситуативных знаний восприятие текста возможно только на лингвистическом уровне [5]. В связи с этим построим модель базы знаний информационной системы в виде ситуативно-синтагматической сети, т. е. смешанного графа, вершинами которого являются псевдоосновы слов, а дугами – семантические связи между ними. Ребрами в сети представлено ситуативное отношение.

1.1. Ситуативное отношение

Пусть V ($V \neq \varnothing$) – некоторое непустое множество (*словарь*), элементы которого будем называть *словами*. Обозначим через a и b произвольные слова из словаря V, через Th_i ($i = \overline{1, n}$; $n \geq 1$) – тематические корпуса текстов, а через Fu ($Fu = Th_1 \cup Th_2 \cup \ldots \cup Th_n$) – полный корпус текстов. (Тематический корпус представляет собой объединение («склейку») текстов по конкретной тематике, а полный корпус – это объединение всех тематических корпусов.) Введем в рассмотрение для каждого корпуса текстов Th_i отношение толерантности (рефлексивное и симметричное бинарное отношение) Ψ_i на множестве V. Считаем, что упорядоченная пара слов (a, b) является элементом отношения Ψ_i тогда и только тогда, когда вероятность совместной встречаемости слов a и b в тематическом корпусе текстов Th_i не меньше некоторого порогового значения (*уровня ситуативной связи*). При выполнении этого условия отношение Ψ_i будем называть *ситуативным отношением в корпусе текстов* Th_i. Под совместной встречаемостью двух слов здесь понимается наличие их псевдооснов (или псевдооснов их синонимов) в одном и том же предложении корпуса Th_i.

1.2. Отношение парадигматического подчинения

Элементами ситуативно-синтагматической сети являются связи между синтагматическими структурами, соответствующие отношению парадигматического подчинения, которое определим следующим образом.

Пусть Δ – рефлексивное и транзитивное бинарное отношение (предпорядок) на множестве V^+ всех непустых цепочек в словаре V. Отношение Δ назовем *отношением парадигматического подчинения в словаре* V, если оно удовлетворяет следующему свойству: для любых цепочек $\beta, \gamma, \delta \in V^+$ существует цепочка $\alpha \in V^+$, для которой справедливы соотношения $(\alpha, \beta) \in \Delta$ и $(\alpha, \gamma) \in \Delta$ при выполнении условий $(\beta, \delta) \in \Delta$ и $(\gamma, \delta) \in \Delta$.

Если для любых цепочек $\mu, \nu \in V^+$ справедливо соотношение $(\mu, \nu) \in \Delta$, то будем говорить, что *цепочка* ν *парадигматически подчинена цепочке* μ *в словаре* V.

Содержательно предпорядок Δ соответствует родовидовому отношению между словами и фразами языка, а также отношениям «часть–целое», «причина–следствие» и т. п.

1.3. Отношение синтаксического подчинения

Для моделирования синтаксической структуры предложений введем понятие отношения синтаксического подчинения.

Пусть $F = \langle V, N, I, R\rangle$ – формальная порождающая грамматика, где V – множество ее терминальных символов (введенный в п. 1.1 словарь); $N = \{I, '\}$ – множество нетерминальных; I – начальный символ, а R – схема грамматики, т. е. множество правил вывода вида $\alpha \to \beta$ (α и β – различные цепочки в полном словаре $V \cup N$). Схема R грамматики F удовлетворяет свойствам:

– 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 $a, b \in V$.

Обозначим через $\pi = a_1 a_2 \ldots a_m$ произвольное предложение, где $a_1, a_2, \ldots, a_m$ – его слова (точнее, вхождения слов в это предложение), а через μ и ν – непустые непересекающиеся (не имеющие общих вхождений слов) подцепочки предложения π. Тогда бинарное отношение Ω_π на множестве всех таких подцепочек предложения π назовем *отношением синтаксического подчинения в предложении* π, если:

– для любых слов a_i, a_j $(i, j = \overline{1, m};\ i \neq j)$ предложения π справедливо соотношение $(a_i, a_j) \in \Omega_\pi$ тогда и только тогда, когда в выводе предложения π из начального символа I присутствуют цепочки $\alpha a_i' \beta$, $\gamma a_i' a_j' \delta$ (или $\gamma a_j' a_i' \delta$), для которых выполняется условие выводимости $\alpha a_i' \beta \square \gamma a_i' a_j' \delta$ (или $\alpha a_i' \beta \square \gamma a_j' a_i' \delta$). Здесь $\square$ – символ выводимости в грамматике F, а $\alpha, \beta, \gamma, \delta$ – цепочки в словаре $V \cup N$. Некоторые из цепочек $\alpha, \beta, \gamma, \delta$ могут быть пустыми (возможно, все). Если $i < j$ (или $j < i$), то цепочку $a_i a_j$ (или $a_j a_i$) будем называть *синтагмой предложения* π. При $j \neq i + 1$ (или $i \neq j + 1$) синтагму $a_i a_j$ (или $a_j a_i$) назовем *разделенной*, а при $j = i + 1$ (или $i = j + 1$) – *неразделенной*;

– для произвольных непустых непересекающихся подцепочек μ и ν предложения π условие $(\mu, \nu) \in \Omega_\pi$ выполнено тогда и только тогда, когда существует синтагма $a_i a_j$ предложения π, такая, что в выводе предложения π из начального символа I цепочка μ получена из a_i', а цепочка ν – из a_j'.

Обозначим через $\prec$ линейный порядок на множестве всех непустых непересекающихся подцепочек предложения π, соответствующий естественному порядку слов в нем, т. е. такой, что для всех $i, j = \overline{1, m-1}$, а также $r, s = \overline{1, m}$ соотношение $a_i a_{i+1} \ldots a_j \prec a_r a_{r+1} \ldots a_s$ справедливо в том и только в том случае, когда $j < r$. Тогда если $\mu \prec \nu$ (или $\nu \prec \mu$), то цепочку $\mu\nu$ (или $\nu\mu$) назовем *синтагматической структурой предложения* π. При этом будем говорить, что μ – *определяемый*, а ν – *определяющий* члены синтагматических структур $\mu\nu$ и $\nu\mu$.

Ориентированный граф сужения отношения Ω_π на множество всех вхождений слов предложения π назовем *синтаксическим графом предложения* π. Синтаксическим графом предложения, состоящего из одного слова a, будем считать граф $(\{a\}, \varnothing)$. Синтаксическим графом

любой цепочки δ, полученной из предложения π транспозицией некоторых ее слов, будем называть синтаксический граф предложения π.

Справедливо [6]

У т в е р ж д е н и е 1. *Синтаксический граф любого предложения языка*, *порождаемого грамматикой F*, *является ордеревом* (назовем его *синтаксическим деревом*).

1.4. Информативность синтагматических структур

Информативность синтагматических структур вычисляется с использованием результатов синтаксической и статистической обработок тематических корпусов текстов Th_i и полного корпуса текстов *Fu*.

Рассмотрим следующую совокупность событий:

S_{Fu} – извлечение случайным образом некоторой синтагматической структуры α из тематического корпуса текстов (или текстового документа) *Th* (*Th* ∈ *Fu*);

S_{Fu} – извлечение синтагматической структуры α из полного корпуса текстов *Fu*;

H_{Th} – появление тематического корпуса текстов (или документа) *Th*.

Пусть $P(S_{Th} / S_{Fu})$ – условная вероятность того, что синтагматическая структура α извлечена из множества *Th* при условии, что она уже извлечена из полного корпуса текстов *Fu*. Эта вероятность вычисляется следующим образом:

$$P(S_{Th} / S_{Fu}) = \frac{P(S_{Th} \cdot S_{Fu})}{P(S_{Fu})} = \frac{P(S_{Th}) \cdot P(S_{Fu} / S_{Th})}{P(S_{Fu})}.$$

Вероятность $P(S_{Th} / S_{Fu})$ будем называть *информативностью* синтагматической структуры α в тематическом корпусе текстов (или текстовом документе) *Th*.

Условная вероятность $P(S_{Fu} / S_{Th}) = 1$, поскольку событие, состоящее в том, что синтагматическая структура α извлечена из полного корпуса *Fu* при условии, что она уже извлечена из тематического корпуса *Th*, является достоверным, так как *Th* – подмножество множества *Fu*. В итоге имеем

$$P(S_{Th} / S_{Fu}) = \frac{P(S_{Th})}{P(S_{Fu})}.$$

Вычислив $P(S_{Th})$ по формуле полной вероятности, получим

$$P(S_{Th} / S_{Fu}) = \frac{P(S_{Th} / H_{Th})}{P(S_{Fu})} \cdot P(H_{Th}).$$

При достаточно больших объемах полного корпуса текстов *Fu* и тематического *Th* можно считать, что

$$P(S_{Th} / H_{Th}) \approx \frac{n_{Th}}{N_{Th}},\ P(S_{Fu}) \approx \frac{n_{Fu}}{N_{Fu}},\ P(H_{Th}) \approx \frac{N_{Th}}{N_{Fu}},$$

где n_{Th}, n_{Fu} – абсолютные частоты встречаемости (с точностью до синонимии и совпадения псевдооснов слов) синтагматической структуры α в тематическом и полном корпусах текстов, а N_{Th}, N_{Fu} – число вхождений всех синтагматических структур типа α (т. е. имеющих синтаксические деревья, совпадающие с синтаксическим деревом цепочки α с точностью до его вершин) в корпуса текстов *Th* и *Fu* соответственно. Тогда формула для вычисления информативности I^{α}_{Th} синтагматической структуры α в тематическом корпусе текстов (или текстовом документе) *Th* примет вид

$$I^{\alpha}_{Th} = \frac{n_{Th}}{n_{Fu}}.$$

1.5. Определение ситуативно-синтагматической сети

Пусть G_1 – множество синтаксических деревьев всех синтагматических структур, информативность которых в тематическом корпусе текстов Th_i превышает некоторый пороговый уровень (*уровень информативности*). Пометим все дуги ордеревьев из множества G_1 символом «Ω», который указывает, что эти дуги соответствуют отношению синтаксического подчинения. Полученный орграф (точнее, совокупность ордеревьев) обозначим через G_2.

Рассмотрим произвольные синтагматические структуры α и β, синтаксические деревья которых имеются в графе G_2. Пусть $(\alpha, \beta) \in \Delta$, а слова *a* и *b* являются абсолютно определяемыми членами структур α и β, т. е. корнями соответствующих им синтаксических деревьев. Соединим каждую такую пару вершин в графе G_2 дугой, направленной от *a* к *b* и помеченной символом, обозначающим тип отношения парадигматического подчинения (например, *Ч–Ц* – часть–целое, *Р–В* – род–вид, *П–С* – причина–следствие). Обозначим полученный граф через G_3. Соединим каждую пару вершин *a*, *b* графа G_3 ребром, если $(a, b) \in \Psi_i$. Полученный в результате граф с помеченными дугами обозначим через Sit_i и назовем его *тематическим ситуативно-синтагматическим графом*. Совокупность $Sit = \{Sit_i \mid i = \overline{1, n}\}$ будем называть *ситуативно-синтагматической сетью*. Пример фрагмента ситуативно-синтагматической сети приведен на рис. 1, где символ «/» указывает, что вершинами графа *Sit* являются псевдоосновы слов.

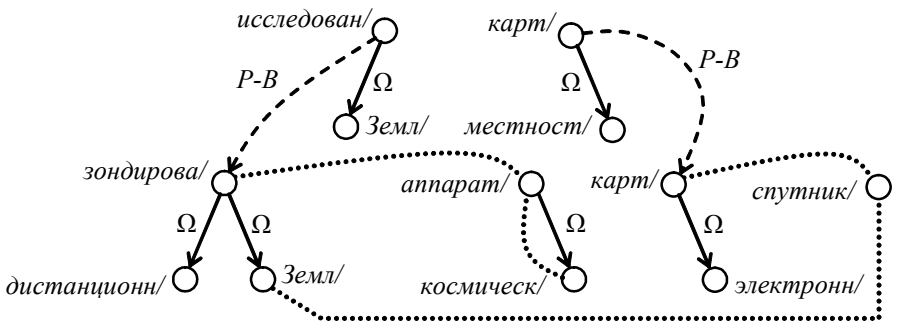

Рис. 1. Фрагмент ситуативно-синтагматической сети

2. Словари базы знаний

При программной реализации информационной системы ситуативно-синтагматическую сеть целесообразно представить в памяти компьютера в виде совокупности словарей.

2.1. Парадигматический словарь

Используя отношение парадигматического подчинения, определим формально понятие категории как класса слов и словосочетаний, объединенных общими семантическими признаками.

Непустое множество $K \subset V^+$ цепочек, для которых существует цепочка $\alpha \in V^+$, такая, что для всех цепочек $\beta \in K$ имеет место соотношение $(\alpha, \beta) \in \Delta$, назовем *категорией с именем* α, или α-*категорией*.

Нетрудно доказать

У т в е р ж д е н и е 2. *Если для имени* α *некоторой категории и для имени* β *любой другой категории справедливо соотношение* $(\alpha, \beta) \in \Delta$, *то* β-*категория является подмножеством* α-*категории*.

Образуем множество категорий *K′*, именами которых являются некоторые синтагматические структуры π_j ($j = \overline{1, l}$). Пусть имеется также множество цепочек *Par* из V^+, обозначающих тип отношения Δ, таких, что для любой цепочки $\tau \in Par$ существует хотя бы одна синтагматическая структура π_r $(1 \le r \le l)$, для которой справедливо соотношение $(\tau, \pi_r) \in \Delta$. В силу утвер-

ждения 2 π_r-категория из множества K' является подмножеством τ-категории. Если множество *Par* включает все такие π_j-категории, то назовем *Par парадигматическим словарем*. Пример фрагмента этого словаря показан на рис. 2, где стрелкой указано направление синтаксической связи, задаваемое отношением Ω_π.

Синтагматическая структура	Тип отношения	Парадигматически подчиненные синтагматические структуры
•••		
автоматизированн/ое ← мест/о, рабоч/ее ← мест/о	*Ч-Ц*	*персональн/ый ← компьютер/*
		программ/ное ← обеспечен/ие
•••		
информационн/ая ← технолог/ия	*Р-В*	*автоматическ/ое ← реферировн/ание*
		информационн/ый ← поиск
		распознаван/ие → образ/ов
•••		
отключ/или → электричеств/о	*П-С*	*перестал/ ← компьютер/, перестал/ → работ/ать*
•••		

Рис. 2. Фрагмент парадигматического словаря

2.2. Словарь синтагматических структур

Словарь включает синтагмы и синтагматические структуры с их частотами в полном и тематических корпусах текстов.

Пусть α – некоторая синтагматическая структура, $P_{ПКТ}$ и $P_{ТК\text{-}i}$ ($i = \overline{1, n}$) – ее абсолютные частоты соответственно в полном и *i*-м тематических корпусах текстов. Тогда совокупность кортежей типа $\langle \alpha, P_{ПКТ}, P_{ТК\text{-}1}, P_{ТК\text{-}2}, \ldots, P_{ТК\text{-}n} \rangle$ будем называть *словарем синтагматических структур*. Пример фрагмента словаря приведен на рис. 3.

Синтагматическая структура	$P_{ПКТ}$	$P_{ТК\text{-}1}$	•••	$P_{ТК\text{-}n}$
•••				
автоматизирован/ое ← мест/о, рабочее/е ← мест/о	8 720	6 239	•••	587
•••				
информационн/ая ← технолог/ия	3 443	2 211	•••	871
•••				
отключ/или → электричеств/о	7 361	643	•••	2 443
•••				

Рис. 3. Фрагмент словаря синтагматических структур

Частным случаем синтагмы является слово. Псевдоосновы слов и их статистические характеристики целесообразно хранить в специальном словаре псевдооснов слов (рис. 4).

Псевдооснова слова	$P_{ПКТ}$	$P_{ТК\text{-}1}$	•••	$P_{ТК\text{-}n}$
•••				
автоматизирован/ое	8 720	6 239	•••	587
•••				
технолог/ия	3 443	2 211	•••	871
•••				
электричеств/о	7 361	643	•••	2 443
•••				

Рис. 4. Фрагмент словаря псевдооснов слов

2.3. Отношение парадигматической эквивалентности. Словарь синонимов

Элементами одной и той же категории являются цепочки (слова, синтаксические конструкции), которые имеют совпадающие смысловые значения независимо от контекста (например, «*языкознание – лингвистика – языковедение*», «*забастовка – стачка*»). Для таких цепочек введем следующее определение.

Бинарное отношение Λ на множестве V^+ назовем *отношением парадигматической эквивалентности в словаре V*, если для любых цепочек $\alpha, \beta \in V^+$ $(\alpha, \beta) \in \Lambda$ тогда и только тогда, когда $(\alpha, \beta) \in \Delta$ и $(\beta, \alpha) \in \Delta$.

Если $(\alpha, \beta) \in \Lambda$, то цепочки α и β будем называть *парадигматически эквивалентными*, или *синонимами*, *в словаре V*.

Нетрудно видеть, что отношение Λ симметрично, поскольку из соотношений $(\alpha, \beta) \in \Delta$ и $(\beta, \alpha) \in \Delta$ следует, что $(\alpha, \beta) \in \Lambda$ и $(\beta, \alpha) \in \Lambda$. Отношение Λ рефлексивно, поскольку рефлексивно Δ. Легко доказывается и транзитивность отношения парадигматической эквивалентности. Действительно, пусть $(\alpha, \beta) \in \Lambda$ и $(\beta, \gamma) \in \Lambda$. Тогда $(\alpha, \beta) \in \Delta$, $(\beta, \alpha) \in \Delta$, $(\beta, \gamma) \in \Delta$ и $(\gamma, \beta) \in \Delta$. В силу транзитивности отношения Δ справедливы соотношения $(\alpha, \gamma) \in \Delta$ и $(\gamma, \alpha) \in \Delta$, откуда и следует, что выполнено соотношение $(\alpha, \gamma) \in \Lambda$. Таким образом доказано

У т в е р ж д е н и е 3. *Отношение* Λ *есть эквивалентность на множестве* V^+.

Из утверждения 1 вытекает

У т в е р ж д е н и е 4. *Если категория имеет имя* β, *а цепочка* $\alpha \in V^+$ $(\alpha \neq \beta)$ *синонимична цепочке* β, *то* α *также является именем этой категории*.

Утверждение верно, поскольку, в силу синонимии цепочек α и β, справедливо соотношение $(\alpha, \beta) \in \Delta$.

Для синонимичных цепочек справедливо

У т в е р ж д е н и е 5. *Произвольные непустые цепочки* β *и* γ *в словаре V*, *такие*, *что имеет место соотношение* $(\beta, \gamma) \in \Lambda$, *являются элементами одной и той же категории.*

Действительно, пусть цепочка β является элементом некоторой α-категории. Докажем, что этой же категории принадлежит и цепочка γ. Согласно определению категории $(\alpha, \beta) \in \Delta$. Поскольку, в силу определения отношения Λ, справедливо соотношение $(\beta, \gamma) \in \Delta$, то, вследствие транзитивности отношения Δ, $(\alpha, \gamma) \in \Delta$, что и означает принадлежность цепочки γ категории с именем α.

С учетом утверждений 3–5 *словарем синонимов* будем называть множество категорий, элементы каждой из которых образуют смежный класс по эквивалентности Λ, а именем категории является синтагматическая структура из этого смежного класса (рис. 5).

Синтагматическая структура	**Синонимичные синтагматические структуры**
•••	
АРМ	*автоматизирова/нное ← мест/о*, *рабоч/ее ← мест/о*
•••	
забастовк/а	*стачк/а*
•••	
ПЭВМ	*персональн/ый ← компьютер/*
	ПК
•••	
языкознан/ие	*лингвист/ика*
	языковед/ение
	наук/а → о, *о → язык/е*
•••	

Рис. 5. Фрагмент словаря синонимов

2.4. Ситуативные словари

Ситуативный словарь как компонент базы знаний информационной системы создается для каждого i-го ($i = \overline{1, n}$) тематического корпуса текстов Th_i и соответствует ситуативному отношению Ψ_i. Элементами словаря являются кортежи вида $\langle \alpha, \beta, P'_{TK\text{-}i} \rangle$, где α и β – псевдоосновы ситуативно связанных слов, а $P'_{TK\text{-}i}$ – частота их совместной встречаемости в корпусе текстов Th_i (рис. 6).

Псевдооснова слова	**Псевдооснова слова**	$P'_{TK\text{-}i}$
•••		
аппарат/	*зондирова/*	3207
•••		
Земл/	*карт/*	1832
•••		
карт/	*спутник/*	2311
•••		

Рис. 6. Фрагмент ситуативного словаря

3. Маршрут информативности и семантический след текста

Пусть имеется текст (т. е. кортеж предложений) T. Установим тематику текста T и выберем соответствующий тематический корпус текстов Th_i ($1 \le i \le n$). (В зависимости от решаемой задачи тематика текстового документа определяется информационной системой автоматически или ее пользователем по рубрикатору.) Вычислим, используя полный корпус текстов Fu и тематический Th_i, информативность синтагматических структур всех предложений текста T. Исключим из T все неинформативные предложения, т. е. предложения, не содержащие информативных структур. В результате получим кортеж предложений (в порядке их следования в T) $T_{inf} = \langle \pi_1, \pi_2, \ldots, \pi_l \rangle$. Кортеж T_{inf} будем называть *маршрутом информативности текста T*.

Обозначим через $D_{inf} = \langle D_1, D_2, \ldots, D_l \rangle$ кортеж синтаксических деревьев предложений из множества T_{inf}. Пометим дуги всех ордеревьев из множества D_{inf} символом «Ω» (по аналогии с ситуативно-синтагматической сетью) и исключим псевдоокончания всех слов из предложений π_i ($i = \overline{1, l}$). Рассмотрим объединение совокупности преобразованных таким образом ордеревьев из D_{inf} и тематического ситуативно-синтагматического графа Sit_i. Полученный кортеж подграфов графа Sit_i назовем *семантическим следом текста T в тематическом ситуативно-синтагматическом графе* Sit_i (рис. 7).

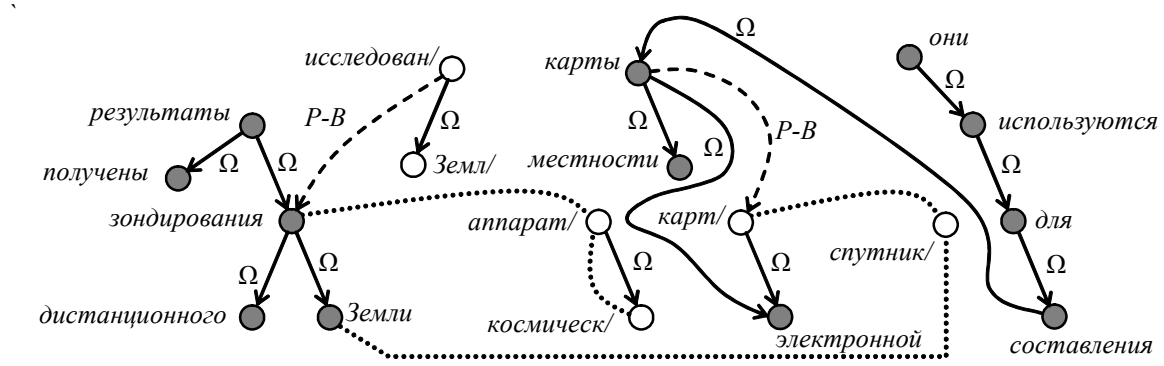

Рис. 7. Фрагмент семантического следа текста

4. Алгоритмы семантического анализа текста

Целью семантического анализа текста является получение его маршрута информативности и семантического следа в релевантном тексту тематическом ситуативно-синтагматическом графе.

4.1. Алгоритм построения маршрута информативности

На входе алгоритма находится текст T, полный и тематические корпуса текстов, словари базы знаний, на выходе – маршрут информативности. Алгоритм работает следующим образом.

1. Выполнить алгоритмы синтаксического анализа текста T из статьи [6], используя словарь синтагматических структур для установления синтаксических связей между словами предложений.

2. Найти релевантный тексту T тематический корпус текстов $Th \in \{Th_i \mid i = \overline{1, n}\}$.

3. Вычислить информативность каждой синтагматической структуры α предложений текста T по формуле

$$I_{Th}^{\alpha} = \frac{n_{Th}}{n_{Fu}},$$

используя зафиксированные в словарях синтагматических структур и псевдооснов слов абсолютные частоты встречаемости в полном корпусе текстов Fu и тематическом Th. Частоты синонимичных структур, найденных в словаре синонимов, суммируются.

4. Исключить из текста T все предложения, не содержащие информативных синтагматических структур.

5. Построить кортеж $T_{inf} = \langle \pi_1, \pi_2, \ldots, \pi_l \rangle$. КОНЕЦ (маршрут информативности текста T построен).

4.2. Алгоритм формирования семантического следа

Процесс формирования семантического следа на практике сводится к построению кортежа синтаксических деревьев информативных предложений входного текста и поиску записей парадигматического и ситуативного словарей. На входе алгоритма создания семантического следа находятся маршрут информативности $T_{inf} = \langle \pi_1, \pi_2, \ldots, \pi_l \rangle$ текста T и словари базы знаний, на выходе – семантический след текста T. Алгоритм включает следующие шаги.

1. $Sem := \varnothing$, $Z := \varnothing$.

2. Провести синтаксический анализ всех предложений π_i ($i = \overline{1, l}$) маршрута информативности T_{inf} текста T [6].

3. Сформировать кортеж $D_{inf} = \langle D_1, D_2, \ldots, D_l \rangle$ синтаксических деревьев информативных предложений π_i. Поместить кортеж D_{inf} в множество (файл) Sem.

4. Найти синтаксические деревья всех информативных синтагматических структур предложений π_i и поместить их в множество Z.

5. Найти в словаре синонимов все синонимы синтагматических структур из множества Z и пополнить их синтаксическими деревьями множество Z.

6. Найти в парадигматическом словаре все синтагматические структуры, которым парадигматически подчинены структуры с синтаксическими деревьями из множества Z, и поместить их в множество Sem.

7. Найти в парадигматическом словаре все синтагматические структуры, которые парадигматически подчинены структурам с синтаксическими деревьями из множества Z, и поместить их в множество Sem.

8. Найти в ситуативном словаре все слова, ситуативно связанные со словами-вершинами синтаксических деревьев из множества Z, и поместить их в множество Sem. КОНЕЦ (семантический след текста T сформирован в виде множества Sem).

Заключение

Предложенные в статье модель базы знаний и метод семантического анализа могут быть использованы для повышения эффективности информационных процессов в различных системах обработки текста. Например, в информационно-поисковых системах указанные модель и метод предназначены для повышения полноты и точности поиска, а также автоматизации процессов индексирования текстовых документов путем выявления в них информативных синтагматических структур. В системах машинного перевода ситуативно-синтагматическую сеть можно применить для разрешения лексической и синтаксической многозначности. В системах реферирования текстовых документов сеть может использоваться для синтеза связного реферата и тематического обобщения нескольких рефератов.

Список литературы

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

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

3. Харламов А.А., Ермаков А.Е., Кузнецов Д.М. Технология обработки текстовой информации с опорой на семантическое представление на основе иерархических структур из динамических нейронных сетей, управляемых механизмом внимания // Информационные технологии. – 1998. – № 2. – С. 26–32.

4. Белянин В.П. Введение в психолингвистику. – М.: Черо, 1999. – 123 с.

5. Пиотровский Р.Г. Инженерная лингвистика и теория языка. – Л.: Наука, 1979. – 111 с.

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

Поступила 15.04.05

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

S.F. Lipnitski

SEMANTIC TEXT ANALYSIS ON THE BASIS OF SITUATION-SYNTAGMATIC NETWORK

A situation-syntagmatic network model of knowledge base of the information system is proposed. Syntactic trees of informative syntagmatic structures as well as semantic and situation ties are elements of the model. Algorithms of semantic analysis by means of the model are developed. Semantic analysis means constructing the self-descriptiveness route of text document and its semantic track in the network.