УДК 004.052.2

ЛИНГВИСТИЧЕСКИЕ АСПЕКТЫ ИНФОРМАТИКИ

Е.Б. Козеренко

В данной работе информатика рассматривается как наука о языковых преобразованиях. Суть этих преобразований заключается в моделировании переходов от сложной многокомпонентной и многоуровневой системы естественного языка к языку двоичных кодов (анализ) и в обратном направлении — от языков машинного уровня — к языкам высокого уровня и естественному языку (синтез). При рассмотрении лингвистических аспектов информатики важно осознавать, что *язык*, понимаемый как некоторое лингвистическое устройство (естественное или искусственное), выступает одновременно как инструмент и объект исследования и разработки. Во всей полноте проблема моделирования языковых преобразований встаёт при создании систем машинного перевода и извлечения знаний из естественно-языкового текста. При проектировании лингвистических процессоров используются системы логических правил различных грамматик и вероятностные механизмы.

1. Введение

Современный этап развития цивилизации многими исследователями сегодня вполне обоснованно квалифицируется как этап перехода к информационному обществу, основанному на знаниях. Поскольку основная часть знаний, которыми сегодня располагает человечество, представлена в виде текстов на естественных языках, то лингвистические аспекты информатики, как фундаментальной науки и важной прикладной сферы деятельности миллионов людей на нашей планете, становятся всё более важной и актуальной проблемой дальнейшего развития человеческого общества.

Информационные процессы в различных аспектах являются основным предметом исследования информатики. Суть информационных процессов в компьютерных системах заключается в «переводах» с языков высокого уровня на язык вычислительной машины. Поскольку исходной информационно-моделирующей системой для человека является естественный язык, то искусственные языки — в данном случае языки программирования — прошли эволюционный путь от машинных кодов до эргономичных языковых систем, которые на каждом этапе развития приобретали всё больше черт, свойственных естественному языку. На современном этапе «языковая компетенция» компьютерных систем достигла такого уровня, что стало возможным обеспечивать обработку естественно-языковых текстов и перевод с одного естественного языка на другой на промышленном уровне. Безусловно, текст, порождаемый компьютерной системой, ещё далёк от совершенства и требует постредактирования, однако, современные системы машинного перевода, в среднем, увеличивают скорость перевода в 2,5 раза.

Естественный язык — это универсальная моделирующая система. Человеческий интеллект включает «лингвистический процессор», который, в основных чертах, может служить *прототипом* для искусственного языкового процессора. Однако, основные трудности при этом заключаются в том, что процессы обработки языка «естественным» устройством (человеческим интеллектом) и «артефактом», т.е. искусственно созданным языковым процессором (двоичная логика, комбинации 0 и 1, сдвиги, присваивания, пересылки данных) существенным образом отличаются.

При рассмотрении лингвистических аспектов информатики важно осознавать, что *язык*, понимаемый как некоторое лингвистическое устройство (естественное или искусственное), выступает одновременно как инструмент и объект исследования и разработки.

Рассмотрим общие черты устройства языка. Очевидно, что его основными необходимыми компонентами являются *лексикон* и правила образования структур (*грамматика*). Лексикон естественного языка — это когнитивная модель мира в сознании человека, отражённая в системе языка. При этом очевидно, что «членение реальной действительности» в языке имеет чёткую структуру, характерную для большинства естественных языков. Основным инструментом при этом является *категоризация*.

Верхним уровнем системы категорий являются классы слов, ассоциируемые со следующими грамматическими категориями («нетерминальными символами», «частями речи»):

- открытые классы:
  - объекты (сущности);
  - действия (предикаты);
  - модификаторы:
    - объектов — адъективные единицы и группы (в том числе меры и степени), комплементы (дополнения),
    - действий — адвербиальные единицы (в том числе меры и степени), локативы;
- закрытые классы:
  - коннекторы:
    - сочинительные (союзы),
    - подчинительные (предлоги);
  - ссылки (местоимения);
  - кванторы;
  - отрицательные частицы;
  - служебные элементы.

Такая система категорий служит основой когнитивной модели мира и её отражения в языке в различных тезаурусах, онтологиях и идеографических словарях, которые, в свою очередь, принимаются за основу когнитивных моделей интеллектуальных систем и баз знаний. Самой сложной задачей при этом является обеспечение семантической эквивалентности и адекватности перевода с одного естественного языка на другой в связи с неоднородным членением реальной действительности разными языками.

Обобщённое определение грамматики, которое справедливо для широкого класса языков (включая естественный), — это определение порождающей контекстно-свободной грамматики.

> Контекстно-свободная грамматика — $G = (N, T, P, S)$, где N — это множество нетерминальных символов, T — множество терминальных символов, P — множество продукций вида $A \to b$, где A — это нетерминальный символ, b — это цепочка символов, S — специальный исходный символ.

Грамматика вычислительной системы должна задавать правила организации последовательностей категориальных единиц: исходное категориальное значение, окказиональные категориальное значение (роли). При этом должны быть отражены такие универсальные механизмы естественного языка как совпадение и наложение значений. Механизм «наложения» ролей заключается в том, что исходная категория единицы не исчезает, а уходит на второй (третий) план — именно её присутствие и даёт тот семантический эффект, который позволяет выразить нужное значение в данном контексте.

Эти сложные особенности естественного языка не могут быть отражены средствами только контекстно-свободной порождающей грамматики. Для этого в настоящее время применяются и развиваются такие формализмы как унификационные и категориальные грамматики. Включение вероятностных расширений в механизмы грамматик и в лексикон позволяют создать интегральный семантико-вероятностный «портрет» языка.

2. Моделирование лингвистических объектов в унификационных и категориальных грамматиках

Методы моделирования лингвистической реальности в унификационных и категориальных грамматиках отражают сложный характер естественных языков и включают возможности уже существующих подходов грамматик составляющих и грамматик зависимостей. Унификационные грамматики основаны на конструктивной идее представления лингвистических сущностей в виде объектов, свойства которых задаются как наборы атрибутов и соответствующих им значений. Значения атрибутов выступают в качестве ограничений на сочетаемость объектов, например, фраза «белые облако» будет распознана как недопустимая, поскольку значения атрибута «число» будет различаться у прилагательного и существительного. Таким образом, ограничения являются очень важным изобразительным средством при построении правил унификационных грамматик, которые имеют также ещё одно название «формализмы, основанные на ограничениях» (constraint-based formalisms).

По сравнению с контекстно-свободными грамматиками и грамматиками зависимостей унификационные грамматики обеспечивают более тонкие и дифференцированные механизмы представления лингвистической информации. Для чего же нужны детальные системы ограничений в представлениях грамматических категорий? Прежде всего, они позволяют реализовать отношения согласования (и исключить случаи нарушения согласования).

Системы ограничений позволяют задавать субкатегориальные признаки (т.е. ожидаемые категории «подчинённых» объектов). Субкатегориальные признаки необходимы для тех лингвистических объектов, которые «держат» структуру фразы, например, матричных глаголов. Унификационные грамматики дают возможность моделировать гораздо более сложные языковые явления, чем контекстно-свободные грамматики, и, наконец, дают возможность эффективно и удобно вычислять семантику синтаксических конструкций. В целом все эти преимущества помогают избежать или, по крайней мере, снизить перепроизводство гипотез, порождаемых системой в процессе грамматического разбора естественно-языковых высказываний.

Основной операцией, которая используется в унификационных грамматиках, является проверка равенства значений. При успешном завершении проверки осуществляется унификация объектов, отсюда и название грамматики. Разработана целая система правил унификации (см. ниже), которые при успешном завершении проверки значений атрибутов могут порождать новые объекты. Если условия унификации не выполняются, то унификация не производится.

Таким образом, базовыми инструментами унификационнных грамматик являются *структуры атрибутов*, ассоциируемые с тем или иным лингвистическим объектом, и *операция унификации*, которая осуществляет комплекс основных операций над структурами атрибутов и значений. Важными расширениями унификационных грамматик являются механизмы *типизации* и *наследования*.

В лингвистической теории система представлений на основе атрибутов и значений впервые была применена в фонологии, затем в семантике и в синтаксисе. В работе Н. Хомского [1] развиваются идеи построения моделей синтаксиса с использованием структур атрибутов и значений. Операция унификации применительно к лингвистическим объектам разрабатывалась для структур атрибутов [2] и термов — унификация термов [3]. Эти исследования и разработки проводились в рамках создания систем машинного перевода. Вехами в развитии систем лингвистических представлений на основе унификации были грамматики, известные как «Расширенные сети переходов» (Augmented Transition Networks, ATN) [4, 5], Q-системы [6]. Использование механизма наследования и условий соответствия для лингвистических знаний были впервые предложены Бобровым и Вебером [7].

Наиболее известные грамматические теории, построенные на основе унификации, — это лексико-функциональная грамматика (Lexical Functional Grammar — LFG [8]), вершинная грамматика составляющих, или, точнее, «грамматика фразовых структур, управляемых головной вершиной» (Head-Driven Phrase Structure Grammar — HPSG [9]), грамматика конструкций (Construction Grammar [10]), унификационная категориальная грамматика (Unification Categorial Grammar [11]).

Категориальные грамматики основаны на математическом исчислении, моделирующем распознавание цепочек слов в качестве правильно построенных предложений некоторого языка. Основополагающими в теории категориальных грамматик, которые вначале предназначались для анализа математического языка, были работы К. Адъюкевича [12] и И. Ламбека [13], затем И. Бар-Хиллел [14] показал применимость механизма категориальной грамматики к предложениям естественного языка. На современном этапе категориальная грамматика получила развитие в работах М. Муртгата [15], М. Стидмэна [16, 17], аппарат категориальной грамматики, включая операцию аппликации, используется в семиотической универсальной грамматике С.К. Шаумяна [18, 19].

2.1. Контекстно-свободная порождающая грамматика и унификационная грамматика. Правила контекстно-свободной грамматики (грамматики составляющих) основаны на простой идее конкатенации, то есть задаётся порядок следования составляющих, например, если за именной группой (NP) следует глагольная группа (VP), то получается предложение (S), и далее каждая из составляющих может быть разложена на группы (нетерминальные символы) и элементы (терминальные символы) в соответствии с правилами заданной грамматики:

$$
\begin{aligned}
\mathrm{G} &= \{\mathrm{T}, \mathrm{N}, \mathrm{P}, \mathrm{S}\},\\
\mathrm{S} &\to \mathrm{NPVP},\\
\mathrm{NP} &\to \mathrm{N} \mid \mathrm{DetN} \mid \mathrm{AdjN} \mid :::,\\
\mathrm{VP} &\to \mathrm{V} \mid \mathrm{VNP} \mid \mathrm{VPP} \mid :::,\\
\mathrm{PP} &\to \mathrm{PrepNP} \mid :::,
\end{aligned}
$$

Однако, для того, чтобы результирующие составляющие были грамматичны, т.е. чтобы исключить предложения вроде «перелётные птицы улетел», необходимо ввести дополнительные проверочные условия (механизм ограничений), которые будут проверять совпадение значений атрибута «число» у предполагаемого субъекта (в данном случае именной группы) и предиката предложения (глагольной группы). Именно этот механизм и обеспечивается унификационной грамматикой через структуры атрибутов и значений. Таким образом, в правила контекстно-свободной грамматики встраиваются расширения в виде структур атрибутов и значений. Эти структуры также имеют название «матрицы атрибутов–значений» (Attribute-Value Matrices — AVM).

Рассмотрим представление простого предложения английского языка на основе унификационно-порождающих правил (т.е. грамматики, стержнем которой является грамматика составляющих, правила которой расширены необходимыми структурами атрибутов и значений).

John runs.

$$\begin{bmatrix}\text{CAT} & \text{NP}\\ \text{Number} & \text{Sg}\\ \text{Person} & 3\end{bmatrix} \text{ унифицируется с } \begin{bmatrix}\text{CAT} & \text{VP}\\ \text{Number} & \text{Sg}\\ \text{Person} & 3\end{bmatrix} = \begin{bmatrix}\text{CAT} & \text{S}\\ \text{Number} & \text{Sg}\\ \text{Person} & 3\end{bmatrix}$$

Здесь производится унификация именной (NP) и глагольной (VP) составляющих простого предложения, при этом проверяются значения атрибутов «число» (Number) и «лицо» (Person) обеих составляющих. В данном случае они совпали, поэтому унификация успешно завершается, порождая новую вершину, — предложение (S), которая наследует атрибуты «число» — «единственное», «лицо» — «третье». Этот простой пример иллюстрирует суть унификационной грамматики, для более сложных случаев необходимы более детальные и тонкие системы представлений атрибутов и значений. Эта запись — некоторый наглядный пример, использующий как традиционную нотацию грамматик составляющих, так и элементы унификационной грамматики — матрицы атрибутов.

В отличие от традиционной порождающей грамматики составляющих, в которой именная группа обозначается как NP, глагольная группа — VP, в унификационной грамматике используются свои обозначения, в частности, категориальные значения также имеют вид матриц атрибутов-значений. Вот, например, как

представляется именная группа в наиболее известных унификационно-грамматических системах [20–22]:

$$\text{NP} \rightarrow [\text{N0}] \mid [\text{N1}] \mid [\text{N2}].$$

Это представление обозначает следующее: [N0] — атомарная именная группа, например, имя существительное или имя существительное, которому предшествует определитель, например, артикль; [N1] — именная группа с одним распространением (например, с модификатором); [N2] — именная группа с более чем одним распространением, возможно, с комплементом.

Рассмотрим матрицу атрибутов-значений одного из подвидов именной группы

$$[\text{N0}] \leftarrow \{[\text{N}+], [\text{V}-], [\text{BAR0}]\}.$$

Данная запись обозначает следующее: [N+] наличие атрибута «номинативность», [V−] — отсутствие признака «глагольность», [BAR0] — ранг группы (атомарный).

Итак, ключевым аспектом унификационной грамматики являются системы представления лингвистической информации в виде структур атрибутов и значений.

2.2. Структуры атрибутов и значений. Рассмотрим, каким образом лингвистические объекты представляются в виде структур атрибутов и значений. Прежде всего, значениями атрибутов могут быть простые (атомарные значения) или структуры атрибутов-значений.

Примеры атрибутов с атомарным значением: [Num: Sg], [Pers: 3-d]. Атрибут, значением которого является структура атрибутов:

$$\text{Agreement1} = \{[\text{Num: Sg}], [\text{Pers: 3-d}]\}.$$

Особую роль играют значения, являющиеся грамматическими категориями (категориальные значения). При этом важно отметить, что конструирование категорий — это сфера инженерно-лингвистического творчества, до некоторой степени, эти конструкции базируются на традиционных частях речи, но, как мы видели в предыдущих примерах, имеют также свои отличия.

Каждый языковой объект (атомарный или структурный) может мыслиться как некоторый узел дерева разбора на основе унификациннно-порождающей грамматики. Эта грамматика

управляет процессом формирования структур атрибутов более крупных составляющих из структур атрибутов их составных частей. Категориальные значения и другие атрибуты синтактико-семантических объектов играют роль условий, определяющих совместимость указанных частей грамматических конструкций. Приведём примеры матриц атрибутов и значений:

$$\begin{bmatrix} \text{CAT} & \text{Noun} \\ \text{Number} & \text{Sg} \end{bmatrix}$$

— пример матрицы атрибутов и значений для именного узла;

$$\begin{bmatrix} \text{CAT} & \text{Verb} \\ \text{Number} & \text{Sg} \\ \text{Pers} & 3 \\ \text{Subcat} & \text{NP} \end{bmatrix}$$

— пример матрицы атрибутов и значений для глагольного узла;

$$\begin{bmatrix} \text{Agreement1} & \begin{bmatrix} \text{Number} & \text{Sg} \\ \text{Person} & 3 \end{bmatrix} \end{bmatrix}$$

— пример матрицы, задающей структуру атрибутов в качестве значения атрибута Agreement1;

$$\begin{bmatrix} \text{CAT} & \text{Verb} \\ \text{Agreement1} & \end{bmatrix}$$

— пример матрицы, где структура Agreement1 выступает как атрибут некоторого узла с категориальным значением Verb.

Если мы рассматриваем некоторый структурный объект (иначе говоря, узел дерева разбора), например, именную, или глагольную, или адъективную, группу, то одна из вершин этого объекта будет *головной*. Понятие *головной вершины* имеет большое конструктивное значение в унификационных грамматиках, поскольку *атрибуты головной вершины* передаются вверх по дереву материнскому узлу, т. е. наследование происходит от «дочери» (daughter node) к «матери» (mother node). Внутри дочернего узла головная вершина «вбирает» в себя все нужные атрибуты других компонентов этого узла, а правила грамматики задают, какие атрибуты должны передаваться.

Эффективным инструментом задания структур атрибутов и значений являются сложные категории («псевдонимы» — Aliases). Эти «псевдонимы» позволяют конструировать необходимые сочетания признаков (атрибутов и значений), которые должны совместно встречаться у определённых объектов, этот инструмент обеспечивает более компактную запись правил грамматики.

Рассмотрим некоторые примеры «псевдонимов»:

$$\text{ALIAS V0} = [\text{V}+, \text{N}-, \text{BAR 0}]$$

— «псевдоним», задающий сочетание признаков единичного глагола;

$$\text{ALIAS N0} = [\text{V}-, \text{N}+, \text{BAR 0}]$$

— «псевдоним», задающий сочетание признаков единичного существительного;

$$\text{ALIAS N2} = [\text{V}-, \text{N}+, \text{BAR 2}]$$

— «псевдоним», задающий сочетание признаков именной группы, например, *white house press secretary*.

При составлении структур атрибутов для глаголов и других типов лексических единиц, у которых могут быть зависимые объекты (иначе говоря, «модель управления», «валентности», «актанты», «аргументы», «распространение»), задаются так называемые «субкатегориальные признаки» (SUBCAT). Субкатегориальные признаки — это ожидаемые категории «подчинённых» объектов, которые должны появиться поблизости от главного слова.

Рассмотрим пример задания субкатегориального фрейма для глагольной фразы *leaves London in the evening*, где глагол имеет два аргумента:

$$\begin{aligned} &\text{Verb} \rightarrow \textit{leaves} \\ &\langle \textit{Verb} \text{ HEAD AGREEMENT NUMBER} \rangle = \text{SG}, \\ &\langle \textit{Verb} \text{ HEAD SUBCAT FIRST CAT} \rangle = \text{NP}, \\ &\langle \textit{Verb} \text{ HEAD SUBCAT SECOND CAT} \rangle = \text{PP}, \\ &\langle \textit{Verb} \text{ HEAD SUBCAT THIRD CAT} \rangle = \text{END}. \end{aligned}$$

Это довольно простой пример, в реальности у глаголов могут быть довольно сложные субкатегориальные фреймы, которые могут состоять из многих различных типов фраз.

Приведём пример унификации более сложной структуры с субкатегориальными атрибутами (пример из Британского национального корпуса):

V=pp_inf:

V1 → H0[VSUBCATPP_VPINF] P2 V1[FIN–, VFORM INF]

pleads with the administration to give no far support.

2.3. Категориальные грамматики. Совершенно особый вид грамматики, отличный от всех генеративистских типов грамматик, — это категориальные грамматики. Категориальная грамматика включает два компонента: *категориальный лексикон*, который ассоциирует каждое слово с некоторой синтаксической и семантической категорией и *комбинаторные правила*, указывающие, каким образом должны объединяться функции и аргументы.

Предусмотрены два вида категорий: функторы (или функции) и аргументы. У аргументов, как, например, у существительных, — простая категория — N. Глаголы или определители (артикли, указательные местоимения и т.п.) действуют как функторы. Например, определитель может рассматриваться как функтор, который применяется к N, стоящему справа от него, и образует именную группу NP. Такие сложные категории строятся с использованием *операторов* X=Y и X\Y. Оператор X=Y означает функцию из Y в X, т.е. нечто объединяется с Y, стоящим справа от него, и производит X. Таким образом, определитель получает категорию NP=N: что-то, что объединяется с N, стоящим справа от него, и производит NP.

Оператор X\Y означает функцию из Y в X, когда нечто объединяется с Y, стоящим слева от него и производит X. Например, глагол и имя, стоящее слева от глагола, будут порождать предложение (S). Таким образом, в категориальных грамматиках глагол получает следующее обозначение: S\NP (т.е. нечто, слева от чего стоит именная группа, и их объединение порождает предложение).

Переходные глаголы могут иметь категорию (S\NP)=NP, т.е. нечто, что объединяется с NP, стоящим справа от него, и порождает S\NP.

Таким образом, привычные категории из других грамматик получают иную систему обозначений в категориальной грамматике:

$$\begin{aligned} \text{Det} &\rightarrow \text{NP=N}, \\ \text{VP} &\rightarrow \text{S\NP}, \\ \text{Vtrans} &\rightarrow \text{(S\NP)=NP}. \end{aligned}$$

Рассмотрим пример порождения предложения в категориальной грамматике [16]:

Harry	*eats*	*apples*
NP	(S\NP)=NP	NP
NP	S\NP	
S		

При порождении вначале срабатывает оператор, который стоит снаружи скобок, порождая функцию S\NP из (S\NP)=NP и NP, затем срабатывает оператор, который стоит внутри скобок, порождая предложение S из NP и S\NP.

Итак, техника категориальных грамматик даёт лаконичный логический аппарат представления языковых структур.

2.4. Существующие аппараты грамматик фразовых структур и унификации. Дадим краткий обзор основных существующих аппаратов унификационных грамматик: это вершинная грамматика — грамматика фразовых структур, управляемых головной вершиной — HPSG [22], обобщённая грамматика фразовых структур — GPSG [21], пересмотренная обобщённая грамматика фразовых структур — RGPSG [23]. Нами также рассматривались аппараты категориальной грамматики и грамматики зависимостей [24].

Механизм категоризации, т.е. порождение основной грамматики атомарных категорий из отчётливых наборов атрибутных связываний был впервые предложен в [25]. Он был использован при создании варианта GPSG — грамматики Alvey [20], в которой были произведены сокращения количества категорий за счёт включения каждой категории в более общую.

Важным принципом грамматики HPSG является строгий лексикализм, утверждающий, что структура слова и структура фразы управляются независимыми механизмами. Роли определяются глагольными валентностям; высказывания — это сплав категориальных значений, ролевых значений и их структурных проекций, особых для каждого конкретного языка. Эта точка зрения на устройство грамматики совпадает с нашими выводами, к которым мы пришли эмпирическим путём в результате работы над рядом проектов по созданию систем извлечения знаний из текстов (ДИЕС, ЛОГОС-Д, Аналитик, ИКС) и машинному переводу [26].

Основным формальным аппаратом для задания грамматики русского языка традиционно является грамматика зависимостей. Считается, что она в большей степени подходит для языка со свободным порядком слов, чем грамматика фразовых структур. Однако, грамматика зависимостей предлагает достаточно сложный механизм линеаризации языковых структур при порождении естественно-языкового текста из внутренних предикатно-актантных представлений. Поэтому к нашим задачам моделирования русско-английского трансфера мы применили подход, сочетающий аппараты грамматик составляющих и грамматик зависимостей для русско-английского машинного перевода.

В GPSG существуют три вида атрибутов с категориальным значением: SLASH, который отмечает путь между незаполненной валентностью и её заполнителем, передавая вниз значение категории ожидаемой единицы, AGR, отмечающий путь между аргументом и функтором, который с ним синтаксически согласуется (например, между субъектом и матричным глаголом), и WH, который отмечает путь от wh-слова (т.е. вопросительного слова — *what*, *why*, *which*, или других подобных слов) к минимальной клаузе которая содержит его, задавая морфологический тип wh-слова.

Формат правил непосредственного доминирования/линейного предшествования GPSG моделирует параметры головной вершины и некоторые свойства свободного порядка слов.

Формальная грамматика HPSG основана на правилах фразовых структур (составляющих), но при этом отношения доминирования реализуются через головные элементы. Фразовые типы также рассматриваются в терминах множественных иерархий наследования, которые позволяют делать обобщения разнообразных типов конструкций.

3. Вероятностные расширения грамматик

В последние несколько лет значительное продвижение в области обработки естественного языка и машинного перевода было достигнуто за счёт применения различных стохастических моделей анализа языковых структур. Эти модели, в частности, успешно использовались в речевых системах перевода [30, 31]. Машинный перевод на основе статистики был впервые предложен в [32, 33].

Отправная точка для любой системы обработки естественного языка — проектирование модуля определения и разметки частей речи (mэrrepa). Различные статистические (вероятностные, стохастические) тэггеры появились в 1980-е годы. Общая идея всех стохастических тэггеров заключается в выборе наиболее вероятного тэга (т.е. частеречной метки) для данного слова. Чаще всего для вероятностных тэггеров используются Марковские модели. Так, например, для некоторого данного предложения или последовательности слов выбирается последовательность тэгов, которая максимизирует следующую формулу:

$$\mathrm{P}(\text{слово} \mid \text{тэг}) * \mathrm{P}(\text{тэг} \mid \text{предыдущие } n \text{ тэгов}).$$

Ещё один подход к машинному обучению, основанный на правилах и стохастическом тэггировании (разметке частей речи), известен как обучение, основанное на трансформациях (Transformation-Based Learning, TBL). TBL — это метод управляемого обучения с использованием некоторого заранее размеченного обучающего корпуса.

Для вероятностного грамматического разбора применяются стохастические грамматики.

— Вероятностная контекстно-свободная грамматика. Определение: $\mathrm{G} = (\mathrm{N}, \mathrm{T}, \mathrm{P}, \mathrm{S}, \mathrm{D})$, где N — это множество нетерминальных символов, T — множество терминальных символов, P — множество продукций вида $\mathrm{A} \to \mathrm{b}$, где A — это нетерминальный символ, b — это цепочка символов, S — специальный исходный симвло, D — это функция, приписывающая значения вероятности каждому правилу из множества P. Как получить необходимые данные для вероятностной контекстно-свободной грамматики? Один из путей — использование корпуса синтаксически размеченных предложений. Такой корпус называется банком синтаксических деревьев (treebank). Например, Penn Treebank [34] содержит деревья разбора для ряда текстовых корпусов (Brown Corpus, Switchboard corpus). Если задан банк деревьев разбора, то вероятность каждой развёртки некоторого нетерминального узла может быть вычислена путём подсчёта количества случаев, когда данная развёртка встречается, с последующей нормализацией:

$$P(\alpha \to \beta \mid \alpha) = \frac{\mathrm{Count}(\alpha \to \beta)}{\sum\limits_{\gamma} \mathrm{Count}(\alpha \to \gamma)} = \frac{\mathrm{Count}(\alpha \to \beta)}{\mathrm{Count}(\alpha)}.$$

— Вероятностная грамматика замещения деревьев. Её определение то же, что и для вероятностной контекстно-свободной грамматики, но здесь мы имеем дело не с правилами, а с фрагментами деревьев произвольной глубины, и значения вероятностей приписываются этим фрагментам, что делает эту грамматику стохастически более мощной, чем вероятностная контекстно-свободная грамматика.

Статистические методы обработки естественного языка расширяют схему основных существующих подходов к машинному переводу — прямого перевода, переноса (трансфера) и подхода на основе языка-посредника (интерлингвы) [35].

Грамматика, применяемая для нашей системы правил, — это вероятностная грамматика замещения функциональных деревьев, задающая правила многовариантного когнитивного переноса.

4. Функционально-семантический подход к реализации многоязычных процессоров

В настоящее время нами ведётся исследовательский проект по созданию многоязычной лингвистической базы знаний с функциями машинного перевода. Мы выбрали подход, заключающийся в разработке системы правил фразовых структур, отражающих также и отношения зависимости через механизм наследования атрибутов головной вершины. Этот подход более практичен с вычислительной точки зрения, и, по нашим данным, не применялся ранее для двуязычной ситуации.

Функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой. Множество языковых структур, представленных в виде синтактико-семантических комплексов, выстраиваются в *иерархию правил*. В нашем случае это разновидность унификационно-порождающей грамматики, в которой структуры атрибутов и значений и их преобразования задаются в виде контекстно-свободных и мягко контекстно-зависимых продукционных правил. Отношения зависимости реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов.

На текущем этапе система правил модифицируется с учётом возможной многозначности синтаксических структур, и разрабатываются механизмы разрешения неоднозначности посредством включения в систему правил статистической информации о возможных контекстах языковых структур. Поскольку естественный язык генерирует бесконечное число цепочек символов (словосочетаний, фраз, клауз, предложений), любая заранее заданная система правил оказывается неполной, и поэтому необходимы механизмы обучения для расширения и модификации правил. Данные, на которых базируется автоматический вывод правил, выявляются путём сопоставления параллельных текстов. Существуют чисто вычислительные методы нахождения соответствий, однако, они дают много шумов, для устранения которых всё равно требуется кропотливая лингвистическая экспертиза, но уже на этапе постобработки.

С нашей точки зрения, для того, чтобы избежать порождения избыточных правил, в систему необходимо заложить не только шаблоны языковых структур, по аналогии с которыми будут порождаться правила, но и принципы установления синонимичных средств языка, что позволяет использовать новый подход на основе *полей функционального переноса* [26]. В настоящий момент нами разрабатывается расширенная спецификация грамматики, в которой задаются категориально-функциональные признаки языковых объектов и структур и уточняются их вероятностные расширения. Функциональные значения языковых объектов кодируются в виде меток (тэгов) фразовых структур, типы атрибутов — значений определяются как категории, например:

[Category: VerbNounIng]
taking the risks, *using films*;

[Category: toPlusInfinitiveSubj]
He is appointed *to be a successor*;

[Category: toPlusInfinitiveObj]
We know *them to be reliable partners*.

В наших представлениях учитываются различные варианты перевода (трансфера) одной и той же структуры с английского языка на русский и обратного перевода. Например, возможны

следующие способы трансфера герундиальной фразы (Category: VerbNounIng) с английского языка на русский:

[Category: VerbNounIng] →
ИЛИ {[Category: VerbInf];
[Category: VerbNounIng];
[Category: Clause]}

Для любой схемы машинного перевода существуют такие серьёзные проблемы, как задание синтактико-семантических ожиданий для глаголов, обработка разорванных структур, адаптация фразовых комплексов и фразеологических единиц. В англо-русском трансфере эти проблемы усугубляются высокой продуктивностью английских фразовых глаголов (и других единиц) и их производных.

В нашем подходе используются как правила фразовых структур, так и словарная информация для этих случаев. При построении системы правил для нас являются важными также соображения практичности, быстроты реализации и, по возможности, низких вычислительных затрат. Разработанная нами система атрибутов и значений позволяет задавать детальную информацию о категориальных и субкатегориальных признаках.

Фактически, процесс трансфера с одного языка на другой происходит через функционально-категориальные значения языковых единиц. Языковые структуры, которые могут быть подвергнуты процедуре трансфера, должны обладать семантической полнотой, с точки зрения их функции. Так, случаи категориального сдвига, в частности, когда применяется приём конверсии, требуют специальной обработки: категориальный сдвиг синтаксической единицы определяется функциональной ролью этой единицы в предложении (например, имя существительное в функции модификатора может переводиться как прилагательное).

Таким образом, только создавая понятия-кентавры, такие как «составляющие–зависимость», «линейность–нелинейность», «форма–функция», и т.п. мы можем прийти к разумной и ясной картине лингвистической реальности [18].

Грамматика, применяемая для нашей системы правил, — это вероятностная грамматика замещения функциональных деревьев, задающая правила многовариантного когнитивного переноса.

Поскольку структуры естественного языка во многих случаях бывают неоднозначными или многозначными, это приводит к множественности возможных переводов с одного языка на другой. Основой системы логико-лингвистических правил, разрабатываемых в нашем проекте, являются *обобщённые когнитивные структуры*, извлекаемые из систем грамматических категорий русского и английского языков и функциональных ролей языковых единиц в предложении. Формальный аппарат представления англо-русских синтактико-семантических соответствий является вариантом вероятностной унификационной грамматики.

Для разрешения неоднозначности эта грамматика предлагает следующее решение: выбор наиболее вероятной интерпретации структуры в данном контексте. Такой подход был применён нами при разработке системы машинного перевода [26], он обеспечивает устойчивую работу лингвистического процессора и позволяет избежать избыточного количества правил разбора языковых структур.

5. Моделирование вариативности структур когнитивного переноса

Внутренние логико-семантические закономерности языкового строя и функционирования языка в разработанном нами формализме основаны на эвристических правилах различной степени детализации. При этом разрешение неоднозначности осуществляется на основе учёта весов структур (деревьев и поддеревьев разбора) в системе правил лингвистического процессора.

Функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой. Например, [Feature, EnumVerb]; [Category, bePlus]; [Category, toPlusInfinitive]; [Feature, verbModal]; [Feature, verbComplex]; и т.п.

Решения таких важных проблем, как разрешение референциальных ссылок, разделённые конструкции, также вырабатываются с помощью аппарата фразовых структур. Система правил, основанная на этом формальном аппарате, может быть названа Когнитивной Трансферной Грамматикой (КТГ). Состоит она из переносимых фразовых структур, а также правил перевода методом трансфера (переноса), которые включаются в один и тот

же блок. Такие блоки, или Структуры Когнитивного Переноса (СКП), являются составными компонентами декларативного модуля синтаксического процессора и задают, как правила линейного порядка, так и отношения зависимости в рамках одной и той же фразовой структуры.

Исходный вариант модуля СКП был реализован в экспериментальной системе машинного перевода и состоял из 222 правил разбора-перевода [26]. В настоящее время на основе функционально-семантического подхода в нашем исследовательском проекте разработана Многовариантная Когнитивная Трансферная Грамматика (МКТГ), учитывающая неоднозначность синтаксических структур и включающая свыше 300 структурно-семантических правил для англо-русского перевода.

Синтаксис МКТ-структуры (МКТС) может быть представлен следующим образом:

МКТС → *МКТС* ⟨*идентификатор*⟩ *МКТС* ⟨*вес*⟩

МКТС ⟨*метка*⟩ ⟨*Входная фразовая структура & набор атрибутов-значений*⟩ →

⟨*Схема трансфера*⟩ →

OR {⟨*Генерируемая фразовая структура & набор атрибутов-значений* 1⟩ ⟨*вес1*⟩ ⟨ *Генерируемая фразовая структура & набор атрибутов-значений* 2⟩ ⟨*вес2*⟩ … ⟨*Генерируемая фразовая структура & набор атрибутов-значений* N⟩ ⟨*вес*N⟩}.

МКТГ-правило — это контекстно-зависимая продукция, и деривационный процесс может определяться переходами И/ИЛИ. Причём эти два механизма вводят лексическую и структурную неоднозначность, что является центральным свойством естественных языков.

При нашем подходе прямое кодирование возможных атрибутов глагольных ожиданий тоже, в основном, осуществляется посредством структур когнитивного переноса. Поскольку фреймы глагольных ожиданий могут быть довольно сложными и состоять из множества фраз различных типов, мы вначале разработали список возможных фразовых типов, которые могут образовывать эти фреймы, например,

VPto «*I want to know*»;

VPing «*He contemplates using them*»;

Sto «*feel themselves to be completely happy*».

Каждый глагол может появляться в нескольких различных фреймах.

Основным отличием наших представлений от известных фреймовых представлений глаголов является *семиотический подход* [19, 26] к представлению функциональных значений глаголов и их реализации в различных контекстах.

Таким образом, МКТГ является функционально-семантическим вариантом вероятностной унификационной грамматики замещения деревьев (PTSG — Probabilistic Tree Substitution Grammar).

6. Заключение

Одной из актуальных и стратегически важных проблем развития современной информатики как фундаментальной науки является проблема моделирования преобразований текстов естественного языка, которая, в частности, возникает при разработке систем машинного перевода и извлечения знаний из текстов, представленных на естественных языках. Перспективное направление решения этой проблемы состоит в создании лингвистических процессоров, в которых используются системы логических правил различных грамматик и вероятностные механизмы. Выше были изложены некоторые результаты исследований данной проблемы, полученные в Институте проблем информатики РАН за последние годы. Эти результаты, по нашему мнению, являются определённым вкладом в развитие лингвистических основ информатики как фундаментальной науки об информации и информационных процессах в природе и обществе.

Система грамматических единиц, классов и категорий совместно с правилами их функционирования служит для передачи обобщённых структур ментального плана, лежащих в основе смысла высказываний и составляющих основу грамматического строя языка [26].

Как было показано в [28], способ кодирования в языке в значительной степени определяется глубинной семантической структурой, и значительное преимущество имеет такой способ представления, который в качестве исходной позиции принимает семантический уровень, и конкретным семантическим единицам сопоставляются выражающие их средства кодирования.

Подход функциональной семантики во многих аспектах согласуется с категориальной грамматикой: её суть сохраняется в новых фактах, квалифицируемых через традиционные категории [29].

Переносимость фразовых структур обусловлена выбором языковых единиц исходного и целевого языка, принадлежащих к одним и тем же полям функционального переноса (ПФП), не взирая на различия или сходство их традиционных категориальных значений [26]. Множество функциональных значений с их категориальными реализациями служат источником ограничений для механизма унификации в формальном представлении нашей грамматики. Разработанный нами формализм предназначен для грамматического разбора на основе атрибутов и значений, механизма наследования атрибутов головной вершины фразовых структур, которые выделяются по признаку функциональной идентичности в исходном и целевом языках.

Техника категоризации, используемая в унификационно-порождающих грамматиках, даёт очень мощный и гибкий инструмент инженерно-лингвистического моделирования для различных классов задач, в которых необходима глубокая проработка языковых объектов. Такие задачи решаются при создании систем машинного перевода, извлечения знаний из естественно-языкового текста и различных видов семантической обработки знаний.

Подход, основанный на сочетании грамматики когнитивного переноса и методов машинного обучения, обеспечивает надёжную и расширяемую платформу для моделирования межъязыкового синтактико-семантического трансфера (переноса) и может быть применён к большему числу языков (особенно имеющих сходные структуры категориальных атрибутов — значений). Однако, проблемы обработки разорванных структур, референциальных ссылок и разрешения неоднозначности, хотя и частично решены, требуют развития используемого механизма.

Наши дальнейшие исследования связаны с разработкой специальных более дифференцированных расширений существующей системы атрибутов-значений для отслеживания отдалённых компонентов разделённых цепочек, уточнением семантики проблемных головных вершин и глагольных фреймов, адаптацией многочисленных фразовых комплексов и фразеологизмов, а также развитием механизмов разрешения неоднозначности фразовых структур с использованием вероятностных методов.

Список литературы

1. *Chomsky N.* Aspects of the Theory of Syntax. MIT Press, Cambridge, MA, 1965. Русский перевод: Н. Хомский. Аспекты теории синтаксиса. — Благовещенск: Благовещенский Гуманитарный Колледж им. И. А. Бодуэна де Куртене, 1999. — 258 с.
2. *Kay M.* Functional grammar // BLS-79, Berkeley, CA, 1979. P. 142–158.
3. *Colmerauer A.* Metamorphosis grammars // Natural Language Communications with Computers, Lecture Notes in Computer Science 63. L. Bolc (Ed.), Berlin: Springer Verlag, 1978. P. 133–189.
4. *Pereira F., Warren D.H.D.* Definite clause grammars for language analysis — a survey of the formalism and a comparison with augmented transition networks // Artificial Intelligence, 13(3), 1980. P. 231–278.
5. *Woods W.A.* What's in a link: Foundations for semantic networks // Representation and Understanding: Studies in Cognitive Science. Bobrow, D.G. and Collins A. M. (Eds.). — N. Y.: Academic Press, 1975. P. 35–82.
6. *Colmerauer A., Roussel P.* The birth of Prolog // History of Programming Languages — II. Bergin Jr., T.J. and Gibson, Jr., R.G. (Eds.). — N. Y.: ACM Press / Addison-Wesley, 1996. P. 331–352.
7. *Bobrow R.J., Webber B.* Knowledge representation for syntactic / semantic processing // AAAI-80, Stanford, CA: Morgan Kaufmann, 1980. P. 316–323.
8. *Bresnan J.* (Ed.) The Mental Representation of Grammatical Relations. Cambridge, MA: MIT Press, 1982.
9. *Pollard C., Sag I.A.* Head-Driven Phrase Structure Grammar. Chicago:University of Chicago Press, 1994.
10. *Kay P., Fillmore C.J.* Grammatical constructions and linguistic generalizations: The What's X Doing Y? construction // Language, 75(1), 1999. P. 1–33.
11. *Uszkoreit H.* Categorial unification grammars // COLING-86, Bonn, 1986. P. 187–194.
12. *Adjukiewicz K.* Die syntaktische Konnexitat // Studia Philosophica, 1, 1935. 1–27. English translation «Syntactic Connection» by H. Weber in McCall, S. (Ed.) Polish Logic. — Oxford: Oxford University Press, 1967. P. 207–231.
13. *Ламбек И.* Математическое исследование структуры предложения // Шрейдер Ю.А., Ревзин И.И., Лахути Д.Г., Финн Ф.К. (ред.) Математическая лингвистика. -М.: Мир, 1964. (Lambek J. The

mathematics of sentence structure. American Mathematical Monthly, 65(3), 1958. P. 154–170.

14. *Бар-Хиллел И.* Разрешающие процедуры для структуры естественных языков // Математическая лингвистика / Под ред. Ю.А. Шрейдера, И.И. Ревзина, Д.Г. Лахути, Ф.К. Финна. — М.: Мир, 1964.
15. *Moortgat M.* Categorial investigations: logical and linguistic aspects of the Lambek calculus. — Dordrecht: Foris, 1988.
16. *Steedman M.J.* Constituency and coordination in a combinatory grammar // Alternative Conceptions of Phrase Structure / Ed. by M.R. Batlin, A.S. Kroch. — Chicago: University of Chicago, 1989. — P. 201–231.
17. *Steedman M.* The syntactic process. Cambridge (Mass.): The MIT Press, 2000.
18. *Shaumyan S.* A Semiotic Theory of Language. Indiana University Press, 1987.
19. *Shaumyan S.* Signs, Mind, and Reality. John Benjamins Publishing Company, USA, 2006.
20. *Grover C., Carroll J., Briscoe T.* The Alvey Natural Language Tools Grammar (4-th Release). Technical Report, 1993, Computer Laboratory, University of Cambridge, 1993.
21. *Gazdar G., Klein E., Pullum G.* Sag,I. Generalized Phrase Structure Grammar. — Oxford: Basil Blackwell, 1985.
22. *Pollard C., Sag I.A.* Head-Driven Phrase Structure Grammar. Chicago:University of Chicago Press, 1994.
23. *Ristard E.S.* Computational complexity of current GPSG theory. Proceedings of the 24-th Annual Meeting of the Association for Computational Linguistics. Columbia University. — N.Y.: Association for Computational Linguistics. 1986. P. 30–39.
24. *Mel'cuk I.A.* Dependency Syntax: Theory and Practice, State University of N.Y. Press, 1988.
25. *Gazdar G., Mellish C.* Natural Language Processing in Prolog. Wokingam, UK: Addison-Wesley, 1989.
26. *Kozerenko E.B.* Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23–26, 2003, Las Vegas, USA // CSREA Press, 2003. P. 49–55.
27. *Бондарко А.В.* Принципы функциональной грамматики и вопросы аспектологии. — М.: УРСС, 2003. — 208 с.

28. *Кибрик А.Е.* Очерки по общим и прикладным вопросам языкознания. — М.: УРСС, 2001. — 334 с.
29. *Золотова Г.А.* Коммуникативные аспекты русского синтаксиса. — М.: УРСС, 2001. — 368 с.
30. *Kay M., Gawron J., Norvig P.* Verbmobil: A Translation System for Face-to-Face Dialog. CSLI; 1992.
31. *Frederking R., Rudnicky A.I., Hogan C.* Interactive speech translation in the DIPLOMAT project. In Proceedings of the ACL-97 Spoken Language Translation Workshop, Madrid. P. 61–66. ACL, 1997.
32. *Brown P.F., Cocke J., Della S.A., Pietra V.J.* Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer & P.S. Roossin. A statistical approach to machine translation. Computational Linguistics 16. P. 79–85, 1990.
33. *Brown P.F., Della S.A., Pietra V.J.* Della Pietra and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2). P. 263–311, 1993.
34. *Marcus M.P., Santorini B., Marcinkiewicz M.A.* Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2), 313–330.
35. Dorr, Bonnie and Nizar Habash. Interlingua Approximation: A Generation-Heavy Approach. AMTA-2002 Interlingua Reliability Workshop. Tiburon, California, USA, 2002.