УДК 681.322

С.Ф. Липницкий

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ СИНТАКСИЧЕСКОГО АНАЛИЗА ТЕКСТА В ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЕ

Предлагается метод синтаксического анализа текста, основанный на моделировании процесса распознавания его синтагм средствами специальной формальной грамматики (штрих-грамматики), которая обеспечивает универсальность метода для различных естественных языков. Формализованы основные семантические отношения в языке, с использованием которых исследованы условия существования в предложениях текста маргинальных синтагм. С учетом этих свойств разработаны алгоритмы синтаксического анализа проективных и непроективных предложений.

Введение

Назначение информационно-аналитической системы состоит в переработке больших файлов информации с целью ее интеллектуального анализа. Независимо от используемых методов аналитической обработки ей предшествует синтаксический анализ текста, который заключается в построении синтаксической структуры каждого его предложения. Различают поверхностный и глубинный анализ. В результате поверхностного устанавливается лишь факт наличия синтаксической связи между словами и ее направление. При глубинном анализе ставится цель дифференциации связей за счет привлечения семантических признаков при реализации анализа.

В статье предлагается метод синтаксического анализа текста, основанный на моделировании процесса распознавания его синтагм средствами специальной формальной грамматики (штрих-грамматики), которая обеспечивает универсальность метода для различных языков.

1. Анализ проективных предложений

Пусть $F = \langle V, N, I, R \rangle$ – формальная порождающая грамматика, где V – непустое множество терминальных элементов (назовем их *словами*), $N = \{I, '\}$ – множество нетерминальных, I – начальный символ, а R – схема грамматики, т.е. множество правил вывода вида $\alpha \to \beta$ (α и β – различные цепочки в словаре $V \cup N$). Схему R грамматики F определим следующим образом [1]:

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

2) 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 F будем называть *штрих-грамматикой* над словарем V. Язык $L(F)$, порождаемый штрих-грамматикой, назовем *входным* языком, его цепочки – *предложениями входного языка*, или *входными предложениями*, а словарь V – *словарем входного языка*, или *входным словарем*. Всякое непустое линейно упорядоченное подмножество языка $L(F)$ будем называть *текстом* этого языка, или *входным текстом*.

1.1. Синтагмы и синтагматические структуры

При моделировании синтаксической структуры предложений языка $L(F)$ будем использовать отношение синтаксического подчинения, понятие которого определим следующим образом.

Пусть $\pi = a_1 a_2 \ldots a_n$ – произвольное предложение языка $L(F)$, где $a_1, a_2, \ldots, a_n$ – вхождения слов в это предложение. (О вхождениях слов в цепочку, а не о словах говорят в связи с тем, что отдельные слова могут встречаться в ней более одного раза.) Обозначим через μ и ν некоторые 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 *языка* $L(F)$, если будут выполнены следующие условия:

1) для любых слов a_i, a_j $(i, j = \overline{1,n};\ i \neq j)$ предложения π $(a_i, a_j) \in \Omega_\pi$ 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 $\gamma a_j'a_i'\delta$), причем $\alpha a_i'\beta \vdash \gamma a_i'a_j'\delta$ (или $\alpha a_i'\beta \vdash \gamma a_j'a_i'\delta$). Здесь $\vdash$ – 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_ia_j (или a_ja_i) будем называть *синтагмой предложения* π *языка* $L(F)$. При $j \neq i+1$ (или $i \neq j+1$) синтагму a_ia_j (или a_ja_i) назовем *разделенной*, а при $j = i+1$ (или $i = j+1$) – *неразделенной*;

2) для произвольных непустых непересекающихся подцепочек μ и ν предложения π $(\mu, \nu) \in \Omega_\pi$ тогда и только тогда, когда существует синтагма a_ia_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 $i, j = \overline{1,n-1}$, $r, s = \overline{1,n}$ $a_ia_{i+1} \ldots a_j \prec a_ra_{r+1} \ldots a_s$ 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 *языка* $L(F)$. При этом будем говорить, что μ – *определяемый*, а ν – *определяющий* члены синтагматических структур $\mu\nu$ и $\nu\mu$.

Объединение $\Omega_{L(F)} = \bigcup\limits_{\pi \in L(F)} \Omega_\pi$ отношений синтаксического подчинения во всех предложениях языка будем называть *отношением синтаксического подчинения в языке* $L(F)$. Синтагмы и синтагматические структуры предложений этого языка назовем *синтагмами и синтагматическими структурами языка* $L(F)$.

1.2. Синтаксическое дерево предложения

Если ab – синтагма некоторого предложения языка $L(F)$ и $(a, b) \in \Omega_\pi$, то будем говорить, что синтаксическая связь *направлена* от слова a к слову b. 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, $\alpha\overrightarrow{a\beta b}\gamma$, $\alpha\overleftarrow{a\beta b}\gamma$). Если же направление связи неизвестно или несущественно, то условимся обозначать ее чертой над синтагмой (например, $\alpha\overline{a\beta b}\gamma$).

Синтаксические связи между словами предложения представляют обычно в виде ориентированного графа, вершинами которого являются вхождения слов в предложение, а дуги соответствуют синтаксическим связям между ними. Формально понятие синтаксического графа определим следующим образом [2].

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.

Установим вид синтаксического графа предложения входного языка $L(F)$.

Утверждение 1. *Синтаксический граф любого предложения языка* $L(F)$ *является ордеревом* (назовем его *синтаксическим деревом*).

Доказательство. Проведем доказательство индукцией по числу n вхождений слов в предложение. При $n = 1$ и $n = 2$ синтаксическими графами слова и синтагмы являются ордеревья. Предположим, что при $n = k$ синтаксический граф предложения, содержащего k вхож-

дений слов, есть ордерево. Докажем, что после включения в это предложение еще одного вхождения слова, т.е. при $n = k + 1$, синтаксический граф предложения останется ордеревом. Обозначим включаемое в предложение вхождение слова через b. Тогда в предложении имеется вхождение слова (обозначим его через a), являющееся, в силу определения грамматики F, определяемым членом синтагмы ab или ba. Если вершину a ордерева с k вершинами соединить с вершиной b дугой (a, b), то, очевидно, снова будем иметь ордерево. Утверждение 1 доказано.

Для описания синтаксической структуры предложения часто используется дерево синтаксического подчинения [3]. В рамках нашей модели это понятие определим следующим образом.

Синтаксическое дерево предложения $\pi = a_1 a_2 \ldots a_n$ называется *деревом синтаксического подчинения* этого предложения, если множеством его вершин является линейно упорядоченное множество $\langle \{a_1, a_2, \ldots, a_n\}, \prec \rangle$.

1.3. Проективные предложения

Проективность является показателем синтаксической правильности предложений естественного языка [4]. При формальном определении этого понятия будем использовать наглядный критерий из [3], согласно которому цепочка является проективной, когда в ее дереве синтаксического подчинения отсутствуют пересекающиеся дуги и корень дерева не лежит ни под одной из них. Введем предварительно некоторые вспомогательные понятия.

Рассмотрим произвольное предложение π языка $L(F)$ и цепочку δ, полученную из этого предложения транспозицией некоторых его слов. Всякую дугу, соединяющую в любом направлении вершины a_i и a_j дерева синтаксического подчинения цепочки δ, назовем *пересекающейся* с любой дугой, соединяющей вершины a_r и a_s (также в произвольном направлении), если $1 \le i < r < j < s \le n$. При $i < r < j$ будем говорить, что вершина a_r *лежит под дугой*, соединяющей вершины a_i и a_j. Число всех дуг, под которыми лежит вершина a_r, называют *степенью гнездования вершины* a_r [3].

Цепочку δ будем называть *проективной*, если в ее дереве синтаксического подчинения нет пересекающихся дуг и корень дерева не лежит ни под какой дугой, и *непроективной* во всех остальных случаях. Текст любого языка будем считать *проективным*, если проективны все его предложения.

Утверждение 2. *Все цепочки языка $L(F)$ являются проективными предложениями.*

Доказательство. Пусть n – число вхождений слов в предложение. При $n = 1$ и $n = 2$ слова и синтагмы соответственно являются проективными предложениями. Пусть при $n = k$ любое предложение языка $L(F)$ проективно. Докажем, что при $n = k + 1$, т.е. после добавления в это предложение одного вхождения слова, проективность не нарушится. Предположим, что в предложении из k вхождений слов добавлено некоторое вхождение слова b. Тогда в предложении должно присутствовать вхождение слова a, такое, что ab (или ba) – синтагма этого предложения. Предположим от противного, что дуга, соединяющая вершины a и b дерева синтаксического подчинения полученного предложения, пересекается с дугой, соединяющей вершины c и d некоторой синтагмы cd предложения. Тогда, в соответствии с определением проективного предложения, один из членов синтагмы ab (или ba) должен находиться между (в смысле линейного порядка $\prec$) вхождениями слов c и d, что противоречит определению схемы R грамматики F и предположению индукции. Утверждение 2 доказано.

1.4. Маргинальные синтагмы

Пусть $\alpha a \beta b \gamma$ (или $\alpha b \beta a \gamma$) – произвольное предложение языка $L(F)$, где $\alpha, \beta, \gamma \in V^*$ (V^* – множество всех цепочек в словаре V грамматики G), ab (или ba) – синтагма этого предложения с определяемым членом a и определяющим b.

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

синтагм αab и ba.

Пусть δ – произвольная цепочка из множества V^+ всех непустых цепочек в словаре V, подцепочкой которой является некоторое предложение π языка $L(F)$. Если $\mu\nu$ – синтагматическая структура предложения π, то будем считать ее и *синтагматической структурой цепочки* δ. Если же ab (или ba) – маргинальная синтагма предложения π с маргинальным словом b, таким, что для любого вхождения слова c ($c \neq b$) цепочки δ пары bc и cb не являются ее синтагмами, то ab (или ba) назовем *маргинальной синтагмой цепочки* δ.

Лемма. *Если* $\rho \in V^+$, *а* ab (*или* ba) – *маргинальная синтагма цепочки* ρ, *причем в схеме* R *грамматики* F *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 предложением языка* $L(F)$ *тогда и только тогда, когда* $\rho \in L(F)$.

Доказательство. *Необходимость.* Пусть цепочка σ является предложением языка $L(F)$. Тогда необходимость, т.е. выполнение соотношения $\rho \in L(F)$, следует из факта существования в схеме R грамматики F правил вывода $a' \to a'b'$ (или $a' \to b'a'$) и $a' \to a$, $b' \to b$.

Достаточность. Пусть имеется синтагма ab с определяемым членом a и определяющим b. Тогда 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 F, где $\alpha, \beta, \gamma, \mu, \nu \in V^*$. Поскольку ab – маргинальная синтагма предложения ρ, то, в силу определения маргинальной синтагмы, для любого слова c предложения ρ цепочка bc не является синтагмой, т.е. при выводе предложения 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 предложения ρ является цепочка ba. Лемма доказана.

Используя лемму, нетрудно доказать

Утверждение 3. *Если* $\mu a'b'\nu$ (*или* $\mu b'a'\nu$) – *некоторая цепочка в словаре* V, *где* $\mu, \nu \in V^*$, ab (*или* ba) – *маргинальная синтагма с определяющим членом* b, *причем в схеме* R *грамматики* F *имеется правило вывода* $a' \to a'b'$ (*или* $a' \to b'a'$), *то цепочка* $\mu a'\nu$ *возводима к начальному символу* I *грамматики* F *тогда и только тогда, когда к символу* I *возводима цепочка* $\mu a'b'\nu$ (*или* $\mu b'a'\nu$).

Доказательство. *Необходимость.* Пусть цепочка $\mu a'\nu$ возводима к начальному символу I. Докажем, что к символу I возводима цепочка $\mu a'b'\nu$. Действительно, применив к цепочке $\mu a'\nu$ правило вывода $a' \to a$, получим цепочку $\mu a\nu$, которая является предложением языка $L(F)$, откуда следует, в силу существования правил вывода $a' \to a$, $b' \to b$, что цепочка $\mu a'b'\nu$ возводима к символу I. Аналогично доказывается необходимость для цепочки $\mu b'a'\nu$.

Достаточность. Пусть теперь цепочка $\mu a'b'\nu$ возводима к начальному символу I. Доказательство того, что к этому символу возводится и цепочка $\mu a'\nu$, следует из достаточности леммы. Применим к цепочке $\mu a'b'\nu$ правила вывода $a' \to a$, $b' \to b$. Получим предложение $\mu ab\nu$ языка $L(F)$. В силу леммы предложением этого языка является и цепочка $\mu a\nu$, откуда следует, что цепочка $\mu a'\nu$ возводима к символу I. Доказательство возводимости к символу I цепочки $\mu a'\nu$ (при возводимости к нему цепочки $\mu b'a'\nu$) аналогично. Утверждение 3 доказано.

В соответствии с утверждением 3 алгоритм синтаксического анализа входных цепочек может быть построен в виде циклической процедуры сведения их к начальному символу по принципу "снизу вверх". При этом правила вывода применяются иначе, чем при порождении предложений: правые части правил заменяются их соответствующими левыми частями. Процедура анализа реализуется следующим образом. На первом шаге все слова входной цепочки "штрихуются", т.е. заменяются соответствующими цепочками с использованием правил вывода вида $a' \to a$ (например, слово a заменяется цепочкой a'). На втором шаге в цепочке ищутся подцепочки вида $a'b'$ или $b'a'$, где ab и ba – маргинальные синтагмы с определяемым членом a, и заменяются с помощью правил вида $a' \to a'b'$ (или $a' \to b'a'$) цепочками вида a'.

Далее второй шаг циклически повторяется. Признаком завершения процесса синтаксического анализа является получение начального символа I или цепочки, включающей более одного символа I. В последнем случае анализируемая цепочка, в силу утверждения 3, не является проективным предложением.

Из утверждения 2 и необходимости леммы 3 следует

Утверждение 4. *Если* $\rho \in L(F)$ *– произвольное предложение, а* ab (*или* ba) *– его маргинальная синтагма, то предложение* σ, *полученное из* ρ *удалением определяющего члена* b *синтагмы* ab (*или* ba), *проективно.*

Утверждение 4 обеспечивает получение проективного предложения после исключения из него определяющих членов всех неразделенных маргинальных синтагм. В соответствии с этим утверждением возведение синтагм с помощью правил вывода грамматики F может быть заменено более эффективной циклической процедурой. На первом шаге этой процедуры в анализируемом предложении выявляются неразделенные маргинальные синтагмы. На втором шаге из этих синтагм исключаются определяющие члены. Далее процесс повторяется аналогичным образом до получения в каждом предложении текста абсолютно определяемого члена в качестве единственного его слова.

1.4.1. Существование маргинальных синтагм. Исследуем условия существования неразделенных маргинальных синтагм проективного предложения $\pi = a_1 a_2 \dots a_n$ в синтагматических структурах следующих десяти типов: 1) $\overrightarrow{a_i a_{i+1}}$ $(i = \overline{1, n-1})$; 2) $\overleftarrow{a_i a_{i+1}}$ $(i = \overline{1, n-1})$; 3) $\overrightarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$ $(i = \overline{1, n-2})$; 4) $\overrightarrow{a_i \overleftarrow{a_{i+1} a_{i+2}}}$ $(i = \overline{1, n-2})$; 5) $\overrightarrow{a_i \dots \overrightarrow{a_j a_{j+1}} a_{j+2}}$ $(i = \overline{1, j-1},$ $j = \overline{2, n-2})$; 6) $\overrightarrow{\overrightarrow{a_i a_{i+1}} a_{i+2} \dots a_j}$ $(i = \overline{1, j-3},$ $j = \overline{4, n})$; 7) $\overrightarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}}}$ $(i = \overline{1, n-3})$; 8) $\overrightarrow{a_i \dots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}}}$ $(i = \overline{1, j-1},$ $j = \overline{2, n-3})$; 9) $\overrightarrow{\overrightarrow{a_i a_{i+1}} \overleftarrow{a_{i+2} a_{i+3}} \dots a_j}$ $(i = \overline{1, j-4},$ $j = \overline{5, n})$; 10) $\overrightarrow{a_i \dots \overrightarrow{a_j a_{j+1}} \overleftarrow{a_{j+2} a_{j+3}} \dots a_k}$ $(i = \overline{1, j-1},$ $j = \overline{2, k-4},$ $k = \overline{6, n})$.

1°. *Синтагма* $\overrightarrow{a_i a_{i+1}}$ *предложения* π *является маргинальной, если для любого* $j = \overline{i+2, n}$ *цепочка* $a_{i+1} a_j$ *не является синтагмой предложения* π.

Доказательство. Докажем сначала, что для любого $j = \overline{1, i-1}$ цепочка $a_j a_{i+1}$ не является синтагмой. Действительно, если предположить противное, то должны существовать правила, обеспечивающие вывод цепочки $a'_j \dots a'_i a'_{i+1}$ из цепочки $a'_i a'_{i+1}$, что противоречит определению синтагмы предложения π и его проективности. С учетом этого обстоятельства, леммы и определения маргинальной синтагмы предложения нетрудно видеть, что для любого $j = 1, 2, \dots, i-1, i+1, \dots, n$ цепочка $a_{i+1} a_j$ не является синтагмой предложения π, т.е. синтагма $\overrightarrow{a_i a_{i+1}}$ этого предложения маргинальная. Свойство 1° доказано.

Из свойства 1° вытекает очевидное, но важное

Следствие. *Синтагма* $\overrightarrow{a_{n-1} a_n}$ *является маргинальной синтагмой предложения* π.

Точно так же доказывается свойство 2°.

2°. *Синтагма* $\overleftarrow{a_i a_{i+1}}$ *предложения* π *является маргинальной, если для любого* $j = \overline{1, i-1}$ *цепочка* $a_j a_i$ *не является синтагмой предложения* π.

Следствие. *Синтагма* $\overleftarrow{a_1 a_2}$ *является маргинальной синтагмой предложения* π.

Замечание. На практике целесообразно использовать не свойства 1° и 2°, а следствия из них, поскольку достаточно сложно установить, что некоторая пара слов предложения не является его синтагмой.

3°. *Синтагматическая структура* $\overrightarrow{\overrightarrow{a_i a_{i+1}}\, a_{i+2}}$ *предложения* π *содержит маргинальную синтагму* $\overrightarrow{a_i a_{i+1}}$.

Доказательство. Пусть от противного цепочка $a_i a_{i+1}$ не является маргинальной син-

тагмой, т.е. в предложении π имеется слово c, которое служит определяемым членом синтагм $a_{i+1}c$ или ca_{i+1}. Тогда в схеме R грамматики F должны существовать правила, обеспечивающие вывод цепочки $c'\ldots a'_{i+1}a'_{i+2}$ из цепочки $a'_{i+1}a'_{i+2}$, что противоречит определению синтагмы предложения π и его проективности. Свойство 3° доказано.

Аналогично свойству 3° доказываются свойства 4° – 10°.

4°. *Синтагматическая структура* $\overrightarrow{\overrightarrow{a_i a_{i+1}}a_{i+2}}$ *предложения* π *содержит маргинальную синтагму* $\overrightarrow{a_{i+1}a_{i+2}}$.

5°. *Синтагматическая структура* $\overrightarrow{a_i\ldots\overrightarrow{a_j a_{j+1}}\,a_{j+2}}$ *предложения* π *содержит маргинальную синтагму* $\overrightarrow{a_j a_{j+1}}$.

6°. *Синтагматическая структура* $\overrightarrow{\overrightarrow{a_i a_{i+1}}a_{i+2}\ldots a_j}$ *предложения* π *содержит маргинальную синтагму* $\overrightarrow{a_{i+1}a_{i+2}}$.

7°. *Синтагматическая структура* $\overrightarrow{\overrightarrow{a_i a_{i+1}}\overrightarrow{a_{i+2}a_{i+3}}}$ *предложения* π *содержит маргинальные синтагмы* $\overrightarrow{a_i a_{i+1}}$ *и* $\overrightarrow{a_{i+2}a_{i+3}}$.

8°. *Синтагматическая структура* $\overrightarrow{a_i\ldots\overrightarrow{a_j a_{j+1}}\overrightarrow{a_{j+2}a_{j+3}}}$ *предложения* π *содержит маргинальные синтагмы* $\overrightarrow{a_j a_{j+1}}$ *и* $\overrightarrow{a_{j+2}a_{j+3}}$.

9°. *Синтагматическая структура* $\overrightarrow{\overrightarrow{a_i a_{i+1}}\overrightarrow{a_{i+2}a_{i+3}}\ldots a_j}$ *предложения* π *содержит маргинальные синтагмы* $\overrightarrow{a_i a_{i+1}}$ *и* $\overrightarrow{a_{i+2}a_{i+3}}$.

10°. *Синтагматическая структура* $\overrightarrow{a_i\ldots\overrightarrow{a_j a_{j+1}}\overrightarrow{a_{j+2}a_{j+3}}\ldots a_k}$ *предложения* π *содержит маргинальные синтагмы* $\overrightarrow{a_j a_{j+1}}$ *и* $\overrightarrow{a_{j+2}a_{j+3}}$.

Свойства 1°– 10° позволяют найти маргинальные синтагмы в предложениях анализируемого текста.

1.5. Алгоритм анализа проективных предложений

На входе алгоритма – цепочка $\delta \in V^+$, на выходе – синтаксическое дерево $D_\delta = (V_\delta, E_\delta)$ цепочки δ. Алгоритм работает следующим образом.

В цепочке δ ищутся маргинальные синтагмы типа $\overrightarrow{a_1a_2}$ и $\overrightarrow{a_{n-1}a_n}$ и исключаются из δ их маргинальные слова. Процесс поиска таких синтагм и исключения маргинальных слов повторяется до тех пор, пока синтагмы указанного типа будут присутствовать в цепочке δ. Далее аналогичная процедура повторяется для синтагматических структур типа $\overrightarrow{\overrightarrow{a_i a_{i+1}}a_{i+2}}$ и $\overrightarrow{\overrightarrow{a_i a_{i+1}}a_{i+2}}$, затем для структур $\overrightarrow{a_i\ldots\overrightarrow{a_j a_{j+1}}\,a_{j+2}}$, $\overrightarrow{a_i\overrightarrow{a_{i+1}a_{i+2}}\ldots a_j}$ и т.д. Алгоритм анализа проективных предложений включает следующие шаги.

1. $V_\delta := \varnothing$, $E_\delta := \varnothing$.

2. Искать в цепочке δ маргинальную синтагму, удовлетворяющую следствию из свойства 1°. Если такая синтагма найдена, то перейти к п.3, иначе – к п.5.

3. Найти объединение синтаксического дерева найденной синтагмы и синтаксического дерева D_δ, считая полученное объединение деревом D_δ.

4. Исключить из цепочки δ маргинальное слово найденной синтагмы. Если в полученной в результате цепочке δ оказалось более одного слова, то перейти к п.2, иначе – КОНЕЦ (синтаксическое дерево входного предложения δ построено).

5. Искать в цепочке δ маргинальную синтагму, удовлетворяющую следствию из свойства 2°. Если такая синтагма найдена, то перейти к п.3, иначе – к п.6.

6. Последовательно искать в цепочке δ маргинальную синтагму, удовлетворяющую свойствам 3°, 4° и т.д. Если маргинальная синтагма найдена, то перейти к п.3, иначе – к п.7.

7. КОНЕЦ (цепочка δ не является проективной).

2. Анализ непроективных предложений

Поскольку любое непроективное предложение является конкатенацией некоторого количества проективных предложений (в предельном случае ими являются все слова непроективного предложения), то к последним может быть применен алгоритм синтаксического анализа проективных предложений. В результате получим совокупность ордеревьев, являющихся синтаксическими деревьями проективных подцепочек анализируемой цепочки. Далее для каждой пары ордеревьев следует проверить существование дуги, соответствующей синтагме, которая не выявлена в процессе работы алгоритма анализа проективных предложений. Определяемым членом такой синтагмы может, очевидно, быть любая вершина одного из ордеревьев указанной пары, а определяющим – корень другого ордерева.

2.1. Алгоритм анализа непроективных предложений

На входе алгоритма – цепочка $\delta \in V^+$ и орграф D_δ в виде совокупности синтаксических деревьев проективных подцепочек цепочки, на выходе – синтаксическое дерево цепочки δ. Алгоритм состоит из следующих шагов.

1. Из совокупности D_δ всех синтаксических деревьев проективных подцепочек цепочки δ составить множество $\mathcal{D}$ всевозможных упорядоченных ордеревьев вида (D', D'').

2. Взять из множества $\mathcal{D}$ произвольную пару (D', D'') ордеревьев. $\mathcal{D} := \mathcal{D} \setminus \{(D', D'')\}$.

3. Искать синтагму, определяющим членом которой является произвольная вершина ордерева D', а определяемым – корень ордерева D''. Если такая синтагма найдена, то перейти к п.4, иначе – к п.5.

4. Исключить из множества $\mathcal{D}$ все упорядоченные пары ордеревьев, вторым элементом которых является ордерево D''. Дополнить орграф D_δ дугой, соответствующей найденной в п.3 синтагме. Если $\mathcal{D} = \varnothing$, то перейти к п.6, иначе – к п.2.

5. Если $\mathcal{D} = \varnothing$, то перейти к п.6, иначе – к п.2.

6. Если орграф D_δ связный, то КОНЕЦ (синтаксическое дерево предложения δ построено), иначе – КОНЕЦ (цепочка δ не является проективной и непроективной).

3. Учет неграмматичности при синтаксическом анализе предложений

Неграмматичные, или некорректные, предложения могут включать фразы с грамматическими ошибками, а также цепочки, корректные с общеграмматических позиций, но неприемлемые в рамках конкретной информационной системы. Различают неграмматичности на лексическом уровне, на уровне предложения, а также диалога пользователя с системой [5].

На уровне лексики неграмматичность проявляется в связи с отсутствием некоторых слов в базе знаний, орфографическими ошибками в словах и неверной сегментацией предложения (например, отсутствием пробелов между словами).

Неграмматичность на уровне предложения возникает из-за пропущенных или случайных слов в нем, нарушения порядка их следования, отсутствия согласования между словами и синтаксической омонимии.

На уровне диалога основным видом неграмматичности является эллипсис: для обеспечения краткости в предложениях могут быть пропущены отдельные слова и целые фразы.

Определим формально понятие неграмматичности.

Предложение некоторого языка назовем *неграмматичным*, если выполняется хотя бы одно из следующих двух условий:

1) оно не является проективным и непроективным и получено из проективного или непроективного предложения удалением определяемых членов одной или более его синтагматических структур;

2) для него существует более одного синтаксического дерева (такое предложение назовем *синтаксическим омонимом*).

3.1. Восстановление некорректных синтагм

Пусть имеется множество текстов K_i $(i \geq 1)$ на входном языке, которые будем называть *тематическими корпусами текстов* (каждый корпус K_i соответствует i-й предметной области). Основное требование к корпусу текстов – достаточный объем, позволяющий получить достоверные статистические характеристики его синтагм. Полагаем также, что для каждого корпуса K_i существует *модифицированный корпус текстов* M_i, в котором вместо предложений представлены их синтаксические деревья. На практике некоторым предложениям корпуса K_i может соответствовать более одного синтаксического дерева (в случаях, когда не все синтаксические связи удалось распознать в процессе анализа). В общем случае считаем, что элементами множества M_i являются множества синтаксических деревьев, такие, что для каждого $i = 1, 2, \ldots$ существует биективное отображение $\Theta_i : M_i \to K_i$.

Всякое нераспознанное в процессе синтаксического анализа слово анализируемого предложения исключается из текста. Процесс восстановления исключенных слов, а также определяемых членов синтагм при эллипсисе сводится к их поиску в модифицированном корпусе текстов, соответствующем тематике входного текста.

3.2. Распознавание синтаксической омонимии

Задача состоит в выборе синтаксической структуры предложения, адекватной контекстному окружению. Процедура распознавания наличия синтаксической омонимии во входном предложении может быть реализована на основе следующего необходимого условия.

Утверждение 5. *Если проективное предложение входного языка является синтаксическим омонимом, то существует синтагматическая структура (подцепочка этого предложения), после исключения определяемого члена которой полученное предложение будет проективным.*

Доказательство. Пусть от противного после удаления определяемого члена любой синтагматической структуры входного предложения полученная в результате цепочка не является проективной. Тогда во входном предложении отсутствует синтагматическая структура, определяющий член которой может быть определяющим членом некоторой другой синтагматической структуры этого предложения. Отсюда следует, что для данного входного предложения существует единственное синтаксическое дерево. Полученное противоречие доказывает утверждение 5.

В силу утверждения 5 рабочим критерием наличия омонимии во входном предложении может служить появление в нем новой синтаксической связи после удаления из предложения определяемого члена некоторой синтагматической структуры. После восстановления предложения, т.е. включения в него удаленной цепочки и замены прежней связи на новую, в результате синтаксического анализа должно быть получено другое синтаксическое дерево, отличное от прежнего. Для выбора адекватной связи будем использовать контекст анализируемого предложения и модифицированный корпус текстов.

Заключение

Предложенный в статье подход к математическому моделированию процессов синтаксического анализа текста может быть использован на предварительном этапе интеллектуальной обработки полнотекстовых документов на различных естественных языках. Такая универсальность метода анализа достигается благодаря независимости его алгоритмов от входного языка. Адаптация алгоритмов и программ, реализующих рассмотренную математическую модель, к каждому конкретному языку осуществляется путем применения в информационно-

аналитической системе специальных словарей синтагм и устойчивых словосочетаний. Процесс построения этих словарей может быть в значительной степени автоматизирован путем статистической обработки корпусов текстов большого объема (порядка 100 млн. словоупотреблений).

Список литературы

1. Липницкий С.Ф. Математическая модель распознавания синтаксической структуры предложений при обработке текстовой информации // Доклады НАН Беларуси. – 2002. – Т. 46. – № 1. – С.60–63.

2. Липницкий С.Ф., Ярмош Н.А. Моделирование интеллектуальных процессов в инженерных информационных системах. –Мн.: Беларуская навука, 1996.–222 с.

3. Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука, 1985. – 144 с.

4. Реформатский А.А. Введение в языковедение: Учебник для вузов. 5-е изд., уточн. / Под ред. В.А. Виноградова. –М.: Аспект Пресс, 2002. –536 с.

5. Карбонелл Дж., Хейз Ф. Стратегии преодоления коммуникативных неудач при анализе неграмматичных языковых выражений // Новое в зарубежной лингвистике. Вып. XXIV. Компьютерная лингвистика. – М.: Прогресс, 1989. – С. 48 – 105.

Поступила 12.01.04

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

S. F. Lipnitski

THE MATHEMATICAL MODEL OF SYNTACTIC ANALYSIS OF THE TEXT IN INFORMATIONAL ANALYTICAL SYSTEM

A mathematical model of semantic relationships in the informational analytical system is proposed. The algorithms of the grammar, semantic and intelligent analysis of the text information can be realized on a basis of this model. A recognition method of syntactic structure of natural language sentences is suggested. The method is based on mathematical simulation of semantic relations and their properties.