

Е. Б. КАЙКОВА, В. Я. ТЕРЗИАН, канд. техн. наук

МЕХАНИЗМ ИССЛЕДОВАНИЯ ВРЕМЕННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТАХ

Пользователя в процессе общения с ЭВМ интересует не только явно заданная системе информация, но и (чаще в большей степени) вся дополнительная информация, которая может быть получена посредством умозаключений. А это предопределяет необходимость адекватного отображения в системе последовательности событий во времени. В статье рассматриваются вопросы анализа связных текстов, предполагающего построение временных последовательностей.

Задачами работы являются: 1 — разработка механизма выявления временной последовательности фактов внутри текста на основе ее явных и неявных признаков; 2 — экспериментальное исследование полученного механизма в рамках действующей диалоговой естественно-языковой (ЕЯ) системы. Заметим, что временная последовательность считается определенной, если любой паре фактов можно поставить в соответствие отношение типа «раньше», «позже», «одновременно».

Дадим классификацию признаков временной последовательности пар фактов в ЕЯ текстах. Основные типы — явно и неявно заданные признаки.

ВМ1: МС1: раньше (МФ3) = (МФ4),
 ВМ2: МС2: одновременно (МФ5) = (МФ4),
 ВМ3: МС3: позже (МФ6) — (МФ5),

Здесь ВМ — метка временного маркера, МС — метка семантического отношения, МФ — метка факта.

Гораздо сложнее выявить временную последовательность в тех случаях, когда она задана с использованием неявных признаков первого и второго классов. Остановимся подробнее на анализе признаков каждого класса.

Наличие причинно-следственной связи между фактом $МФ_\alpha$ и $МФ_\beta$ может быть установлена по следующему правилу:

$\forall МФ_\alpha \forall МФ_\beta \exists \{МФ1, МФ2, \dots, МФN\} \in KNW \exists \{МС1, МС2, \dots, МСN\} \in KNW$ (МС1: результат (МФ $_\alpha$) = (МФ1); МС2: условие (МФ2) = (МФ1); ...; МС(N-1): результат (МФ(N-1)) = (МФN); МСN: условие (МФ $_\beta$) = (МФN) \vdash $МФ_\alpha \Rightarrow МФ_\beta$, (1)

где KNW — множество фактов и отношений базы знаний; \vdash — «приводит к»; \Rightarrow — обозначение причинно-следственной связи между фактами.

Данная цепочка фактов отражает временную последовательность событий:

$$МФ_\alpha \rightarrow МФ2 \rightarrow МФ4 \rightarrow \dots \rightarrow МФ(N-1) \rightarrow МФ_\beta. \quad (2)$$

Присутствие промежуточных звеньев указанной цепи фактов в исходном тексте позволяет и для них определить место либо на псевдофизической, либо на семантической временной оси.

При анализе неявных признаков временной последовательности второго класса выполнение двух фактов $МФ_\alpha$ и $МФ_\beta$ несовместимо по времени, если

$\forall МФ_\alpha \forall МФ_\beta \exists МФ1 \in KNW \exists МФ2 \in KNW \exists МС1(МС1: \text{условие } (МФ_\alpha) = (МФ1)) \exists МС2(МС2: \text{условие } (МФ_\beta) = (МФ2)) (МФ1 \neq \neq МФ2) \vdash МФ_\alpha \neq МФ_\beta,$ (3)

где $МФ1 \neq МФ2$ означает противоречивость фактов $МФ1$ и $МФ2$; \neq — несовместимость фактов по времени.

В диалоговой системе помимо понятия факта широко используется понятие синтактико-семантического отношения (ССО) [2]. В дальнейшем тоже будем оперировать этим понятием, поэтому введем предварительно формат ССО:

$$R_i: МСИ_t(R'_i) = (R'_i). \quad (4)$$

Здесь МСИ — морфолого-синтаксическая информация словоформы из правой части ССО; R_i — метка ССО; R'_i — левая часть ССО; R'_i — правая часть ССО.

Пример: $R1$: какой (семинар) = (всесоюзный).

Остановимся подробнее на вопросах противоречивости фактов, которая определяется противоречивостью ССО, составляющих факт. Противоречивость ССО, в свою очередь, определяется противоречивостью составляющих их элементов.

Рассмотрим противоречивость ССО, определяемую их правой частью. Выбор правой части ССО в качестве определяющей вызван тем, что каждая словоформа некоторой естественно-языковой конструкции обязательно должна присутствовать в правой части одного из ССО, составляющих данную естественно-языковую конструкцию (ЕЯК). При определении противоречивости ССО будем использовать классификацию противоречий, изложенную в [3].

Определение 1. Два ССО R_i и R_j такие, что $R_i^r = R_j^r$, являются противоречивыми, если выполняется одно из следующих условий:

- 1) $(R_i^r = \text{HE}(R_j^r))$ или $(\text{HE}(R_i^r) = R_j^r)$;
- 2) $(R_i^r \in \text{MVAL}_{\langle \text{противоположно} \rangle}(R_j^r))$ или $(R_j^r \in \text{MVAL}_{\langle \text{противоположно} \rangle}(R_i^r))$,

где R^r — элемент из правой части ССО R ;

- 3) $\exists H_k ((H_k \in \text{KNW})(R_i^r \neq R_j^r)(R_i^r \in$

$\in \text{MVAL}_{\langle \text{элемент класса} \rangle}(H_k))(R_j^r \in \text{MVAL}_{\langle \text{элемент класса} \rangle}(H_k))$,

где $A \in \text{MVAL}_L(B)$ означает, что A связано с B с помощью СМНО L ; H — естественно-языковая конструкция.

При этом, если для данной пары ССО выполняются одновременно два условия противоречивости элементов правых частей, то эта пара ССО не считается противоречивой.

Рассмотрим примеры. Пара ССО: **что-делать** (Петя) = (читать) и **что-делать** (Петя) = (не, читать) является противоречивой по первому условию противоречивости. Пара ССО: **как** (бежать) = (быстро) и **как** (бежать) = (медленно) является противоречивой по второму условию противоречивости, поскольку: **противоположно** (быстро) = (медленно). Пара ССО: **какой** (шар) = (красный) и **какой** (шар) = (синий) является противоречивой по третьему условию противоречивости, поскольку: **элемент класса** (цвет) = (красный) и **элемент класса** (цвет) = (синий), в то время как пара ССО: **какой** (шар) = (красный) и **какой** (шар) = (воздушный) не является противоречивой, поскольку характеристики «красный» и «воздушный» относятся к разным классам. Пары ССО: **какой** (шар) = (красный) и **какой** (шар) = (не, синий); **как** (бежать) = (быстро) и **как** (бежать) = (не, медленно) не являются противоречивыми, поскольку для них выполняются одновременно два условия противоречивости.

Введем понятие взаимно однозначного семантического соответствия естественно-языковых конструкций.

Определение 2. Две ЕЯК H_i и H_j находятся во взаимно однозначном семантическом соответствии (обозначение — $H_i \leftrightarrow H_j^s$), если выполняется следующее условие:

$$(|MVAL(H_i)| = |MVAL(H_j)| = n) \exists Z_n (\forall L_k \in Z_n) \exists! (H_{im}, H_{jl}) \\ ((H_{im} \in MVAL(H_i)) (H_{jl} \in MVAL(H_j)) (H_{jl} \in MVAL_{L_k}(H_{im}))),$$

где $A \in MVAL(B)$ означает, что A входит в состав B .

Определение 3. Два естественно-языковых выражения H_i^p и H_j^p являются противоречивыми, если выполняется хотя бы одно из следующих условий:

$$a) (H_i^p \leftrightarrow H_j^p) \exists R_k (H_i^p) \exists R_l (H_j^p) ((R_k \in MVAL_{\langle \text{противоположно} \rangle} (R_l)) \\ \text{или } (R_l \in MVAL_{\langle \text{противоположно} \rangle} (R_k))),$$

где $R(H)$ — это ССО, принадлежащее ЕЯК H ;

$$b) (H_i^p \in MVAL_{\langle \text{противоположно} \rangle} (H_j^p)) \text{ или} \\ (H_j^p \in MVAL_{\langle \text{противоположно} \rangle} (H_i^p)).$$

В тех случаях, когда во фрагменте ЕЯ текста вообще отсутствуют какие-либо признаки временной последовательности фактов, последние рассматриваются по очередности их расположения в тексте. Для анализа подобных фрагментов строится дополнительная временная ось, которая объединяет временные маркеры появления фактов в связном тексте.

Анализ признаков всех четырех классов и определение очередности расположения фактов в тексте составляют основу алгоритма выявления временной последовательности в тексте.

Первым шагом алгоритма является разбиение текста на логически связанные фрагменты, относящиеся к описанию одной ситуации. Процедура разбиения текста использует правило соотнесения пары фактов к описанию одной ситуации, которое имеет вид

$$\forall M\Phi_i \forall M\Phi_j (M\Phi_i \xrightarrow{0} M\Phi_j) \exists S ((\exists R_k (M\Phi_i) \exists R_l (M\Phi_j)) \\ (GL(M\Phi_i) \xrightarrow{R_k(M\Phi_i)} S) (GL(M\Phi_j) \xrightarrow{R_l(M\Phi_j)} S))) \vdash SIT(M\Phi_i, M\Phi_j), \quad (5)$$

где $\xrightarrow{0}$ — означает, что факты следуют непосредственно друг за другом в некотором фрагменте текста; S — словоформа; $GL(M\Phi)$ — сказуемое в факте $M\Phi$; $R_i(M\Phi)$ — i -е ССО в разбиении факта $M\Phi$; $S_1 \xrightarrow{R_i(M\Phi)} S_2$ — означает, что в i -м ССО факта $M\Phi$ объединяются словоформы S_1 и S_2 ; SIT — означает, что факты относятся к описанию одной ситуации.

Среди всех предложений во фрагменте описания ситуации выделяются так называемые ядерные и подчиненные предложения.

Вторым шагом алгоритма как раз и является разбиение множества предложений (естественно-языковых описаний фактов) на подмножества ядерных и подчиненных предложений в соответствии с правилом:

$$\forall M\Phi_\alpha \forall M\Phi_\beta \exists MC (MC : ST (M\Phi_\alpha) = (M\Phi_\beta)) \vdash Я (M\Phi_\beta) \& П (M\Phi_\alpha). \quad (6)$$

Здесь ST — имя временного отношения; $Я(M\Phi)$ означает, что факт $M\Phi$ является ядерным; $П(M\Phi)$ означает, что факт $M\Phi$ является подчиненным.

Следующие шаги алгоритма — построение фрагментов временных осей для временных маркеров ядерных предложений. Последовательность построения определяется наличием во фрагменте описания ситуации признаков временной последовательности различных классов в соответствии с описанной выше классификацией. Наиболее полный алгоритм включает в себя 3 шага построения.

1. Формирование общей структуры фрагмента текста, т. е. расположение на основных временных осях $ВМ$ ядерных предложений, содержащих явные признаки временной последовательности 1-го или 2-го классов. Сформированная общая структура является базой для построения полной логической модели времени текста.

2. Пополнение общей структуры текста $ВМ$ предложений, содержащих неявные признаки временной последовательности первого класса.

3. Включение в структуру $ВМ$ ядерных предложений, содержащих неявные признаки второго класса, с учетом расположения этих $ВМ$ на дополнительной временной оси.

Затем модель времени дополняется промежуточными отрезками логических осей, на которых располагаются $ВМ$ фактов, описанных в подчиненных предложениях фрагмента текста. Полученная полная модель времени фрагмента текста, описывающего одну ситуацию, включается в логическую модель времени текста. Таким образом, каждому тексту в базе знаний диалоговой системы будет соответствовать логическая модель временной последовательности, адекватно отражающая последовательность выполнения действий во времени.

Нагляднее работа предложенного механизма может быть продемонстрирована на примере анализа следующего текста: «Сегодня с 3-х часов до 4-х часов Максимов читал лекцию в университете. Вечером он беседовал со студентом Вороновым о дипломе. Вчера утром до 10 часов он редактировал доклад. За неделю до этого, 3-го апреля, он выступал на Международной конференции в Москве».

Предварительно в системе формируется понятие «сегодняшняя дата». Для этого пользователь в режиме локального диалога должен объяснить: «Сегодня — 10 апреля 1985 года, 12

часов». На базе этого понятия формируются временные маркеры времени поступления естественно-языковых описаний фактов МФ (предложений текста) на вход диалоговой системы — ВМТ. Полученные ВМ располагаются на дополнительной временной оси (рис. 1).

Согласно правилу 5 все предложения рассматриваемого текста относятся к описанию одной ситуации. Разбиваем все множество предложений на подмножества ядерных и подчиненных. Последнее состоит в данном случае из одного предложения: «За неделю до этого, 3-го апреля, он выступал на Международной конференции в Москве». Однако в этом предложении помимо явной связки временного отношения «...до этого» (2-й класс явных признаков) присутствует временная конструкция «3 апреля» (1-й класс явных признаков); поэтому

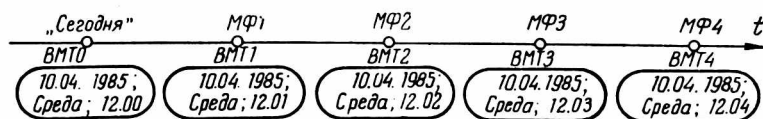


Рис. 1

система просто игнорирует связку временного отношения. Факт будет выглядеть следующим образом: «3-го апреля он выступал на Международной конференции в Москве». Таким образом, все предложения текста в данном случае ядерные.

Так как все предложения рассматриваемого текста содержат явные признаки временной последовательности, алгоритм сворачивается до одного шага построения — формирования общей структуры текста. Факту МФ1: «Максимов читал лекцию в университете» соответствуют два ВМ:

ВМ1: 11, 00, 001000000, 10, 04, 1985, 15, 00, 00, МФ1;

ВМ2: 11, 00, 000000001, 10, 04, 1985, 16, 00, 00, МФ1,

где 11 — код грамматического времени глагола; 00 — признак отсутствия квантора; 001000000 — признак начала временного интервала; 000000001 — признак конца временного интервала; 10, 04, 1985 — дата выполнения действия; 16, 00, 00 — время выполнения действия; МФ1 — метка факта, к которому относится временной маркер.

По умолчанию считается истинным тот факт, что в тексте понятие «сегодня» соотносится со временем поступления текста на вход системы. Если же есть дополнительное указание на то, что факты соотнесены со временем составления текста, а не со временем поступления в систему, то временные точки маркеров смещаются по псевдофизической оси соответственно влево или вправо.

В описании фактов МФ2 и МФ3 в качестве явных признаков 1-го класса используются размытые временные харак-

теристики «вечером», «утром». Для их преобразования подключаются специальные модули врожденных или сформированных в процессе общения с пользователем процедурных знаний. Словоформа «вечером» заменяется временным интервалом «с 18 часов до 22 часов», а словоформа «утром» — временным интервалом «с 8 часов до 11 часов». Аналогично преобразовывается относительная временная характеристика 1-го класса явных признаков «вчера». После преобразования факту МФ2 «Вечером он беседовал со студентом Вороновым о дипломе» приписываются ВМ интервала действия:

ВМ3: 11, 00, 001000000, 10, 04, 1985, 18, 00, 00, МФ2;

ВМ4: 11, 00, 000000001, 10, 04, 1985, 22, 00, 00, МФ2,

а факту МФ3 «Вчера утром до 10 часов он редактировал доклад» —

ВМ5: 11, 00, 001000000, 09, 04, 1985, 08, 00, 00, МФ3;

ВМ6: 11, 00, 000000001, 09, 04, 1985, 10, 00, 00, МФ3,

причем правая граница интервала скорректирована в соответствии с дополнительной временной характеристикой «до 10 часов». И, наконец, формируется ВМ для факта МФ4 «Он выступал на Международной конференции в Москве»:

ВМ7: 11, 00, 000001000, 03, 04, 1985, МФ4.

Все полученные ВМ сравниваются по абсолютному значению элементов формата и в соответствии с этим значением располагаются на временной псевдофизической оси (рис. 2). Логическая модель времени данного текста, таким образом, будет представлять собой отрезки на псевдофизической основной и дополнительной временных осях.

Если добавить к рассматриваемому тексту два новых предложения: «В ходе выступления Максимова на конференции были высказаны пожелания по переориентации созданной им

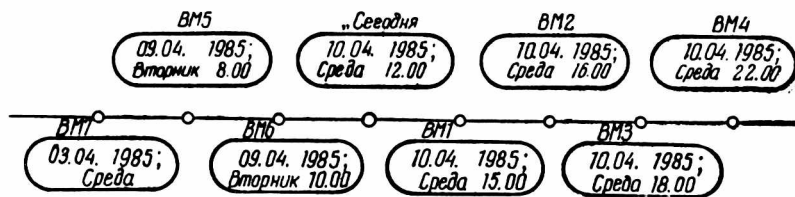


Рис. 2

системы. Ответы на высказанные пожелания Максимов включил в свой доклад», то в этом случае к работе будет подключен полный алгоритм. Так как в описании факта МФ5 «В ходе выступления Максимова...» используется временная конструкция «в ходе» (2-й класс явных признаков), то для пары фактов МФ4 и МФ5 будет сформировано временное от-

ношение «одновременно» и факту МФ5 будет приписан ВМ, совпадающий с ВМ7: ВМ8: 11, 00, 000001000, 03, 04, 1985, МФ5.

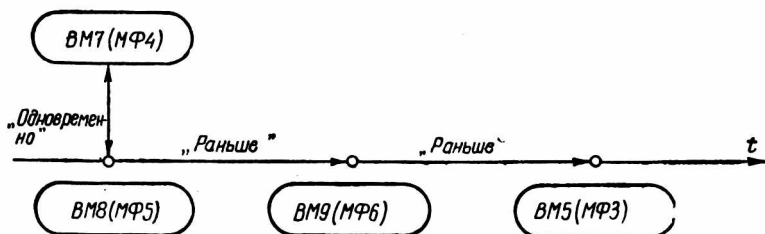


Рис. 3

Проанализируем описание факта МФ6 «Ответы на высказанные пожелания Максимов включил в свой доклад». В предложении нет явных ссылок на временную последовательность. Поэтому необходимо выявить причинно-следственные связи фактов между собой (1-й класс неявных признаков).

Рассмотрим обобщенные факты:

Ф1: X1 (кто) включил X2 (что-в) в X3 (куда),

Ф2: X1 (кто) имеет X2 (что),

Ф3: X3 (кто) высказал X2 (что-в) X1 (кому),

Ф4: X1 (кто) редактирует X3 (что-в),

Ф5: X1 (кто) имеет полный X3 (что-в),

которые связаны в базе знаний диалоговой системы с помощью СМНО:

условие (Ф1) = (Ф2);

результат (Ф3) = (Ф2);

условие (Ф4) = (Ф5);

результат (Ф1) = (Ф5).

Отсюда, согласно правилу 1, следует причинно-следственная связь между парами фактов:

Ф3 → Ф1 и Ф1 → Ф4, которая определяет последовательность выполнения действий во времени. Применительно к рассматриваемому тексту последовательность будет выглядеть так, как это показано на рис. 3. Она представлена отрезком на семантической оси, так как факту МФ6 не может быть приписана абсолютная дата времени выполнения действия. Мы можем лишь сказать, что факт МФ6 произошел позже факта МФ5 (МФ4), но раньше факта МФ3. ВМ факта МФ6 будут иметь вид:

ВМ9: 11, 00, 001000000, ВМ5, МФ6;

ВМ10: 11, 00, 000000001, ВМ8, МФ6,

что означает, что факту МФ6 приписан временной интервал от ВМ8 до ВМ5.

Построенная логическая модель времени для связного ЕЯ текста представляет собой отрезки на псевдофизической и семантической основных и дополнительной временных осях; позволяет диалоговой системе отвечать на любой вопрос пользователя, касающийся временных характеристик событий из текста.

Предложенный алгоритм был программно реализован и апробирован в составе действующей диалоговой системы. Полученные результаты позволяют судить о целесообразности изложенного подхода.

Список литературы: 1. *Kaikova E. B.* Grammatical analysis and synthesis of time structures in computing structures // Symposium on Grammars of analysis and synthesis Summaries.— Tallinn, 1983.—Р. 33—35. 2. *Ловицкий В. А.* Обучаемая диалоговая естественно-языковая система // Техническая кибернетика: Изв. АН СССР.— 1983.— Вып. 5.—С. 114—127. 3. *Головина Е. А., Колмычек К. Н., Терзиян В. Я.* Принципы проверки семантической правильности естественно-языковых высказываний // Пробл. бионики.— 1984.— Вып. 32.— С. 64 — 72.

Поступила в редколлегию 23.01.86