РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 1 Москва 2015

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

УДК 004.89 : 510.64

М.И. Забежайло

К вопросу о достаточности оснований для принятия результатов интеллектуального анализа данных средствами ДСМ-метода

Обсуждается проблема контроля достаточности оснований для принятия результатов порождения средствами ДСМ-метода эмпирических зависимостей из данных. Для реализации такого контроля формулируется процедурная конструкция, базирующаяся на анализе систем классов эквивалентности, покрывающих соответствующие классы сходства. Демонстрируются взаимосвязи предлагаемого механизма контроля достаточности оснований с базовыми конструкциями ДСМ-метода (правилами правдоподобного вывода первого и второго рода).

Ключевые слова: *интеллектуальный анализ данных, ДСМ-метод автоматического порождения гипотез, контроль достаточности оснований для принятия результатов ДСМ-ИАД*

Интеллектуальный анализ данных (ИАД) сегодня позиционируется как одно из быстро развивающихся направлений не только теоретических изысканий, но и реальных промышленных приложений, востребованное в том числе и для решения трудных практически значимых задач в целом ряде прикладных предметных областей [1-3].

Несложно убедиться, что достаточно обширный класс задач интеллектуального анализа, ориентированных на восстановление эмпирических зависимостей, неявным образом представленных в тех или иных коллекциях экспериментальных данных, в самом общем виде может быть описан следующей схемой.

Дано:

1) множество (примеров) *объектов*, обладающих целевыми *свойствами*, и множество[1] (контр-

[1] В исходном состоянии, – возможно, пустое.

примеров) объектов, не обладающих целевыми свойствами,

а также

2) *новый объект* (или же некоторое явным образом представленное множество таких объектов).

Требуется: оценить *наличие* (или отсутствие) *целевых свойств* у нового *объекта* (новых объектов из заданного множества), т.е.

- дать соответствующий *прогноз* (о наличии целевых свойств);

- предъявить *основания* (неоспоримые *аргументы*), позволяющие *принять* этот прогноз.

Формализованные постановки задач этого вида могут использовать (т.е. опираться на) те или иные математические техники – методы и модели статистического анализа данных, машинного обучения, нейронных сетей, распознавания образов или автоматической классификации и др.[2] Имеется обширная литература, представляющая особенности формирования и применения в приложениях как подобных математических моделей, так и реализующих их инструментальных программных систем[3].

Однако критически важным аспектом, определяющим возможности использования каждого такого формализованного уточнения обсуждаемой общей задачи, остается вопрос о существовании корректного[4] решения, порождаемого соответствующим методом (формализованным подходом). Фактически речь здесь идет о существовании надежных «*источников доверия*» к получаемым соответствующим инструментарием результатам (другими словами – о возможности предъявить *достаточные основания* для принятия получаемых результатов). Наличие (в конкретном случае – при решении той или иной прикладной задачи) таких *достаточных оснований* – весомый аргумент в пользу доверительного отношения к результатам, получаемым с применением соответствующего инструментария. (Отсутствие возможностей предъявить подобные *достаточные основания* – дополнительный повод удостовериться в обоснованности использования в данном конкретном случае именно этого формального инструментария интеллектуального анализа данных).

Классический пример элегантного решения (математическими средствами – !) вопроса о существовании *достаточных оснований* для принятия результатов интеллектуального анализа данных был дан Ю.И.Журавлевым для обширного класса задач распознавания образов (см., например, работы [4-6] и др.). Предложенные здесь корректные алгебры над множеством некорректных эвристических алгоритмов, использующих технику вычисления оценок при решении задач распознавания образов, обеспечивают доказательство существования (при выполнении определенных условий) корректного[5] алгоритма распознавания и предлагают соответствующие инструменты[6] для его восстановления по исходно заданному набору, вообще говоря, некорректных (т.е. вычисляющих корректные результаты лишь на некоторых – характерных для каждого из них – подмножествах возможных исходных данных) алгоритмов.

Следуя мотивации, только что представленной на примере использования алгебраических соображений при формировании достаточных оснований для принятия результатов интеллектуального анализа данных в соответствующем классе задач распознавания образов, представляется естественным попытаться сформулировать подобную (в части мотивации[7]) конструкцию и для исходной рассматриваемой нами задачи (оценки наличия у вновь предъявляемого объекта целевых свойств, характерных для уже заданного множества объектов-примеров).

Итак, наша цель – не только научиться

- использовать имеющуюся (накапливаемую, доступную, …) информацию об уже идентифицированных экземплярах (примерах) объектов, обладающих целевыми свойствами, причем – вместе с информацией об уже идентифицированных экземплярах (контрпримерах) объектов, не обладающих целевыми свойствами, для идентификации новых (ранее не классифицированных) экземпляров, предъявляемых для идентификации объектов,

но и развить «навыки», позволяющие

- формировать соответствующие (желательно – быстро[8]) проверяемые условия, выполнимость которых обеспечивала бы наличие *достаточных оснований* для *принятия результатов* идентификации (используемыми конкретными инструментальными средствами) вновь предъявленных объектов.

ИДЕЯ (ЦЕЛЕВОГО) ПОДХОДА К РЕШЕНИЮ ПОСТАВЛЕННОЙ ЗАДАЧИ

1. Каждый представленный в исходных данных объект может быть охарактеризован набором признаков (например, определенным набором тех или иных последовательностей 0 и 1 в его «теле»).

2. Все множество таких признаков может быть собрано в (пополняемый и расширяемый) каталог – «алфавит» («атомарных») признаков. Каждый уже идентифицированный ранее объект может быть «закодирован» (представлен) определенным набором (множеством) таких признаков.

3. На множестве идентификационных признаков могут быть заданы отношения (т.е. описания исходно заданных объектов могут стать не множествами признаков, а, например, графами, цепочками и т.п.). Описания изучаемых объектов могут быть расширены также числовыми значениями тех или

[2] См., например, ссылки к работе [3] и др.

[3] См., например, [2, 3] и др.

[4] Причем – ***доказуемо*** корректного!

[5] Формирующего математически корректное решение предъявленной прикладной задачи.

[6] Корректные алгебры (см., например, работы [4-6] и др.).

[7] Т.е. в части поиска и формирования *проверяемых условий*, позволяющих отвечать на вопрос о наличии (или же об отсутствии) в каждом конкретном случае *достаточных оснований* для того, чтобы *доверять* получаемым в ходе выполняемых ИАД-вычислений *результатам*.

[8] Т.е. средствами полиномиально сложных вычислений.

иных параметров (*множества признаков + отношения + числовые характеристики*).

4. В случае пустоты исходно заданного множества контрпримеров, к имеющимся коллекциям объектов (примерам изучаемых объектов) путем дополнительного анализа[9] изучаемой предметной области могут быть добавлены также наборы контрпримеров – комплексов признаков из исходного «алфавита», которые, тем не менее, не являются объектами, обладающими целевыми признаками.

5. На множествах примеров и контрпримеров может быть определена (алгебраическая) операция сходства *, удовлетворяющая трем условиям:

- a*a = a
- a*в = в*a
- (a*в)*c = a*(в*c) = a*в*c

Например: * определяемая как пересечение множеств, или же как выделение множества всех максимальных общих подграфов двух графов, множества всех максимальных общих подцепочек двух цепочек и т.п.

6. С помощью такой операции * на примерах и контрпримерах строятся классы сходства: для каждого заданного объекта исходной коллекции примеров (и, соответственно, контрпримеров) это – множество всех известных объектов, которые сходны[10] с ним.

7. Диагностика (идентификация) новых объектов производится с помощью порождаемых таким образом классов сходства: возможность отнесения нового – диагностируемого – объекта к какому-либо из уже реконструированных (по примерам[11]) классов сходства позволяет утверждать, что это – еще один объект, аналогичный (см. класс сходства) ранее уже наблюдавшимся (в исходной обучающей выборке примеров и контрпримеров).

8. Идея техники отнесения нового объекта к классу выглядит следующим образом:

- класс сходства есть объединение некоторых классов эквивалентности[12] [7];
- по построенному классу сходства можно восстановить (с учетом имеющегося контекста – т.е. множества примеров и контрпримеров, задействованных в обучении) порождающие его классы эквивалентности;
- каждый класс сходства при восстановлении классов эквивалентности, объединением которых он является, будет характеризоваться непустыми (ненулевыми) результатами вычисления уже заданной операции сходства. Таким образом, каждый восстанавливаемый здесь класс эквивалентности будет определяться теми примерами из исходной выборки, результат вычисления операции сходства для которых будет одним и тем же[13] (например, в случае простого сходства на множествах – всех примеров, содержащих общее подмножество образующих);

следовательно:

- каждый класс эквивалентности в таком случае – орбита (множество примеров-родителей) для соответствующей ДСМ-гипотезы. (При этом, разумеется, один и тот же пример из исходной ДСМ-выборки может попадать в несколько пересекающихся, но не совпадающих классов эквивалентности[14]);
- в дополнение к этому порождаемые на исходном множестве примеров и контрпримеров классы сходства и соответствующие классы эквивалентности (как и в случае рассмотрения ДСМ-рассуждений) могут характеризоваться выполнимостью дополнительных логических условий – в частности: запрета на контрпримеры, условия единственности причины и т.п.;
- здесь важно заметить, что в ДСМ-рассуждении используются одновременно классы сходства/эквивалентности, порожденные как на позитивных, так и на негативных примерах (контрпримерах). Критерий достаточности оснований для принятия результатов ДСМ-рассуждения допускает принятие лишь тех результатов правдоподобного вывода одного знака, которые не фальсифицируемы результатами правдоподобного вывода другого знака. Таким образом, в дополнение к традиционной для распознавания образов технике порождения решающих правил добавляются критически важные условия нефальсифицируемости порождаемых результатов в смысле К.Поппера. (При этом позитивные и негативные классы эквивалентности при их порождении в определенном смысле равноправны: результаты одного знака используются при фальсификации результатов другого знака и – наоборот. На достаточном основании принимаются лишь непротиворечивые – нефальсифицируемые – результаты, а процедура взаимной фальсификации требует исчерпывающего просмотра зависимостей – классов эквивалентности, – порождаемых на заданном множестве примеров и контрпримеров);
- каждый набор классов эквивалентности можно «сузить» до фактор-множества. В данном случае каждый класс эквивалентности и его представитель в фактор-множестве будет задаваться соответствующим результатом вычисления операции сходства (ДСМ-гипотезой), а фактор-множество – множеством непротиворечивых ДСМ-гипотез;
- проверкой «аналогии» между элементами фактор-множества и новым – диагностируемым – объектом исполняется идентификация последнего как нового объекта, соотносимого с подмножеством (классом

---

[9] Проводимого с привлечением экспертов по изучаемой предметной области.

[10] Т.е. операция сходства в каждом таком случае дает непустой (ненулевой) результат.

[11] Или же контрпримерам.

[12] Порождаемых всеми возможными в данном случае версиями соответствующего отношения эквивалентности.

[13] Именно этим фактом дополнительно к рефлексивности (идемпотентности соответствующей операции сходства) и симметричности обеспечивается также и транзитивность соответствующего (уже используемой операции сходства) отношения сходства (суженного представленным дополнительным условием до эквивалентности).

[14] Для, разумеется, различных – см. различные ДСМ-гипотезы – конкретных вариантов отношений эквивалентности, участвующих в формировании соответствующих классов сходства.

сходства как объединением соответствующих классов эквивалентности) исходно заданных для обучения примеров как *подобный* им;

- при этом необходимо, чтобы вновь идентифицируемый объект попадал лишь в классы эквивалентности одного знака. В противном случае при идентификации порождается противоречие (что в силу формулировки Правил Правдоподобного Вывода II рода (ППВ-II) в ДСМ-методе порождает противоречивое доопределение по правилу аналогии);
- наконец, идентификация в порожденных на исходно заданном для обучения множестве примеров классах эквивалентности одноэлементных – т.е. вырожденных – классов (что в рамках ДСМ-рассуждения сигнализирует о невыполнении Условия Каузальной Полноты – см., например, [8, 9] и др. – и также о необходимости пополнять исходную обучающую выборку новыми примерами, позволяющими обеспечить выполнение названного Условия), указывает на необходимость расширения исходной выборки примеров такими новыми примерами, которые были бы сходны с примерами, образующими вырожденные классы эквивалентности[15].

9. Поддерживая (формируя, сопровождая и пополняя) базы описаний анализируемых объектов (например, *сигнатур вирусов при анализе вредоносного программного обеспечения* и т.п.) можно оперативно диагностировать новые (потенциально – наделенные искомыми свойствами) объекты, проверяя (в том числе – в режиме глубокого «распараллеливания») сходство каждого такого объекта с уже накопленными в соответствующих базах признаками (а также их «существенными» – см. обучение на примерах и контрпримерах – *комбинациями* и т.п.). При этом собственно проверка такого сходства реализуется как проверка попадания[16] нового (идентифицируемого в настоящий момент) объекта последовательно в каждый из порожденных на текущий момент (невырожденных – !) классов эквивалентности.

10. Стабильность системы классов сходства[17] (и вместе с ними – соответствующих классов эквивалентности) при расширении исходной обучающей выборки примеров и контрпримеров новыми объектами может рассматриваться как сигнал о том, что накопленную выборку можно оценить как *насыщенную* в части порождения (определяемого соответствующей системой классов эквивалентности) множества эмпирических зависимостей (в т.ч. – тенденций и закономерностей – см., например, [10]).

* * *

Нашей основной целью будет продемонстрировать, как представленная выше схема рассуждений о возможностях аргументированного отнесения нового объекта к множеству уже изученных (имеющих достаточно информативное[18] описание их «структуры» и обладающих целевыми свойствами) может специфицироваться (уточняться). Причем такие уточнения должны быть возможны как в части формирования и комбинирования (т.е. «сборки» в те или иные *стратегии* рассуждения) разного вида однородных[19] процедур интеллектуального анализа данных, так и в части взаимно согласованной[20] обработки предлагаемыми процедурами интеллектуального анализа данных (ИАД) различных типов представлений данных. Спектр доступных здесь вариантов использования дескриптивных возможностей для поиска зависимостей, в неявном виде содержащихся в накапливаемых эмпирических данных, может варьироваться от достаточно «грубых»[21], но «обсчитываемых» процедурами невысокой вычислительной сложности, до весьма детальных[22], однако, как правило, требующих экспоненциально сложных вычислений.

Обратимся к более детальному анализу только что представленных соображений алгебраического характера, а также их взаимосвязей с базовыми компонентами ДСМ-метода автоматического порождения зависимостей из эмпирических данных.

Прежде чем двигаться вперед, сформулируем некоторое предположение (своеобразный «*постулат значения*»), которому отводится критически важная роль во всех наших дальнейших построениях: мы будем считать, что у каждого свойства (или же множества свойств) рассматриваемых нами объектов, в описаниях последних существует некоторый фрагмент, отвечающий за наличие этого свойства (множества свойств). Другими словами, вне зависимости от выбора конкретного языка представления анализируемых объектов (их представления, например, множествами признаков[23] или же множествами значений признаков с отношениями на признаках – графами, цепочками – и т.п.) в случае наличия у объекта анализируемых свойств в его описании может быть выделен определенный подобъект

---

[15] Что (см. описание структуры ДСМ-рассуждения, например, в [8, 9] и др.) предоставляет возможности для целенаправленного расширения исходной выборки примеров для обучения.

[16] Причем – непротиворечивого (см. выше) попадания.

[17] См. также замечание о свойствах соответствующих пространств толерантности в работе [7].

[18] Здесь представляется уместным напомнить о требованиях к дескриптивным возможностям используемого языка представления исходных данных (см., например, [8,10] и др.): задействованный способ описания данных должен иметь соответствующие дескриптивные возможности для адекватного представления (т.е. записи в виде формальных выражений используемого языка) восстанавливаемых из данных зависимостей.

[19] Выстроенных в единой «логике» рассуждения.

[20] В данном случае подразумевается, что различные (по своим выразительным возможностям) формализованные описания одних и тех же эмпирических данных обрабатываются однотипными по «логике» рассуждения процедурами ИАД.

[21] Например, в виде булевских векторов значений характерных для рассматриваемой ситуации признаков.

[22] Например (в случае работы с биохимическими данными), в виде пространственных графов (рассматриваемых физиологически активных соединений), дополненных числовыми значениями релевантных физико-химических параметров (доз, концентраций, температур и т.п.).

[23] Булевскими векторами значений этих признаков.

(подмножество, подграф, подцепочка и т.п.), представляющий собой «*структурный носитель*» соответствующего множества свойств.

РАЗДЕЛ I.

Начнем с простейшего случая.

Дано: множество **Ω** объектов (примеров) **C** , каждый из которых обладает некоторым свойством **A**. Каждый элемент **C** из **Ω** характеризуется некоторым (структурным) описанием[24].Также задан (соответствующим описанием) новый объект $\mathbf{C_0}$, о наличии (либо отсутствии) у него анализируемого свойства **A** нет информации.

Требуется: сформировать аргументированное суждение относительно наличия (или же отсутствия) у $\mathbf{C_0}$ свойства **A**.

При формировании такого прогноза и поиске соответствующих аргументов представляется уместным воспользоваться следующими простейшими соображениями:

1) *если все объекты* **C** *из* **Ω** *обладают свойством* **A**, *то естественно предположить, что в их структурных описаниях имеется некоторая «общая часть»* **V**, *которая (в случае единственности причины для* **A***) «вкладывается» во все примеры из* **Ω;**

2) в свою очередь в, «*многопричинной*» ситуации *в* **Ω** *должны содержаться подмножества объектов (примеров) из множества* $\mathbf{\Omega_i}$, *где каждый элемент для каждого такого* $\mathbf{\Omega_i}$ *будет иметь в виде общего «фрагмента» - «носителя свойства* **A***» - соответствующий подобъект (подмножество образующих, подграф, подцепочку и т.п.)* $\mathbf{V_i}$.

Мы начнем наши построения с уточнения простейшего случая и будем рассматривать ситуацию, когда в нашем распоряжении имеется множество Ω объектов $C_1, C_2, \dots, C_n$, каждый из которых обладает заданным свойством A. В исходном состоянии нам не важно, как именно представлены объекты C_i своими формализованными описаниями: необходимо лишь, чтобы в этих описаниях можно было бы выделять общие части[25]. Другими словами, необходимо, чтобы на множестве объектов (включая их составные части) была бы определена операция сходства ⊗ со следующими свойствами:

(i) $c \otimes c = c$

(ii) $c_r \otimes c_s = c_s \otimes c_r$

(iii) $(c_r \otimes c_s) \otimes c_t = c_r \otimes (c_s \otimes c_t) = c_r \otimes c_s \otimes c_t$.

Таким образом, пара $<\Omega,\otimes>$ определяет пространство толерантности в смысле [7, 11] и др., так как всякому непустому[26] результату применения операции ⊗ к каждой конкретной паре объектов $<c_r,c_s>$ можно сопоставить точно такую же пару объектов, принадлежащих (сопряженному с операцией ⊗) бинарному отношению ⊗', которое по условиям (i)-(iii) оказывается отношением толерантности (сходства).

Для каждого конкретного объекта C_r классом сходства $T_\Omega(C_r)$ по определению будем считать множество всех тех примеров из Ω, которые сходны с этим C_r в смысле определенного нами отношения ⊗'. Если при этом зафиксировать конкретное значение V_0 результат применения операции ⊗ к соответствующим элементам $T_\Omega(C_r)$, и рассматривать лишь все содержащие максимальную общую часть V_0 объекты, то порождаемое таким способом отношение ⊗'' оказывается отношением эквивалентности. Имеет место простое

Утверждение 1.

Пусть заданы: множество Ω, удовлетворяющая ограничениям (i)-(iii) операция ⊗, некоторый C_r из Ω а также некоторый подобъект V, характеризующий сходство объекта C_rпо крайней мере с одним из элементов класса сходства $T_\Omega(C_r)$. Множество всех таких элементов C_s из $T_\Omega(C_r)$, что

$$C_r \otimes C_s = V$$

представляет собой класс эквивалентности $E_{\Omega,v}(C_r)$.

Доказательство.

Рефлексивность и симметричность порождаемого фиксацией сходства V отношения ⊗'' определяется соответствующими свойствами отношения ⊗'. Транзитивность вытекает из следующего простого рассуждения: будучи общей частью одновременно пар $<C_r,C_s>$ и $<C_s,C_t>$, заданный подобъект V оказывается также общей частью примеров C_r и C_t .

□

В представленной ситуации для каждого порождаемого из множества примеров Ω применением операции ⊗ сходства V_j можно подобрать максимальное по вложению подмножество примеров из Ω, таких, что все они образуют класс эквивалентности по отношению $\otimes''(V_j)$. Таким образом, нетрудно убедиться, что для представленного Утверждения 1 имеет место

Следствие 1.

Пусть заданы множество примеров Ω и отношение ⊗'. Соответствующее пространство толерантности $<\Omega,\otimes'>$ может быть представлено в виде объединения классов эквивалентности, формируемых по всем возможным порождаемым на $<\Omega,\otimes'>$ отношениям эквивалентности ⊗''.

□

Таким образом, для того чтобы иметь основания утверждать[27], что новый (предложенный для прогно-

[24] Например, множеством значений признаков – множеством образующих (букв) некоторого алфавита **U**, некоторым графом, цепочкой символов, числовым вектором и т.п.

[25] При этом представляется естественным ради сокращения перебора рассматривать лишь соответствующие максимальные общие части анализируемых объектов (примеров).

[26] Т.е. дополнив множество Ω некоторым специальным примером $C_\varnothing$, который по определению будем называть *пустым* объектом, можно говорить о пустоте (отсутствии общих частей у сравниваемых объектов, выражающемся в равенстве $C_\varnothing$ результата вычисления операции ⊗ на этих объектах) и непустоте соответствующего сходства.

[27] По-видимому, более аккуратно было бы говорить об

за свойств) объект $\mathbf{C_0}$ также имеет свойство **A**, в обсуждаемой нами схеме рассуждений следует убедиться, что данный $\mathbf{C_0}$ «попадает» (т.е. может быть помещен как новый элемент) хотя бы в один из классов эквивалентности $E_{\Omega,v}(C_r)$, сформированных на исходном множестве примеров Ω всеми возможными в данном случае отношениями эквивалентности ⊗''.

Процедурно этот процесс можно было бы реализовать следующим образом:

- перечислим все классы эквивалентности, покрывающие соответствующие классы сходства пространства толерантности <Ω,⊗'>,

- для каждого такого класса эквивалентности (номер *l*) выделим характеризующую его «общую часть» V_l входящих в него примеров из Ω;

- проверим, найдется ли среди всех таких V_l по крайней мере одно, вкладывающееся в $\mathbf{C_0}$.

(При этом для «экономной» организации такого процесса «диагностики» свойства **A** у $\mathbf{C_0}$ достаточно ограничиться, например, порождением лишь минимальных по взаимному вложению общих частей вида V_l). Тем не менее, при формировании (переборе) множества всех возможных (для заданного Ω) максимальных общих частей вида V_{l0} необходимо:

(1) выбрать из Ω некоторое множества примеров $C_{i1}, C_{i2}, \dots, C_{ik}$,

(2) убедиться, что сходство объектов $C_{i1}\otimes C_{i2}\otimes\dots\otimes C_{ik}= V_{l0}$ не является пустым подобъектом $C_\varnothing$;

(3) проверить, что все элементы в множестве $\{C_{i1}, C_{i2}, \dots, C_{ik}\}$ попарно различны;

(4) убедиться, что в множество $\{C_{i1}, C_{i2}, \dots, C_{ik}\}$ включены все примеры из исходного множества Ω , содержащие V_{l0} как подобъект, и, наконец, что

(5) в множестве $\{C_{i1}, C_{i2}, \dots, C_{ik}\}$ имеется не менее двух примеров (т.е. выделяемое сходство V_{l0} не является тривиальным): $k \geq 2$.

Рассмотрим четыре переборные задачи, характеризующие комбинаторные свойства предложенной конструкции.

Определение 1.

1. Задачу, где

Дано: множество примеров Ω (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства ⊗ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества Ω классами эквивалентности, сформированным по всем возможным отношениям ⊗'', такой класс $E_{\Omega,V0}$, который содержит ровно k различных элементов из Ω,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** *k*.

2. Задачу, где

Дано: множество примеров Ω (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства ⊗ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества Ω классами эквивалентности, сформированным по всем возможным отношениям ⊗'', такой класс $E_{\Omega,V0}$, который содержит не менее k различных элементов из Ω,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ*** *k*.

3. Задачу, где

Дано: множество примеров Ω (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства ⊗ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества Ω классами эквивалентности, сформированным по всем возможным отношениям ⊗'', такой класс $E_{\Omega,V0}$, который содержит не более k различных элементов из Ω,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ*** *k*.

4. Задачу, где

Дано: множество примеров Ω (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства ⊗.

Требуется: установить, сколько элементов (различных классов эквивалентности) имеется в покрытии множества Ω классами эквивалентности, сформированным по всем возможным отношениям ⊗'',

мы далее будем называть задачей ***ЧИСЛО КЛАССОВ ЭКВИВАЛЕНТНОСТИ***.

□

Характер сложности этих комбинаторных задач (в частности трудноразрешимость Задач 1 и 4) уже в простейшем случае – при рассмотрении объектов C_i в множестве Ω как непустых подмножеств букв некоторого конечного алфавита U, – иллюстрируют следующие три утверждения (см. подробнее, например, работу [12]):

Утверждение 2.

Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** *k* принадлежит классу ***NPC***[28] (***NP***-полных задач).

Утверждение 3.

Задача ***ЧИСЛО КЛАССОВ ЭКВИВАЛЕНТНОСТИ*** принадлежит классу #***PC*** (перечислительно полных задач).

Утверждение 4.

Задачи ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ*** *k* и ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ*** *k* разрешимы полиномиально быстро.

□

отсутствии аргументов, позволяющих оспорить такое заключение.

[28] См. подробнее [13-16] и др.

РАЗДЕЛ II.

Теперь обратимся к более сложному случаю: к ситуации, когда объекты (примеры) из исходно заданного множества Ω могут обладать наборами свойств, т.е. к ситуации, когда **A** – это уже не одноэлементное множество свойств $\mathbf{A} = \{A_1, A_2, \dots, A_m\}$, и каждому объекту C_i из Ω сопоставлено свое (непустое) подмножество $\mathbf{A}_i$ свойств из **A**.

В предложенном контексте сформулированные выше требования (1) – (5) нам придется дополнить естественным новым требованием об отборе (в каждый из классов эквивалентности, формируемых нами для «диагностирования» вновь предложенных на прогноз объектов) тех и только тех объектов из исходного множества Ω, которые имеют одни и те же (общие - !) свойства из **A**. Таким образом, нам придется рассматривать пары $<C_i,\mathbf{A}_i>$, причем такие, что для каждого удовлетворяющего требованиям (1) – (5) набора объектов $C_{i1}, C_{i2}, \dots, C_{ik}$ множество их общих свойств $\mathbf{A}_V$ не пусто, т.е.

(6) для каждого удовлетворяющего требованиям (1) – (5) множества примеров $\{C_{i1}, C_{i2}, \dots, C_{ik}\}$, общая часть которых V не является пустым подобъектом $C_\varnothing$; множество их общих свойств $\mathbf{A}_V$ удовлетворяет условию $\mathbf{A}_V \neq \varnothing$.

При этом в каждом таком случае пара $<V,\mathbf{A}_V>$ будет представлять V как *причину* наличия у объектов из множества $\{C_{i1}, C_{i2}, \dots, C_{ik}\}$ (т.е. «*структурный носитель*»[29]) множества свойств $\mathbf{A}_V$.

Теперь отнесение нового (предложенного для прогноза свойств) объекта $C_\mathbf{0}$ к классу эквивалентности $E_{\Omega,V}$ будет означать, что свойства этого объекта полностью определяет множество $\mathbf{A}_V$ из соответствующей пары $<V,\mathbf{A}_V>$. (Разумеется, для попадания $C_\mathbf{0}$ в класс $E_{\Omega,V}$ требуется, чтобы $C_\mathbf{0}$ содержал этот непустой подобъект V).

Воспользуемся для описания связи *объект* C *обладает свойствами* **A** частично определенным отношением $=>_1$, а для описания связи *подобъект* V *есть причина* (*структурный носитель*) *множества свойств* $\mathbf{A}_V$ – частично определенным $=>_2$. Далее, используем предложенные в [8 и др.] логические средства (язык описания решающих предикатов и правил правдоподобного вывода ДСМ-метода), затем выберем для описания примеров из Ω их представления в виде множеств букв некоторого алфавита U, а в качестве операции сходства ⊗ используем операцию ∩ пересечения множеств, после чего запишем с их помощью представленные выше требования (1) – (2). Нетрудно убедиться, что это будет[30] формула следующего вида:

$$M^+_a(V,W) = \exists k \underline{M^+_a}(V,W,k) =$$

$$= \exists k\, \exists C_1 \dots \exists C_k \exists A_1 \dots \exists A_k \Big([\mathbf{J}_{<1,r>}(C_1 ==>_1 A_1) \,\&$$

$$\&\forall A\, (\mathbf{J}_{<1,r>}(C_1 ==>_1 A) \rightarrow (A \subseteq A_1))\,] \,\& \dots$$

$$\dots \&[\, \mathbf{J}_{<1,r>}(C_k ==>_1 A_k) \,\&$$

$$\&\forall A\, (\mathbf{J}_{<1,r>}(C_k ==>_1 A) \rightarrow (A \subseteq A_k))\,] \,\&$$

$$\&[\, (\bigcap_{i=1}^{k} C_i = V) \,\&\, (V \neq \varnothing)\,] \,\&$$

$$\&\,[\,\forall i \forall j\, (\,[(i \neq j)\&(1 \leq i,j \leq k)] \rightarrow (C_i \neq C_j)\,)\,] \,\&$$

$$\&\forall X\, \forall Y \,\big[\, [\mathbf{J}_{<1,r>}(X ==>_1 Y) \,\&$$

$$\&\forall A\, (\mathbf{J}_{<1,r>}(X ==>_1 A) \rightarrow (A \subseteq Y))\, \&\, (V \subset X)] \rightarrow$$

$$\rightarrow [\, (W \neq \varnothing) \,\&\, (W \subseteq Y) \,\&$$

$$\&\, ((X = C_1) \vee \dots \vee (X = C_k))\,]\,\big] \&(k \geq 2)\,\Big)$$

Другими словами, спецификация (уточнение) предложенной выше схемы диагностики свойств новых объектов путем отнесения их (если это возможно) к порождаемым на исходно заданной для обучения выборке примеров, реализованная формальным уточнением соответствующих

- языка представления данных (в данном случае тип данных – множества признаков) и

- формального представления операции сходства на используемом типе представления данных как операции пересечения множеств, привела нас к формализованной записи одной из базовых конструкций ДСМ-метода[31] – так называемого предиката простого сходства (см. подробнее [17, 8] и др.). Зафиксируем этот факт и двинемся дальше.

РАЗДЕЛ III.

Перейдем к анализу ситуации, когда исходное множество Ω сформировано как *примерами* (т.е. объектами, которые обладают подлежащими изучению свойствами), так и *контрпримерами* (т.е. объектами, которые, – как известно из каких-либо «внешних» по отношению к выполняемому интеллектуальному анализу источников, – не обладают подлежащими изучению свойствами, однако имеют «подобную» *примерам* структуру[32]): $\Omega = \Omega^+ \cup \Omega^-$, но, $\Omega^+ \cap \Omega^- = \varnothing$. В такой ситуации особого обсуждения заслуживают следующие два случая:

а) в множествах $\mathbf{E}_{\Omega^+}$ и $\mathbf{E}_{\Omega^-}$ классов эквивалентности, построенных соответственно на Ω^+ и Ω^-, найдутся соответственно два класса E_{Ω^+,V^+} и E_{Ω^-,V^-}, для которых $V^+ = V^-$ при одном и том же множестве свойств $\mathbf{A}_V$;

[29] Разумеется, с точностью до по-элементного вида исходного множества примеров Ω.

[30] С точностью до индексов у J-функций, характеризующих в ДСМ-предикатах типы истинностных значений и номер текущего шага ДСМ-рассуждения, а также отбора в соответствующий класс эквивалентности из каждого множества всех примеров с одинаковым описанием структуры C_i того из них, который имеет максимальное по вложению множество свойств **A**.

[31] Заметим, сформированной из совершенно других понятийных оснований – из уточнений логическими средствами известных индуктивных канонов рассуждения Д.С.Милля (см., например, [17] и др.)

[32] В их структурах встречаются такие же, как и в структурах *примеров* фрагменты (подобъекты).

b) предлагаемый для прогноза свойств новый объект C_0 «попадает» одновременно как в некоторый «позитивный» класс эквивалентности $E_{\Omega+,V+}$, так и в некоторый негативный класс эквивалентности $E_{\Omega-,V-}$.

В первом из выделенных случаев представляется естественным считать, что с точки зрения решаемой задачи о прогнозе свойств новых объектов оба названных класса эквивалентности $E_{\Omega+,V+}$ и $E_{\Omega-,V-}$ являются носителями некоторой структурной «патологии»[33] (ведь соответствующий структурный фрагмент $V = V^+ = V^-$ оказывается одновременно и *причиной* и *антипричиной*[34] соответствующего множества свойств $\mathbf{A}_V$) и, следовательно, должны быть исключены из представленной выше процедуры прогнозирования.

Во втором случае от прогноза свойств нового объекта C_0 отнесением его к классам $E_{\Omega+,V+}$ и $E_{\Omega-,V-}$ также следует отказаться (либо признать такой прогноз «противоречивым»[35], ведь достаточными основаниями отнести C_0 только к расширению множества примеров Ω^+ или же только к расширению множества контрпримеров Ω^- мы в данной ситуации не располагаем).

РАЗДЕЛ IV.

Теперь перейдем к анализу ситуации, когда предложенный для прогноза свойств новый объект C_0 соотносим сразу с несколькими порождаемыми на исходном множестве примеров Ω классами эквивалентности *одного знака*. Здесь, если (в случае работы исключительно с объектами знака α[36]) выделены все классы эквивалентности $E^1_{\Omega\alpha V1,\alpha}$, $E^2_{\Omega\alpha V2,\alpha}$, … $E^k_{\Omega\alpha Vk,\alpha}$, соотносимые с заданным C_0 (т.е. все соответствующие $V_{1,\alpha}$, $V_{2,\alpha}$, … , $V_{k,\alpha}$ содержатся как фрагменты в C_0), то множество свойств $\mathbf{A}_0$ этого анализируемого объекта C_0 представляется естественным сформировать как объединение множеств свойств $\mathbf{A}_{V1,\alpha}$, $\mathbf{A}_{V2,\alpha}$, … ,$\mathbf{A}_{Vk,\alpha}$, каузально обусловленных их структурными носителями $V_{1,\alpha}$, $V_{2,\alpha}$, … , $V_{k,\alpha}$:

$$\mathbf{A}_0 = \mathbf{A}_{V1,\alpha} \cup \mathbf{A}_{V2,\alpha} \cup \ldots \cup \mathbf{A}_{Vk,\alpha}$$

При этом (как мы уже договорились выше в разделе **III**), если в C_0 окажутся вкладываемыми «структурные носители» разных знаков (т.е. как причины, так и антипричины наличия соответствующих свойств), то представляется естественным либо вообще отказаться от прогноза свойств нового объекта C_0 на имеющейся сходной выборке примеров Ω, либо объявить результаты такого прогноза вновь установленным *эмпирическим противоречием*.

Несложно убедиться, что

- представленный в Разделе **III**a) случай в рамках стандартной конструкции ДСМ-метода (см., например, [8] и др.) воспроизводится в структуре ДСМ-правил правдоподобного вывода первого рода (так называемых ППВ-I), где непротиворечивость порождаемой эмпирической зависимости обусловлена выполнимостью используемого решающего предиката лишь одного знака. В противном случае – это либо порождение эмпирического противоречия, либо отказ от прогноза (сохранение недоопределенности);
- представленные в Разделах **III**b) и **IV** случаи воспроизводятся в структуре ДСМ-ППВ второго рода (ППВ-II).

Зафиксируем и эти два факта, после чего продолжим двигаться дальше.

РАЗДЕЛ V.

Обратимся к анализу ситуации, когда исходное множество Ω содержит как *примеры*, так и *контрпримеры*: $\Omega = \Omega^+ \cup \Omega^-$. Попробуем найти возможности выделять «патологии» представленных в разделах **III**a) и **III**b) типов без исчерпывающего формирования из исходного множества примеров Ω всех возможных классов эквивалентности рассматриваемого вида.

Прежде всего, расширим множество $\mathbf{E}_\Omega$ всех порождаемых из каждого Ω классов эквивалентности ровно $|\Omega|$новыми (одноэлементными) классами $E_{\{C1\}}$, $E_{\{C2\}}$, …, $E_{\{Cn\}}$, (где $n=|\Omega|$ – число элементов в Ω), считая по определению, что для каждого C_i из множества Ω $\{C_i\}$ есть *вырожденный* (одноэлементный) класс эквивалентности. Тогда условие **III**a) может быть упрощено до вида:

(7) ни один из примеров не должен содержать в себе ни один из порождающих классы эквивалентности на контрпримерах «структурных носителей» анализируемых свойств. Симметричное условие должно связывать также контрпримеры с порождаемыми на множестве примеров «структурными носителями» анализируемых свойств.

Легко видеть, что на языке, используемых в ДСМ-методе Правил Правдоподобного Вывода I рода это будет (см., например, [8] и др.) условие так называемого *запрета на контрпримеры*.

Наконец, подобными только что приведенным выше для случая *запрета на контрпримеры* вариациями «взаимодействия» позитивных и негативных классов эквивалентности несложно воспроизвести «логику» формирования процедур ИАД, являющихся аналогами и для остальных входящих в состав ДСМ-метода решающих предикатов для ППВ-I. При этом, как и ранее, *операционная структура* таких процедур (т.е. своего рода *базовая логика* исполнения ИАД) является общей для различных типов представления анализируемых данных а также для различных уточнений операции сходства $\otimes$. В части же *достаточности оснований* для принятия порождаемых в ходе исполняемого ИАД результатов – прогнозов свойств новых объектов – базовым элементом используемой конструкции оказывается во всех рассматриваемых случаях *непротиворечивое* (т.е. не по-

---

[33] Например, в части своих выразительных (дескриптивных) возможностей выбранный язык представления данных может оказаться в конкретном анализируемом случае подобного типа на столько «бедным», что его средствами *структурные различия* (по крайней мере некоторых)*причин* и *антипричин* могут оказаться здесь попросту *не представимыми*.

[34] Т.е. причиной отсутствия.

[35] Разумеется, понимая это противоречие не как *логическое*, а как новый *эмпирический факт*.

[36] Где α есть либо +, либо -.

рождающее «конфликтов» при «*позиционировании*» конкретного объекта *относительно* релевантных ему *классов эквивалентности*) отнесение соответствующих объектов к классам эквивалентности, покрывающим исходное множество примеров Ω (а с ним – и порождаемые соответствующими отношениями сходства ⊗ пространства толерантности <Ω,⊗>).

* * *

Итак, нами продемонстрирована достаточно общая схема *контроля корректности* осуществления ИАД, формализованного средствами ДСМ-метода. Представленная схема предлагает собственное *корректное* решение вопроса о существовании *достаточных оснований* для принятия результатов выполняемого ИАД. Принципиально важные характерные особенности данной схемы– это:

1) возможности при восстановлении эмпирических зависимостей, содержащихся в неявном виде в накапливаемых экспериментальных данных, использовать *семейство однородных* (выстроенных в *единой «логике»* анализа данных) *вычислительных процедур*;

2) возможности *варьировать* в ходе осуществляемого ИАД (т.е. входе применения предложенного семейства однородных вычислительных процедур) «*выразительные возможности» средств представления* изучаемых *данных* (в том числе – возможность «перемещаться» при выборе подходящих дескриптивных средств используемого языка представления данных от простейших – быстро обрабатываемых полиномиально сложными алгоритмами, до весьма детальных, – однако, требующих исчерпывающего и, как правило, экспоненциально сложного перебора вариантов),

3) при этом *выделять* и *прослеживать* (например, на последовательно формируемых расширениях новыми объектами исходной обучающей выборки примеров) соответствующие *инварианты* – те (*устойчиво сохраняющиеся* – т.е. порождаемые средствами ДСМ-решателя задач) классы эквивалентности и сопоставленные им *эмпирические зависимости*, которые (в силу их устойчивого «поведения»[37]) можно рассматривать как обнаруженные в имеющихся данных *эмпирические закономерности* (см. подробнее [8, 10] и др.).

СПИСОК ЛИТЕРАТУРЫ

1. Интеллектуальный анализ данных. – Википедия. – URL: http://ru.wikipedia.org/wiki/Data_mining
2. Predictive Analytics and Data Mining – URL: http://www.sas.com/technologies/analytics/datamining/index.html
3. Забежайло М.И., Синякова Е.В. К вопросу об «интеллектуальности» интеллектуального анализа данных // Научно-техническая информация. Сер. 2. – 2013. – № 3. – С. 1-9.
4. Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть I // Кибернетика. – 1977. – № 4. – С. 5-17.
5. Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть II // Кибернетика. – 1977. – № 6. – С. 21-27.
6. Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть III // Кибернетика. – 1978. – № 2. – С. 35-43.
7. Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // Известия АН СССР. Сер. Техническая Кибернетика. – 1987. – № 5. – С.42-63.
8. Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // Искусственный интеллект и принятие решений. – 2010. – Часть I: №3, С.3 -21; Часть II: № 4, С. 14-40.
9. Автоматическое порождение гипотез в интеллектуальных системах / ред. В.К.Финн. – М.: Либроком, 2009. – 528 с.
10. Финн В.К. Об определении эмпирических закономерностей посредством ДСМ-метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. – 2010. – №4. – С.41-48.
11. Шейдер Ю.А. Равенство, сходство, порядок. – М: Наука, 1971. – 255 с.
12. Забежайло М.И. О некоторых оценках сложности вычисления в ДСМ-методе // Искусственный интеллект и принятие решений. – 2014 (в печати)
13. Гэри М., Джонсон Д.С. Вычислительные машины и трудно-решаемые задачи. – М.: Мир, 1982. – 416 с.
14. Simon J. On the difference between one and many // Lect. Not. Comp. Sci. –1977. –Vol. 52. – P. 480-491.
15. Valiant L.G. The complexity of enumeration and reliability problems // SIAM J.Comput. – 1979. – Vol. 8, № 1. – P. 410-421.
16. Valiant L.G. The complexity of computing the permanent // Theoretical Computer Science. – 1979. – № 8. – P. 189-201.
17. Милль Д.С. Система логики силлогистической и индуктивной. – М.; Книжное дело, 1900. – 781 с.

Материал поступил в редакцию 14.10.14.

Сведения об авторе

ЗАБЕЖАЙЛО Михаил Иванович – кандидат физико-математических наук, старший научный сотрудник, Управляющий директор НП «Центр прикладных исследований компьютерных сетей» (Москва, Сколково). e-mail: zmivan@gmail.com

[37] Т.е. в силу сохранения («*стабильности*») выявленной «структуры» классов эквивалентности при расширении исходной выборки примеров (и контрпримеров) новыми объектами.

УДК 004.89

А.С. Клещев, М.Ю. Черняховская, Е.А. Шалфеева

Особенности автоматизации интеллектуальной деятельности*

Рассматривается поиск путей преодоления проблем практического применения разрабатываемых интеллектуальных программных систем. Анализируется организация повседневной интеллектуальной деятельности и процесс управления ее качеством, направленный на контроль принимаемых решений, уточнение и расширение используемых знаний. Описывается спектр трудозатрат на автоматизацию интеллектуальной деятельности, анализируются существующие парадигмы автоматизации интеллектуальной деятельности и предлагается новая парадигма.

Ключевые слова: *интеллектуальная деятельность, поддержка принятия решений, экспертная система, базы знаний, качество знаний, оценка базы знаний, системная инженерия, программное обеспечение, трудозатраты на автоматизацию*

Появление в 1970-х гг. интеллектуальных систем, или систем, основанных на знаниях, породило большие надежды, связанные с их практическим применением. Однако за прошедшие 40 лет, несмотря на значительные успехи в области теории и технологии создания таких систем, они не получили заметного практического применения.

Цель настоящей работы – анализ причин сложившейся ситуации и поиск путей ее преодоления за счет изменения парадигмы автоматизации интеллектуальной деятельности.

ОСОБЕННОСТИ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Основные понятия

Под *интеллектуальной деятельностью* далее будет пониматься деятельность, состоящая в *принятии* взаимосвязанных *решений на основе знаний* по отношению к некоторым объектам действительности. Специфика принятия решений на основе знаний состоит в том, что алгоритмы такого принятия решений неизвестны; известны лишь алгоритмы «применения знаний» для принятия решений, при этом «качество» (правильность и точность) решений зависит от «качества» знаний.

От исполнителей интеллектуальных видов деятельности требуется обладание необходимыми знаниями (зачастую постоянно обновляемыми) и умение их применять. Чем более правильными и обширными являются знания и чем более правильно они применяются, тем более качественным может быть результат, но тем более сложно его получение.

На рис. 1 показаны примеры принятия разных типов решений: сбор информации, диагностика, ремонт, прогноз, и т.п.

Для иллюстрирования особенностей интеллектуальной деятельности далее будет рассматриваться медицина как типичный пример предметной области с наиболее сложными и ответственными задачами принятия разных типов решений. Объектом деятельности в этой предметной области является пациент. После первичного сбора информации о нем врач принимает решение о диагнозе, планирует лечение, прогнозирует изменение состояния и планирует дальнейшее наблюдение за ним. Если результаты такого наблюдения не соответствуют прогнозу, принимаются решения о коррекции диагноза и/или плана лечения, что ведет к корректировке прогноза и т.д.

Качество интеллектуальной деятельности и управление им

Поскольку полезность интеллектуальной деятельности определяется ее качеством, то параллельно с выполнением интеллектуальной деятельности, должно осуществляться управление ее качеством.

Качество интеллектуальной деятельности определяется величиной риска ошибочных решений и характером ущерба от их последствий. Правильность решений определяется степенью надежности процесса применения знаний. Эта степень надежности зависит от нескольких факторов – *правильности знаний* (как часто знания приводят к правильным решениям при правильном их применении), *точности знаний* (как часто знания приводят к однозначным решениям при правильном их применении), полноты индивидуальных знаний (насколько часто лица, принимающие решения, владеют знаниями, необходимыми для решения конкретных задач интеллектуальной деятельности), правильности применения знаний (насколько

* Работа выполнена при финансовой поддержке РФФИ (проекты 15-07-03193 и 14-07-00270-а).

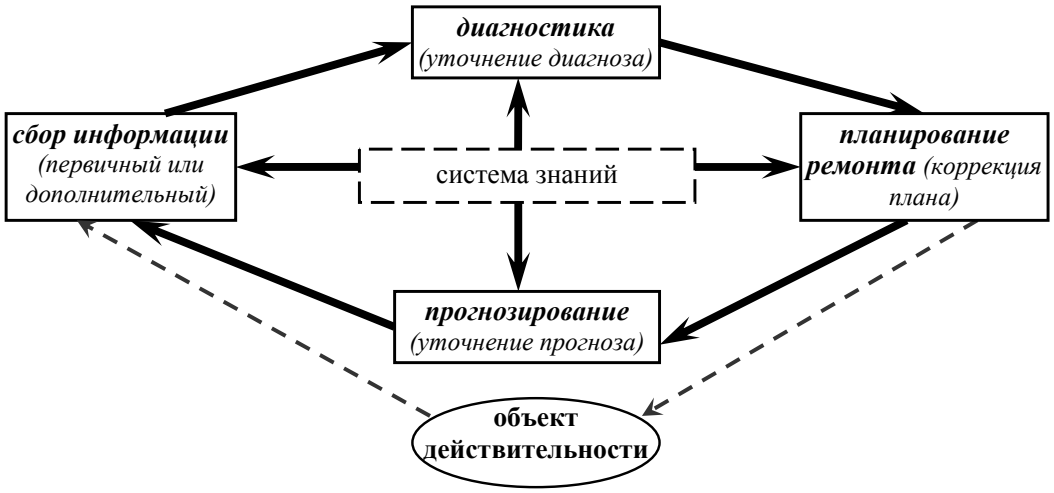

Рис. 1. Схема типичной интеллектуальной деятельности:
сплошными прямоугольниками показаны задачи принятия решений; пунктирными прямоугольниками – необходимые знания; овалом – объект; передача информации – сплошными стрелками; пунктиром – передача информации, связанная с непосредственным взаимодействием с объектом.

часто правильные знания приводят к правильным решениям), точности применения знаний (насколько часто точные знания приводят к однозначным решениям). Стоит отметить, что индивидуальная степень надежности применения знаний определяется систематическими и случайными ошибками. Но вне зависимости от индивидуумов, *качество интеллектуальной деятельности* определяется качеством используемых знаний. Поэтому *процесс управления качеством* состоит, прежде всего, в повышении *качества* (*правильности* и *точности*) *знаний*, используемых специалистами в своей деятельности.

Далее будет рассматриваться только интеллектуальная деятельность, которая не связана с получением прибыли, финансируемая государством. Примером является государственная (бесплатная) медицина.

ОРГАНИЗАЦИЯ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Интеллектуальная деятельность сосредоточена в первичном звене отрасли, где каждый специалист решает всю совокупность задач принятия решений (соответствующую профилю[1] этого звена) относительно объектов действительности, порученных этому специалисту. В медицине примером первичного звена может служить отделение больницы.

Основное звено отрасли объединяет несколько первичных звеньев разного профиля и решает все задачи принятия решений, соответствующие профилю этого основного звена, относительно объектов действительности, находящихся в зоне ответственности этого основного звена, а также включает специализированное первичное звено, которое распределяет все объекты действительности, по отношению к которым должна выполняться интеллектуальная деятельность, по его первичным звеньям. В медицине примером основного звена может служить больница, а специализированного первичного звена в нем – приемное отделение.

Основное звено широкого профиля отрасли решает все задачи отрасли относительно объектов действительности, находящихся в зоне ответственности этого основного звена. Для этого оно содержит набор первичных звеньев всех профилей. В медицине примером основного звена может служить больница широкого профиля. В звеньях широкого профиля решаются задачи любой степени сложности (от простых, типичных – до сложных).

Узкоспециализированное высокотехнологичное основное звено отрасли имеет более узкий профиль, чем основное звено широкого профиля, но более широкую зону ответственности. Как правило, узкоспециализированное высокотехнологичное звено имеет более квалифицированных специалистов, чем основное звено широкого профиля, более высокотехнологичное оборудование и предназначено для решения наиболее трудных задач интеллектуальной деятельности. В медицине примером узкоспециализированного высокотехнологичного звена может служить специализированный НИИ с клиниками и специализированный медицинский центр.

Интеллектуальная деятельность в отрасли выполняется на основе общих знаний – все первичные звенья одного и того же профиля (как основных звеньев «широкого профиля», так и узкоспециализированных высокотехнологичных основных звеньев) при принятии решений должны использовать одни и те же современные знания. Интеллектуальная деятельность и решаемые в ней задачи имеют типовой для отрасли

[1] Профилем здесь считаем раздел в предметной области, например, в медицине – офтальмология, урология и т.д.

характер – характер деятельности всех первичных звеньев в отрасли и решаемые ими задачи различаются лишь профилем первичного звена и зоной ответственности основного звена.

Таким образом, структура отрасли обычно включает два уровня: *первичные* звенья и *основные* звенья, возможно, распределенные по территории.

УПРАВЛЕНИЕ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТЬЮ

Естественным критерием качества интеллектуальной деятельности в отрасли является доля правильных решений во всей отрасли за определенный отрезок времени. Этот же критерий может относиться и к основным, и к первичным звеньям отрасли, и к отдельным специалистам. Цель управления интеллектуальной деятельностью в отрасли – повышение этого показателя, т.е. снижение доли ошибок. Для этого государство предписывает выполнение ряда мероприятий во всех звеньях отрасли: управление текущей деятельностью, контроль качества текущей деятельности, управление знаниями и доведение современных знаний до исполнителей.

Управление текущей деятельностью осуществляется территориальными органами управления отраслью, а также руководством основных и первичных ее звеньев. Среди прочего, важными мероприятиями управления текущей деятельностью, направленными на повышение ее качества, являются организация консультаций ведущих специалистов, либо мозговых штурмов (консилиумов) с участием всех имеющихся специалистов по решению задач интеллектуальной деятельности, вызвавших определенные трудности, а также передача наиболее трудных задач из звеньев широкого профиля в высокотехнологичные специализированные звенья.

Контроль качества текущей деятельности осуществляется центральными органами управления отраслью, а также территориальными органами управления отраслью и руководством основных и первичных ее звеньев. Для оценки текущего уровня качества интеллектуальной деятельности и управления этим качеством обычно используются различного рода отчеты о результатах деятельности, включающие решения, принимаемые специалистом в процессе своей деятельности, и заключительные (подтвержденные) решения.

Деятельность, связанная с управлением знаниями (расширением, уточнением, совершенствованием знаний) относится к области отраслевой науки и выполняется научными организациями отрасли, которые могут быть одновременно и специализированными высокотехнологичными основными звеньями, и отраслевыми высшими учебными заведениями, а также отдельными специалистами основных звеньев.

Доведение современных знаний до исполнителей интеллектуальной деятельности осуществляется различными организациями. Отраслевые высшие учебные заведения ведут базовую подготовку специалистов отрасли. Кроме того, специалисты отраслевых научных учреждений и высокотехнологичных специализированных основных звеньев проводят повышение квалификации специалистов отрасли.

Таким образом, существующие в интеллектуальной деятельности управленческие меры направлены на *уточнение знаний* (по мере выявления в них неточностей), на *контроль принимаемых решений* (выявление случаев неверных решений, чтобы разобраться в их причинах), на *расширение знаний* (обязательное внедрение новых апробированных научных достижений) и на *обучение* новых специалистов с учетом обновляемых знаний.

Вместе с тем, современные системы управления качеством интеллектуальной деятельности в отраслях, основанные преимущественно на «бумажных» технологиях обработки информации, сталкиваются с целым рядом проблем. Исполнители интеллектуальной деятельности обладают различными «человеческими» особенностями, некоторые из них могут негативно влиять на качество их решений. Консультации ведущих специалистов проводятся выборочно, и даже для тех задач, где они необходимы, оказываются не всегда возможными. Кроме того, их полезность зависит от уровня компетентности консультанта. Мозговые штурмы (консилиумы) не всегда анализируют все возможные решения задач интеллектуальной деятельности. Поиск необходимой информации в первичных «бумажных» документах, как правило, довольно труден. Статистические отчеты позволяют лишь оценить уровень качества интеллектуальной деятельности, но не способствуют его улучшению. Знания, которые студенты получают в отраслевых высших учебных заведениях, часто далеки от современных знаний, используемых в специализированных высокотехнологичных основных звеньях. Традиции и средства обучения на современном этапе позволяют осуществлять лишь выборочный контроль знаний выпускаемых специалистов, что не гарантирует усвоение всего требуемого объема знаний выпускниками вузов. Новейшие достижения в знаниях, необходимых для решения задач интеллектуальной деятельности, иногда слишком долго доходят до исполнителей этой деятельности. Еще более трудный вопрос – в какой мере специалисты пользуются этими достижениями в своей повседневной работе? Наконец, проблема, которая не может быть решена при использовании «ручных» и «бумажных» технологий, – это правильность применения знаний при решении задач интеллектуальной деятельности: алгоритм применения знаний слишком сложен, поэтому знания применяются неточно, приблизительно.

ИНСТРУМЕНТЫ И ТЕХНОЛОГИИ АВТОМАТИЗАЦИИ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Одним из средств повышения качества профессиональной деятельности, в том числе и интеллектуальной, является ее автоматизация с помощью компьютеров. К настоящему времени создан целый ряд инструментов, которые могут использоваться (и во многих случаях используются) для разработки систем автоматизации различных видов деятельности: СУБД, системы программирования (CASE, IDE), локальные сети, технология облачных вычислений, средства создания корпоративных систем.

Цель автоматизации интеллектуальной деятельности – повышение качества принятия ответственных решений за счет информационной поддержки этого процесса. Эта поддержка может состоять в автоматической выработке рекомендаций (в том числе возможных альтернативных решений) интеллектуальными системами на основе баз знаний (БЗ) для исполнителя и в построении объяснений таких рекомендаций. Автоматизация имеет следующее ограничение – организационная структура не может перестраиваться под автоматизацию, поэтому автоматизация должна встраиваться в организационную структуру и поддерживать ее. Достижение этой цели связано, во-первых, с *поддержкой повседневной деятельности* сотрудников, встраиваемой в организационную структуру учреждения/предприятия или даже отрасли, во-вторых, с *поддержкой процесса управления качеством* этой деятельности.

Поддержка интеллектуальной деятельности

Для автоматизации интеллектуальной деятельности разрабатываются экспертные системы (ЭС), которые предназначены для решения задач интеллектуальной деятельности некоторых классов на основе баз знаний, как правило, сформированных экспертами [1,2]. Результатом работы экспертной системы является *объяснение*, которое содержит информацию о соответствии гипотез о вариантах решения задачи информации об объекте (входным данным) и базе знаний. Например, в медицине важно объяснение того, какие гипотезы-диагнозы могут быть отвергнуты, а какие гипотезы могут быть приняты для конкретной истории болезни с учетом конкретных знаний о медицинской диагностике [3]. С помощью экспертных систем осуществляется *поддержка принятия решений*, что означает принятие специалистом собственного решения с учетом объяснения, генерируемого экспертной системой (см. рис. 2).

Объяснение экспертной системы в определенной степени может рассматриваться как аналог консультации. И в том, и в другом случае проводится независимый анализ входных данных на соответствие их тем или иным гипотезам о решении задачи в свете знаний консультанта или экспертной системы. При этом ЭС применяет базу знаний правильно (использует правильный алгоритм решения задачи) и может провести полный анализ всех гипотез, чего нельзя требовать от консультации. Однако результаты анализа, полученные консультантом и экспертной системой, в значительной степени зависят от качества применяемых в этом анализе знаний. Если база знаний имеет высокое качество, то ЭС позволяет снизить долю ошибок специалистов из-за неправильного применения знаний.

Качество знаний

Созданная экспертом база знаний может содержать дефекты – быть неполной (некоторые варианты могут быть упущены экспертом или ему неизвестны), неточной (приводить к неоднозначным решениям) или даже неправильной (из-за заблуждений, предубеждений). Качество и полезность базы знаний определяются полнотой, точностью и правильностью содержащихся в ней знаний. Однако очевидна необходимость более объективного оценивания качества баз знаний, поскольку из литературы следует, что до сих пор основными средствами оценивания баз знаний (БЗ) являются средства контроля формальных свойств правильно построенной базы знаний и привлечение экспертов для оценки «решений, предлагаемых системой» [4-6].

Каждое решение, принимаемое специалистом в ходе интеллектуальной деятельности, обычно в дальнейшем проходит процедуру подтверждения, которая может потребовать определенного времени. В результате этой процедуры решение специалиста (или системы) признается правильным или ошибочным. В медицине подтверждение осуществляется по результату выздоровления пациента в результате лечения, либо по результатам оперативного вмешательства, либо решением судебно-медицинских экспертов.

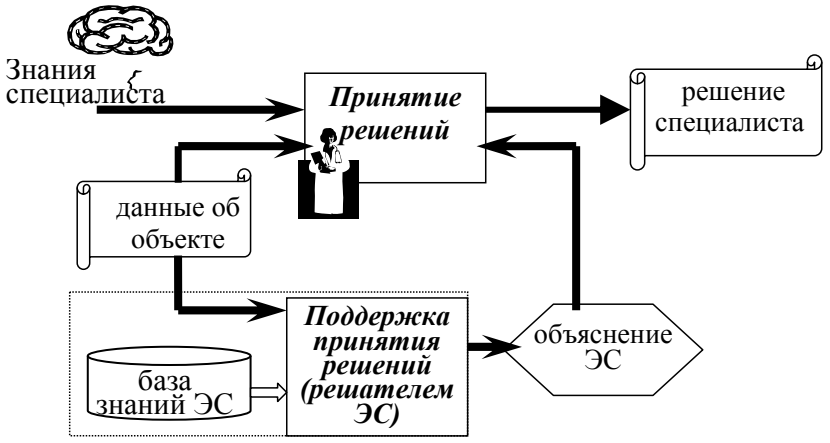

Рис.2 Схема принятия решений с поддержкой экспертной системы:
прямоугольником с ярлычком (здесь и далее) показана деятельность человека;
шестиугольником – сгенерированное системой объяснение решения; цилиндром –хранимая в системе информация;
«папирусами» – документируемая информация.

Оценка правильности базы знаний может определяться множеством задач, которые ЭС правильно решает на основе этой БЗ. *Оценка точности БЗ* может определяться подмножеством этого множества, для которого решения ЭС однозначны.

Если *оценкой знаний специалиста* считать множество задач, для которых он знает правильное решение, то ЭС можно считать *полезной для некоторого специалиста*, если оценка точности ее БЗ лучше оценки знаний специалиста, т.е. *множество задач, решение которых известно этому специалисту,* является *подмножеством множества точно решаемых экспертной системой задач*. То же справедливо и для группы специалистов: ЭС можно считать *полезной для группы специалистов*, если множество задач, решение которых известно хотя бы одному специалисту из этой группы, является подмножеством множества точно решаемых экспертной системой задач. Определяемая таким способом оценка *правильности и точности базы знаний* зависит от *множества задач с известным решением (базы всех прецедентов)*.

Поддержка управленческих мер

Необходимость изменения используемых знаний на основе новых научных исследований и опыта практической работы требует автоматизации контроля их качества. Если база знаний со временем не будет совершенствоваться, то она будет устаревать в том смысле, что *оценка* ее *правильности и точности* не будет изменяться, в то время как *база прецедентов* (множество задач с известным решением) будет расширяться в процессе дальнейшей практики. Поэтому ЭС должна иметь *систему управления* ее *БЗ* [7], цель этой системы управления – обеспечивать полезность ЭС для специалиста (группы специалистов) в течение всего времени ее эксплуатации.

Для достижения этой цели система управления БЗ должна выполнять следующие функции:

- Накапливать базу прецедентов.
- Классифицировать все прецеденты в базе.
- Находить все возможные способы модификации БЗ для включения новых прецедентов в оценку ее правильности и точности.
- Модифицировать БЗ одним из таких возможных способов.

Накопление базы прецедентов возможно за счет включения ЭС в электронный документооборот интеллектуальной деятельности. В этом случае накопление базы прецедентов не потребует от специалистов дополнительных усилий в их повседневной интеллектуальной деятельности. Кроме того, такой электронный документооборот позволит получать объективные оценки качества решения задач (увеличивающаяся в процессе практики специалистов *база прецедентов* используется для автоматизированной проверки того, насколько БЗ или вариант ее модификации удовлетворяет накопленным прецедентам).

Каждый прецедент должен быть отнесен системой управления БЗ к одному из следующих классов решений экспертной системы:

1) правильное и точное решение;
2) правильное, но неточное решение (несколько возможных альтернатив, среди которых было и правильное решение), однако входные данные задачи допускают ее точное решение;
3) правильное, но неточное решение (среди альтернатив было и правильное решение), но входные данные задачи допускают некоторое его уточнение (уменьшение числа альтернатив);
4) правильное, но неточное решение (среди альтернатив было и правильное решение), однако входные данные задачи не допускают его уточнения;
5) неправильное решение (множество альтернатив, возможно пустое, среди которых не было правильного решения).

Функция классификации прецедентов состоит в отнесении каждого прецедента к одному из указанных классов. Решение об отнесении прецедента к классам 2-4 не может быть принято автоматически, поэтому оно должно приниматься экспертами, входящими в группу управления БЗ (рис.3).

Прецеденты, отнесенные к классам 1 и 4, образуют *оценку правильности и точности БЗ*. Новые прецеденты, отнесенные к этим классам, могут быть включены в эту оценку без модификации БЗ, в отличие от новых прецедентов, отнесенных к классам 2, 3 и 5. Допустимой является такая модификация БЗ, которая не ухудшает ее оценку, т.е. классы, к которым отнесены входящие в нее прецеденты после допустимой модификации БЗ, не изменяются, или некоторые прецеденты из класса 4 переходят в класс 1. Новые прецеденты из классов 2 и 3 требуют уточнения БЗ, т.е. такой ее допустимой модификации, при которой прецеденты из класса 2 переходят в класс 1, а прецеденты из класса 3 переходят в классы 1 или 4; уточнение БЗ имеет целью включить эти прецеденты в оценку правильности и точности уточненной БЗ. Новые прецеденты из класса 5 требуют исправления или расширения БЗ, т.е. такой ее допустимой модификации, при которой прецеденты из класса 5 переходят в классы 1 или 4; исправление или расширение БЗ также имеет целью включить эти прецеденты в оценку правильности и точности исправленной или расширенной БЗ. В поиске всех возможных вариантов таких допустимых модификаций БЗ для всех новых прецедентов и состоит функция поиска возможностей включения новых прецедентов в оценку БЗ. Эти варианты допустимых модификаций БЗ должны вычисляться системой управления БЗ автоматически.

В результате выполнения предыдущей функции может быть получено несколько вариантов допустимых модификаций базы знаний для новой группы прецедентов, либо такие варианты могут вообще отсутствовать. Поэтому функция модификации БЗ состоит в выборе одного такого варианта допустимой модификации (если они есть) и его выполнения, либо в пересмотре некоторых ранее принятых решений при модификации БЗ (если таких вариантов нет). Реализация этой функции должна осуществляться экспертами, входящими в группу управления БЗ при поддержке системы управления. От системы автоматизации требуется поддержка получения формализо-

ванных новых знаний (рис. 4): удобные для эксперта средства формирования обучающей выборки из прецедентов классов 2 и 3, средства автоматического формирования очередного варианта модификации БЗ и его оценивания. Удовлетворительный результат означает, что получен вариант БЗ с оценкой не хуже оценки специалиста (а при обобщении подхода к автоматизации на группу специалистов оценка единой БЗ должна становиться не хуже совокупной оценки специалистов), поэтому он становится новой базой знаний экспертной системы (вместо «текущей БЗ», т.е. используемой на этот момент).

Дополнительным источником совершенствования знаний, используемых при решении задач интеллектуальной деятельности, являются новые научные результаты, относящиеся к этой интеллектуальной деятельности. Естественно, что система управления БЗ должна допускать включение в неё новых научных результатов (без ухудшения ее оценки). Такая модификация может выполняться только экспертами, входящими в группу управления БЗ. Если *оценка модифицированной БЗ* становится не хуже при включении в нее новых научных знаний, то такой вариант модификации становится новой базой знаний экспертной системы.

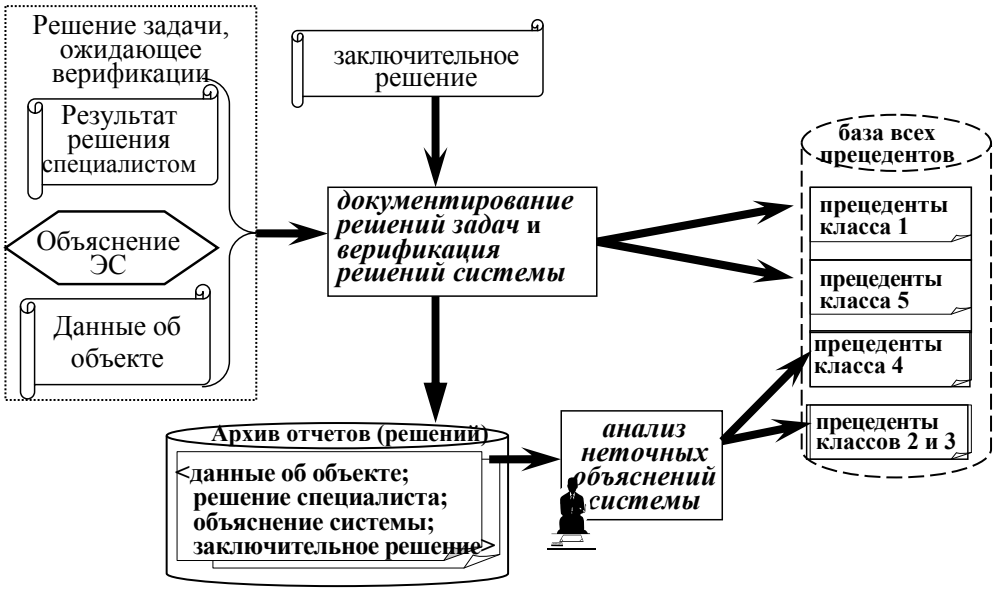

Рис. 3. Схема связи документооборота с поддержкой системы управления

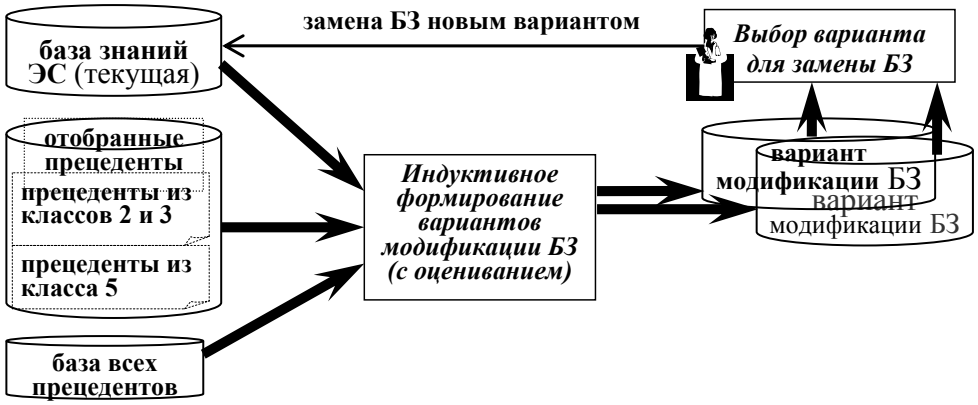

Рис. 4. Поддержка автоматического управления качеством баз знаний

Трудно надеяться, что поддержка принятия решений с помощью ЭС для задач лишь одного класса (например, поддержка диагностики без поддержки планирования лечения или ремонта) может существенно улучшить качество всей интеллектуальной деятельности. Поэтому естественным развитием технологии автоматизации интеллектуальной деятельности являются семейства ЭС: для каждой задачи интеллектуальной деятельности разрабатывается своя система; разные ЭС семейства связываются по входной/выходной информации, а также могут иметь некоторые общие базы знаний или их части или другие информационные ресурсы. Электронный документооборот, о котором выше шла речь, также имеет смысл интегрировать с семейством ЭС.

Семейства компьютерных тренажеров

При обучении решению задач интеллектуальной деятельности, контролю знаний студентов и умений их применять могут использоваться семейства интеллектуальных программных систем, называемых интеллектуальными тренажерами [8]. Задача такого тренажера – генерация виртуального объекта интеллектуальной деятельности, предоставление студенту электронного документооборота, позволяющего ему решать все задачи интеллектуальной деятельности над этим виртуальным объектом, получение оценки правильности принятых студентом решений и, в случае, если среди этих решений имеются неверные, то объяснение причин его ошибок, либо выдача полного объяснения.

Например, медицинский тренажер для очередного сеанса обучения генерирует виртуальный объект «пациент», в частности, все значения признаков, которые необходимы для диагностики, лечения и последующих наблюдений (рис. 5). В тренажерах для обучения принятию решений студент «видит» объект, должен выявить значения важных признаков и поставить диагноз. При обнаружении неверных решений тренажер указывает на них и объясняет студенту причины ошибок (например, визуализирует анализ неверной гипотезы).

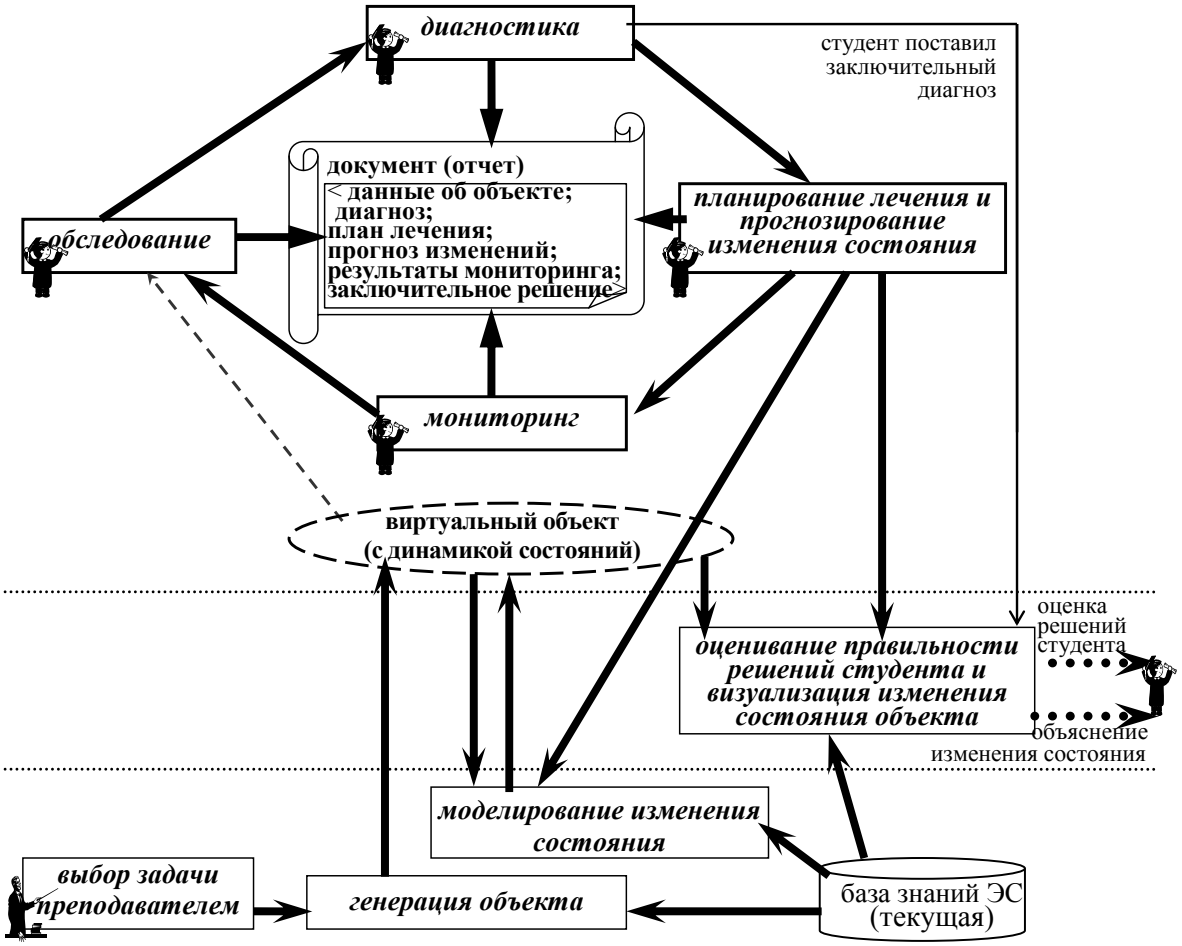

Рис. 5. Схема процесса обучения принятию решений с помощью интеллектуального тренажера

Таким образом, тренажеры моделируют объект, процесс, и выполняют анализ ошибок. Но для этого база знаний должна быть такой, какую рекомендовано использовать в реальной практике, т.е. высокого качества. Применение семейств интеллектуальных тренажеров позволяет уменьшать использование реальных объектов в процессе обучения, полнее проверять знания студентов и умение их применять.

Преимущества автоматизации интеллектуальной деятельности и управления знаниями

Автоматизация интеллектуальной деятельности и управления базами знаний имеет преимущества как для рядовых специалистов, так и для руководства.

Преимущество автоматизации в повседневной интеллектуальной деятельности:

- специалистам доступен результат полного анализа всех входных данных каждой задачи на соответствие их всем существующим гипотезам или возможность рассмотреть полный анализ интересующих гипотез.

Преимущества в управлении качеством знаний:

- получение базой знаний экспертной системы оценки качества знаний не хуже, чем оценка качества знаний специалистов, которые пользуются этой ЭС, и достижение монотонного роста оценки качества базы знаний ЭС;
- фиксация системы знаний с наилучшей оценкой, ее доступность для изучения (в качестве справочника), для использования на практике (в качестве стандарта) и в обучении (в качестве учебного пособия). Наличие такой системы знаний может значительно расширить возможность своевременного доведения новейших знаний до специалистов.

Типичные трудозатраты на автоматизацию интеллектуальной деятельности

При автоматизации отдельной задачи принятия решений на основе знаний для одного профиля требуется разработать один решатель экспертной системы с пользовательским интерфейсом, обычно одну базу знаний и один редактор знаний для нее, а также подсистему оценивания варианта БЗ. Современная система управления БЗ должна включать также автоматизированное рабочее место (АРМ) для анализа прецедентов и формирования обучающей выборки для модификации БЗ и подсистему ее модификации. При построении семейства ЭС с управляемыми БЗ и интеграцией их со средствами документооборота трудозатраты существенно возрастают. Типичные трудозатраты на автоматизацию интеллектуальной деятельности могут быть классифицированы как трудозатраты на *системную инженерию* [9, 10], на *программное обеспечение* и на *базы знаний* (включая управление БЗ) [7], а также на вспомогательные и организационные процессы, например, на процесс создания и сопровождения инфраструктуры (в том числе администрирование сетей).

Трудозатраты на системную инженерию складываются из: системного анализа профессиональной деятельности в предметной области; концептуального проектирования системы автоматизации; документирования требований на разработку всей системы; интеграции всех подсистем и БЗ в единую систему; процессов комплексирования, верификации, технического обслуживания и других процессов, относящихся к системной инженерии [10].

Трудозатраты на программное обеспечение состоят в разработке, а затем в сопровождении следующих видов программных средств: решатель (обычно с пользовательским интерфейсом); редактор знаний (с встроенным контролем внутренних свойств знаний); тренажер; подсистема документирования (со специализированными редакторами баз данных); автоматизированное рабочее место для анализа прецедентов и формирования обучающей выборки для модификации БЗ; подсистема модификации БЗ; подсистема оценивания варианта БЗ.

Трудозатраты на формирование базы знаний складываются из затрат: на разработку первого варианта каждой базы знаний (с помощью редакторов знаний и подсистемы оценивания вариантов); на сопровождение и управление базами знаний (с помощью редакторов знаний, автоматизированных рабочих мест, подсистемы формирования вариантов модификации БЗ и подсистемы оценивания этих вариантов).

Первичной разработкой баз знаний занимаются эксперты по знаниям. Они же продолжают их сопровождение и управление – являются пользователями системы управления качеством знаний в процессе всего периода эксплуатации системы.

Таким образом, автоматизация интеллектуальной деятельности (даже в рамках одного раздела предметной области) связана с типичным спектром трудозатрат на системную инженерию и разработку программного обеспечения, а также с особого рода затратами на разработку баз знаний, поскольку их (БЗ) роль и способ организации отличаются от других хранилищ данных.

СУЩЕСТВУЮЩИЕ ПАРАДИГМЫ АВТОМАТИЗАЦИИ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ И ОЖИДАЕМЫЙ ЭФФЕКТ ОТ ИХ ПРИМЕНЕНИЯ

Цель автоматизации интеллектуальной деятельности – повышение качества этой деятельности, т.е. снижение доли неправильных решений (в отрасли), а также повышение эффективности системы управления качеством такой деятельности (в отрасли).

К настоящему времени можно указать несколько парадигм автоматизации интеллектуальной деятельности, для каждой из которых укажем ожидаемые эффект и затраты.

Автоматизация первичного звена отрасли

Эта парадигма состоит в разработке комплекса ЭС для решения взаимосвязанных задач всех классов первичного звена отрасли. Специалистам первичного звена предоставляется возможность взаимодействовать с решателями ЭС, а также вводить информацию с помощью специализированных редакторов. Экспертам предоставляются средства создания и управления базами знаний. В случае интеграции всех этих подсистем, устанавливаемых на рабочих местах со-

трудников, с подсистемой документооборота руководителям предоставляется возможность знакомиться с результатами работы специалистов и обрабатывать эти результаты.

Ожидаемый эффект от этой парадигмы состоит в получении перечисленных преимуществ (в повседневной деятельности и в совершенствовании знаний) только специалистами первичного звена отрасли.

Однако когда речь идет о совершенствовании знаний специалистами одного первичного звена, следует иметь в виду, что оценка качества БЗ зависит от сложности задач, которые специалистам приходится решать в этом первичном звене. Если чаще всего это несложные задачи (например, в медицинском первичном звене – районной поликлинике), то БЗ будет адаптироваться к решению «типичных», распространенных задач. И, наоборот, в высокотехнологичном первичном звене отрасли БЗ будет адаптироваться к решению преимущественно сложных задач.

Трудозатраты на автоматизацию отдельного первичного звена отрасли включают:

1) все типичные затраты на программное обеспечение (на разработку, тестирование и сопровождение), а именно:

- решатели (и их пользовательские интерфейсы) – n штук (по числу разных интеллектуальных задач),
- подсистему документирования – одна.

Если первичное звено связано с процессом обучения или повышения квалификации, то понадобится один тренажер (с n штук генераторами и подсистемами оценивания);

2) затраты на базы знаний (разработку, оценивание и сопровождение), а именно:

- БЗ – приблизительно тоже n штук,
- редакторы знаний – один универсальный или приблизительно n штук специализированных,
- АРМ для анализа прецедентов и формирования обучающей выборки – n штук,
- подсистему формирования вариантов модификации БЗ – n штук,
- подсистему оценивания варианта БЗ – n штук.

3) все затраты на системную инженерию.

Вывод: такая автоматизация экономически не оправдана, так как затраты на ее разработку измеряются «десятками человеко-лет», затраты на сопровождение также значительны.

Автоматизация основного звена широкого профиля

Эта парадигма состоит в разработке экспертной системы для всех классов задач интеллектуальной деятельности, решаемых в основном звене (например, поликлиники или больницы, как основного звена отрасли). Если все задачи интеллектуальной деятельности в каждом первичном звене основного звена отрасли будут решаться с использованием БЗ, оценки правильности и точности которых выше, чем оценки знаний специалистов, то общий эффект станет заметен.

Аналогично, когда речь идет о совершенствовании знаний специалистами основного звена отрасли, следует иметь в виду, что качество улучшаемых БЗ будет зависеть от сложности задач, которые специалистам приходится решать (в основном звене широкого профиля БЗ будут адаптироваться к решению «типичных», распространенных задач).

Для получения ожидаемого эффекта в рамках такого звена, оно должно «понести» и все необходимые затраты. Трудозатраты на автоматизацию для m профилей деятельности (основного звена отрасли) по сравнению с автоматизацией отдельного первичного звена несколько возрастают, а именно:

- на базы знаний и системы управления ими – их $n * m$ штук,
- на интеграцию n штук решателей с подсистемой документооборота и подсистемами управления качеством $n * m$ баз знаний (в рамках локальной сети учреждения).

При этом с каждым профилем (первичным звеном) связана своя команда управления качеством совокупности баз знаний (m команд), каждая команда занимается первичной разработкой своей совокупности баз знаний и управлением ею в процессе всего периода эксплуатации системы. Для этого привлекаются дополнительные специалисты, поскольку у работающих сотрудников нет свободного времени.

Такая автоматизация экономически не оправдана. Затраты на ее разработку и на сопровождение (по сравнению с первичным звеном) становятся заметно больше, эффект остается таким же, как в первичном звене, только распространен на большее число специалистов (работающих во всех первичных звеньях основного звена).

Автоматизация узкоспециализированного высокотехнологичного звена

Эта парадигма состоит в разработке ЭС для всех классов задач интеллектуальной деятельности, решаемых в узкоспециализированном высокотехнологичном основном звене отрасли. Специалисты именно такого звена могут вносить в БЗ новые научные знания

Ожидаемый эффект от этой парадигмы сводится к преимуществам в совершенствовании знаний специалистами. Преимущества в повседневной деятельности не будут заметны, поскольку качество работы специалистов здесь высокое.

Оценка качества баз знаний, формируемых в узкоспециализированном высокотехнологичном основном звене, отличается от такой оценки в основном звене широкого профиля. В узкоспециализированном высокотехнологичном основном звене БЗ будут адаптироваться к решению преимущественно сложных задач.

Трудозатраты на автоматизацию в этом звене аналогичны трудозатратам на автоматизацию интеллектуальной деятельности основного звена.

Такая автоматизация экономически не оправдана по тем же причинам, что и автоматизация основного звена широкого профиля.

НОВАЯ ПАРАДИГМА АВТОМАТИЗАЦИИ ИНТЕЛЛЕКТУАЛЬНОЙ ДЕЯТЕЛЬНОСТИ

Парадигма автоматизации интеллектуальной деятельности отрасли состоит в разработке единой для этой отрасли интеллектуальной программно-информационной системы для решения всех задач принятия

решений на основе знаний, а также всех задач управления качеством знаний.

«Облачная» реализация такой системы означает, что на центральных серверах отрасли устанавливаются все решатели и все БЗ, подсистема документооборота и подсистема управления, тренажеры, а также единый архив решенных задач. Всем пользователям интеллектуальной программно-информационной системы на своих рабочих местах достаточно иметь доступ в Интернет. Специалисты из основных звеньев отрасли используют «облачные» ЭС, коллектив экспертов (команда управления качеством) каждого профиля управляет качеством баз знаний посредством «облачно» доступных инструментов. Экспертами становятся специально выделенные высококвалифицированные специалисты, соответствующие по уровню квалификации узкоспециализированному высокотехнологичному звену. Управление деятельностью может осуществляться на любом уровне – звена, регионального отделения или всей отрасли.

Ожидаемый эффект от этой парадигмы состоит в получении всех перечисленных преимуществ (в повседневной деятельности и в совершенствовании знаний) в каждом основном звене отрасли. Достижимо наивысшее качество БЗ, так как оно определяется полным спектром решаемых в отрасли задач – от типичных до самых сложных. (Оценки правильности и точности единой БЗ выше, чем оценки знаний любых специалистов отрасли).

Трудозатраты на автоматизацию интеллектуальной деятельности для всей отрасли (для m профилей деятельности в ней) близки к затратам на автоматизацию этой деятельности для основного звена. Здесь ожидается такой же объем работ, как и при автоматизации интеллектуальной деятельности основного звена (но выполняемый в одном «месте», а не во всех звеньях отрасли). Кроме этих затрат, естественно, будут и затраты на создание (и администрирование) общего вычислительного ресурса, объединяющего устройства хранения данных, серверы, сети передачи данных и т.д., на развертывание на центральных серверах отрасли нескольких решателей, баз знаний, комплекса программных средств для управления качеством знаний, других подсистем.

Заметное увеличение трудозатрат будет иметь место при организации распределенного хранилища отчетов, которое должно накапливать отчеты о задачах, решаемых во всех звеньях отрасли. Объем, структура и методы доступа к этому хранилищу должны быть приспособлены к потенциально большой географически распределенной группе пользователей, а также иметь соответствующие средства защиты, поддержки целостности данных, их безопасности и оптимизации.

В этой парадигме становится реально осуществимым появление высококвалифицированных экспертов (из узкоспециализированного высокотехнологичного звена), основной работой которых станет обеспечение качества баз знаний по своему профилю.

При этом возможна поэтапная автоматизация. Например, в случае медицинской деятельности проводить автоматизацию можно поочередно – сначала для одного отдельно взятого профиля, затем для следующего и т.д.

Такая автоматизация экономически целесообразна, так как эффект ощущает вся отрасль, а затраты практически не увеличиваются по сравнению с парадигмой автоматизации основного звена (а в случае поэтапной автоматизации – соизмеримы на каждом этапе с затратами на автоматизацию первичного звена).

* * *

Таким образом, анализ соотношения *затрат* на получаемый специалистами *эффект* показывает, что это соотношение различно в разных парадигмах.

В первичном звене широкого профиля управление БЗ является непосильной задачей: все современные знания вряд ли могут быть внесены в базу знаний экспертом или командой управления качеством *первичного звена*. Это «по силам» лишь высококвалифицированным экспертам высокотехнологичного узкоспециализированного первичного звена.

Так же велики затраты по отношению к ожидаемому эффекту при автоматизации *основного звена*. Внесение во все базы знаний основного звена новых знаний, выявляемых научным сообществом, в каждом основном звене является непосильной задачей (особенно для основных звеньев широкого профиля). И затраты на сопровождение систем могут превысить возможности предприятия (так как интеграция требует много усилий).

Экономически целесообразной по сравнению с ними является парадигма автоматизации *отрасли*. В этом случае затраты на сопровождение программной части сводятся к сопровождению одной программно-информационной системы, используемой по всей отрасли (за счет облачной технологии). То же – с затратами на сопровождение баз знаний, при этом возможность найти (и назначить) лучших экспертов, ответственных за качество знаний на уровне отрасли, становится реальной.

Новая парадигма встраивается в существующую организацию интеллектуальной деятельности и предлагает новые механизмы управления ею, согласованные с существующими в отрасли механизмами (пока осуществляемыми «вручную»). Более эффективное управление возможно только при внедрении единого комплекса ЭС и сопряженных с ними программ для всей отрасли. Это может привести к такому распределению ресурсов в отрасли: узкоспециализированные учреждения смогут стать не только местом принятия лучших решений, но и центром формирования стандарта знаний для отрасли.

СПИСОК ЛИТЕРАТУРЫ

1. Джарратано Д., Райли Г. «Экспертные системы: принципы разработки и программирование», 4-тое изд. – М.: "Вильямс", 2007. – 1152 с.
2. Кобринский Б.А. Ретроспективный анализ медицинских экспертных систем // Новости искусственного интеллекта. – 2005. – №2. – С. 6-17.

3. Клещев А.С., Черняховская М.Ю., Москаленко Ф.М. Модель онтологии предметной области «Медицинская диагностика». Часть 1. Неформальное описание и определение базовых терминов // Научно-техническая информация. Сер. 2. – 2005. – № 12. – С. 1-7.
4. Соловьев С.Ю., Соловьева Г.М. Методы отладки баз знаний в системе ФИАКР // Сб. Автоматизация и роботизация призводства с применением микропроцессорных средств. – Кишинев, 1986. – С.36-37.
5. Тельнов Ю.Ф. Интеллектуальные информационные системы // Московский государственный университет экономики, статистики и информатики. – М.: МЭСИ, 2004. – 246 с.
6. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.
7. Клещев А.С., Грибова В.В. Управление интеллектуальными системами // Известия РАН. Теории и системы управления. – 2010. – № 6. – С. 122-137.
8. Грибова В. В., Федорищев Л.А. Обучающие виртуальные системы и средства их создания // Вестник информационных и компьютерных технологий. – М: «Издательский дом "СПЕКТР"», 2012. – №3. – С. 48-51.
9. ГОСТ Р ИСО/МЭК 15288-2005 Информационная технология. Системная инженерия. Процессы жизненного цикла систем. – М.: Стандартинформ, 2005. – 57 с.
10 Клещев А.С., Шалфеева Е.А. Системный анализ при автоматизации интеллектуальной профессиональной деятельности // XIII Национальная конференция по искусственному интеллекту с международным участием «КИИ-2012», Труды конференции, т.2. – Белгород: Изд-во БГТУ, 2012. – С.128-135.

Материал поступил в редакцию 27.10.14.

Сведения об авторах

КЛЕЩЕВ Александр Сергеевич – доктор физико-математических наук, профессор, главный научный сотрудник Института автоматики и процессов управления Дальневосточного отделения РАН, г. Владивосток
e-mail: kleschev@iacp.dvo.ru

ЧЕРНЯХОВСКАЯ Мери Юзефовна – доктор медицинских наук, главный научный сотрудник Института автоматики и процессов управления Дальневосточного отделения РАН, г. Владивосток
e-mail: chernyah@iacp.dvo.ru

ШАЛФЕЕВА Елена Арефьевна – кандидат технических наук, доцент по специальности, старший научный сотрудник Института автоматики и процессов управления Дальневосточного отделения РАН, г. Владивосток
e-mail: shalf@iacp.dvo.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322

А.Б. Кутузов, Е.А. Кузьменко

Использование корпусных технологий для изучения ошибок: *learner corpora* на факультете филологии НИУ ВШЭ

Представлен проект факультета филологии НИУ ВШЭ по созданию двух корпусов текстов, предоставляющих исследователям лингвистический материал с разнообразными ошибками. Первый корпус, состоящий из академических текстов, написанных студентами НИУ ВШЭ, содержит материал по типам ошибок, совершаемых в английском языке носителями русского языка. Второй корпус, содержащий англо-русские и русско-английские переводы, выполненные студентами переводческих специальностей, предоставляет возможность исследовать ошибки и вариативность в переводе.

Несмотря на широкое распространение корпусов ошибок, для русского языка подобные ресурсы создаются впервые.

Ключевые слова: *учебные корпусы, параллельные корпусы, тесты на основе корпусов, корпусная эрратология*

ТЕХНИЧЕСКИЕ АСПЕКТЫ ОРГАНИЗАЦИИ КОРПУСОВ И РАБОТЫ НАД РАЗМЕТКОЙ

В настоящей статье описывается организация и использование двух корпусов текстов, создаваемых лингвистической лабораторией по корпусным технологиям НИУ ВШЭ в рамках проекта Центра фундаментальных исследований. Это – корпус академических текстов на английском языке, написанных русскоязычными студентами НИУ ВШЭ в рамках курсов по академическому письму (далее Russian Error-Annotated Learner English Corpus, REALEC), и параллельный корпус англо-русских и русско-английских переводов, выполненных студентами переводческих специальностей восьми российских вузов (далее Russian Learner Translator Corpus, RusLTC).

В работе над обоими корпусами принимали участие студенты, преподаватели и научные сотрудники НИУ ВШЭ, в работе над параллельным корпусом участвовали также преподаватели и студенты кафедры перевода и переводоведения Тюменского государственного университета.

Два этих корпуса объединены общей целью: предоставить исследователям лингвистический материал, изобилующий разнообразными ошибками. Создание корпусов ошибок (в англоязычной лингвистике *learner corpora*) получило в последнее время широкое распространение (ср., например, [1]), однако для русского языка подобные ресурсы, доступные широкому пользователю, ещё не создавались.

Параллельный корпус RusLTC расположен в Интернете по адресу http://rus-ltc.org. Корпус REALEC доступен на сайте http://realec.org/.

Тексты, составляющие корпусы, хранятся в формате plain text, но для облегчения непосредственной исследовательской работы с ними они снабжены лингвистической разметкой. Оба корпуса обладают морфологической разметкой, корпус REALEC и часть корпуса RusLTC также размечены по ошибкам в системе онлайн-аннотирования brat [2]. Аннотации хранятся в отдельном файле для каждого текста.

Для поиска переводов по оригиналам и наоборот в параллельном корпусе RusLTC он был снабжен надстройкой в виде двуязычного файла в формате TMX (Translation Memory eXchange), содержащего связанные друг с другом оригиналы и переводы, а также метаинформацию.

ОШИБКИ НОСИТЕЛЕЙ РУССКОГО ЯЗЫКА ПРИ ИЗУЧЕНИИ АНГЛИЙСКОГО: МАТЕРИАЛ КОРПУСА *REALEC*

В сфере исследований Second Language Acquisition (SLA) и в сфере преподавания английского языка носителям русского в настоящее время существует проблема нехватки научных ресурсов, посвящённых процессу освоения иностранного языка и систематизации данных об ошибках, допускаемых в английском языке русскоговорящими студентами. В рамках проекта по созданию корпуса ошибок, инициирован-

ного лингвистической лабораторией по корпусным технологиям НИУ ВШЭ, была предпринята попытка решить эту проблему. Был создан корпус текстов с допущенными ошибками (в настоящее время 794 текста, около 225 тыс. словоупотреблений, его пополнение продолжается), и авторы получили возможность проанализировать статистику совершения ошибок на обширном материале. Анализ включал разработку классификации ошибок и статистическое исследование ошибок некоторых типов. Также были разработаны тренажёры для отработки грамматических правил определённого типа; конкретный тип ошибок, нуждающийся в отработке, определялся на основе корпусных данных о частотности ошибок.

Классификация ошибок

Разработанная классификация ошибок состоит из четырёх подклассификаций по типу информации, которую можно извлечь из допущенной ошибки:

1. Тип ошибки — грамматическое правило, которое нарушается этой ошибкой;
2. Предполагаемая причина совершения ошибки;
3. Критичность ошибки с грамматической точки зрения;
4. Критичность ошибки с точки зрения её влияния на понимание текста в целом.

Ошибке могут приписываться один или несколько параметров из всех четырёх классификаций, поскольку в некоторых случаях затруднительно определить причину ошибки или же ошибка не относится к какому-либо типу грамматического правила (например, ошибки типа Typo ‘опечатка’).

Классификация по типу нарушенного грамматического правила включает в себя 151 подтип. Эти подтипы распределены по шести категориям: лексика, грамматика, синтаксис, дискурс, пунктуация, орфография. Каждая категория распределяется на дальнейшие, более детализированные подкатегории, т.е. классификация имеет древесную структуру. Ниже следуют примеры предложений с ошибками, к которым была применена эта классификация (в квадратных скобках [...] содержится ошибочный текст, подчёркиванием отмечены исправления преподавателя):

(1) To summarize, critical thinking [play] plays a very big role in our lives.

'Подводим итоги: критическое мышление играет очень большую роль в нашей жизни'.

Тип – синтаксис, ошибка согласования – лицо и число, критичность с точки зрения грамматики – средняя, критичность с точки зрения понимания текста – незначительная.

(2) This type of trust is essential for a formation of organizational [identify] identity.

'Такой тип доверия необходим для формирования корпоративного духа'.

Причина – опечатка, критичность с точки зрения грамматики – большая.

Вторая из подклассификаций распределяет ошибки по причинам их совершения. Эта классификация базируется на работах [3, 4], где был предложен анализ закономерностей в совершении ошибок в изучаемом языке и были высказаны гипотезы о возможных причинах допущения ошибок. Классификация причин совершения ошибок выделяет следующие:

1. **L1 interference** (влияние, оказываемое родным языком изучающего, на изучаемый язык) – в большинстве случаев это калькирование конструкций из родного языка на английский язык.

(3) If you did so, your work would [tend to be] become a stereotyped pattern.

'Если бы вы поступили так, ваша работа стала бы стереотипным образцом'.

Подкатегорией *L1 interference* является *Absence of Category from L2 in L1* (отсутствие данной грамматической категории в родном языке изучающего):

(4) If you have a skill like that you will definitely get [ ] a prestigious profession and achieve all the goals.

'Если вы обладаете подобным навыком, вы определённо получите престижную профессию и добьётесь всех своих целей'.

2. **L1 & L2 interaction** (взаимодействие правил родного и изучаемого языков, приводящее к порождению неправильной грамматической структуры

(5) Real friendship is often born [since] in childhood and lasts throughout all your life.

'Настоящая дружба рождается в детстве и длится всю жизнь'.

3. **Typo** (опечатка)

(6) This reality becomes a constituent of the contemporary lifeworld; therefore, it affects the [personalit] personalities, world views and values.

'Эта реалия становится компонентом современной жизни; следовательно, она влияет на личностные качества, идеологию и ценности'.

Кроме того, при разметке ошибок в корпусе был добавлен пункт “Другое”, который выбирался в тех случаях, когда с уверенностью определить причину совершения ошибки не представлялось возможным.

В классификациях критичности ошибок с точки зрения грамматики и смысловой целостности текста ошибкам присваивается количественный коэффициент от 0,5 до 1,5. Величина 0,5 присваивается ошибкам, которые не оказывают большого влияния на понимание текста в одной категории и являются незначительными с грамматической точки зрения в другой категории; 1 – ошибкам средней тяжести в грамматическом и смысловом планах; 1,5 – ошибкам, грубо нарушающим грамматические правила в одной категории, и критичным для понимания текста в другой категории.

Определение смысловой критичности ошибки отдаётся на усмотрение разметчику, что делает разметку довольно субъективной. Суждение о грамматической тяжести ошибки выносится в соответствии с консенсусом разных мнений, высказанных в методологической литературе по SLA.

Статистика ошибок на материале корпуса

В настоящее время в корпусе размечено более 10000 ошибок. Членами исследовательской группы был проведён анализ некоторых типов ошибок:

1) в употреблении артиклей (Е. Кузьменко);
2) в употреблении местоимений (А. Новосёлова);
3) в предложных глаголах и существительных (И. Якименко).

Результаты кратко представлены ниже в качестве примеров возможных исследований на материале нашего корпуса.

Ошибки в употреблении артиклей

Ошибки этого типа – одни из наиболее типичных ошибок русскоговорящих студентов в английском языке. Причиной их высокой частотности является отсутствие морфологического маркирования категории определённости/неопределённости в русском языке и, как следствие, отсутствие артиклей.

Носители русского языка совершают всего три типа ошибок в постановке артиклей:

1. Неправильное использование нулевого артикля:

a) использование нулевого артикля вместо определённого артикля;

b) использование нулевого артикля вместо неопределённого артикля.

2. Использование определённого артикля вместо неопределённого и наоборот.

3. Замена нулевого артикля:

a) определённым артиклем;

b) неопределённым артиклем.

Распределение встреченных ошибок по типам показано на рис. 1.

Из данных рис. 1. следует, что замена определённого или неопределённого артиклей нулевым составляет 71% всех встреченных на исследуемом материале ошибок, т. е. более чем 2/3 всех случаев (т.е. происходит генерализация нулевого артикля).

Из рисунка также видно, что ошибки, связанные с заменой определённого артикля на неопределённый и наоборот не являются частотными, что говорит о том, что в принципе носители безартиклевых языков различают категории определённости и неопределённости.

Ошибки в употреблении местоимений

Наиболее частотными являются ошибки, допущенные в указании неправильного рода местоимения и в употреблении притяжательных местоимений. Нередки ошибки в употреблении указательных и возвратных местоимений; примерно такое же количество ошибок допущено в согласовании числа местоимений и в употреблении неопределенных местоимений (см. рис. 2).

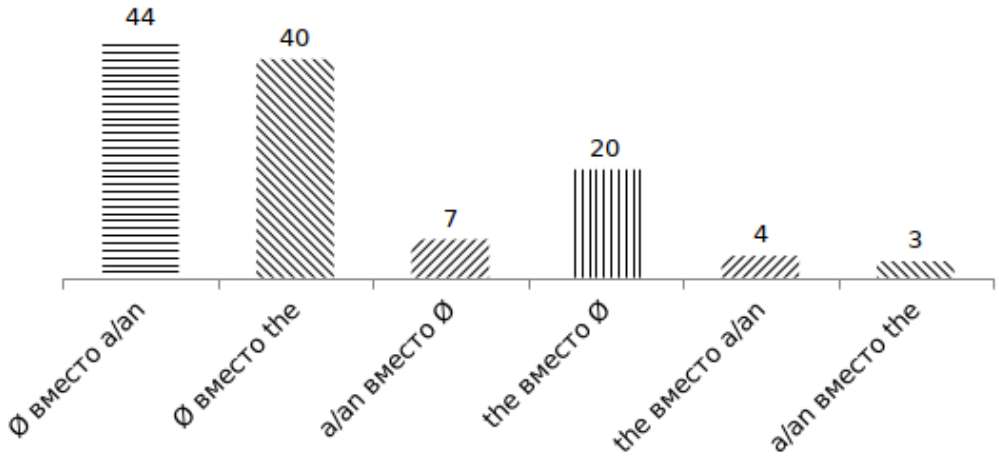

Рис. 1. Распределение типов ошибок в употреблении артиклей

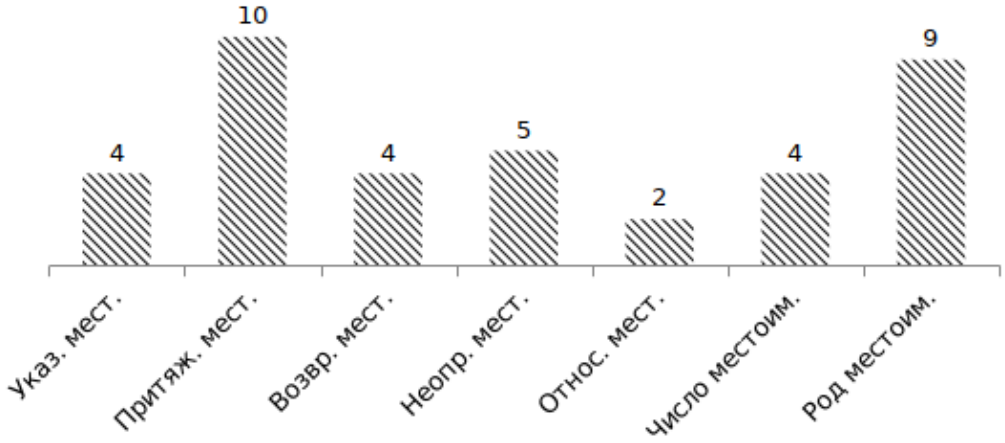

Рис. 2. Распределение типов ошибок в употреблении местоимений

Мы выявили следующие причины, по которым изучающие английский язык делают ошибки:

1. Роль указательных местоимений часто совпадает с ролью определенного артикля, что вызывает использование артикля вместо местоимения;

2. Притяжательные местоимения заменяются личным местоимением в притяжательном падеже;

3. В L1 отсутствуют различные местоимения для маркирования *some* и *any*.

Ошибки в предложных глаголах и существительных

Выбор нужного предлога после глаголов и существительных, наряду с правильным употреблением артиклей и владением сложной системой глагольных времен, относится к основным трудностям для русскоговорящих в изучении английского языка. Статистика ошибок разных подтипов представлена на рис. 3:

Причины, по которым допускаются ошибки в предлогах, следующие:

1. Отсутствие той или иной грамматической категории в родном языке (отсутствие фразовых глаголов в русском языке);

2. Сложность того или иного языкового феномена (в английском языке очень много предложных глаголов и существительных, часто значение таких конструкций не складывается из составных частей, а запоминается).

Тренажёры

В рамках нашего проекта также предполагалось создать языковые тренажёры, учитывающие специфичность совершения ошибок русскоговорящими, изучающими английский язык. Тренажёры основывались на данных об ошибках, представленных в нашем корпусе. Для тестирования было выбрано правило постановки запятой в придаточном предложении:

(7) The river whose bridge is in front of us is called the Cam.

(8) The river Cam, whose bridge is in front of us, looks great at the sunset.

Чтобы оценить эффективность разработанных тренажёров, мы провели следующий эксперимент: 40 студентам третьего курса факультета филологии НИУ ВШЭ было предложено написать небольшое сочинение на английском языке, близкое по формату к разделу Writing в IELTS, и затем в их работах были размечены ошибки. После этого 20 участников эксперимента прошли тренажёры, содержащие предложения из корпуса, в которых были допущены ошибки в постановке запятой в придаточных предложениях. Другие 20 участников (контрольная группа) отрабатывали правило на стандартном материале из учебников. Затем студенты написали второе сочинение аналогичного формата. В собранных работах также были размечены ошибки, после чего мы приступили к проверке статистической значимости между двумя выборками.

Результаты анализа представлены в таблице.

Как показывает таблица, экспериментальная группа показала лучший результат при написании второго эссе, что поддерживает гипотезу об эффективности грамматических тренажёров на материале корпуса.

Анализ эффективности тренажёров

Группа	Среднее количество ошибок на единицу слова до эксперимента	Среднее количество ошибок на единицу слова после эксперимента	Статистическая значимость различия по t-критерию
экспериментальная	0,005122	0,001605	0,046853
контрольная	0,002785	0,004274	0,082329

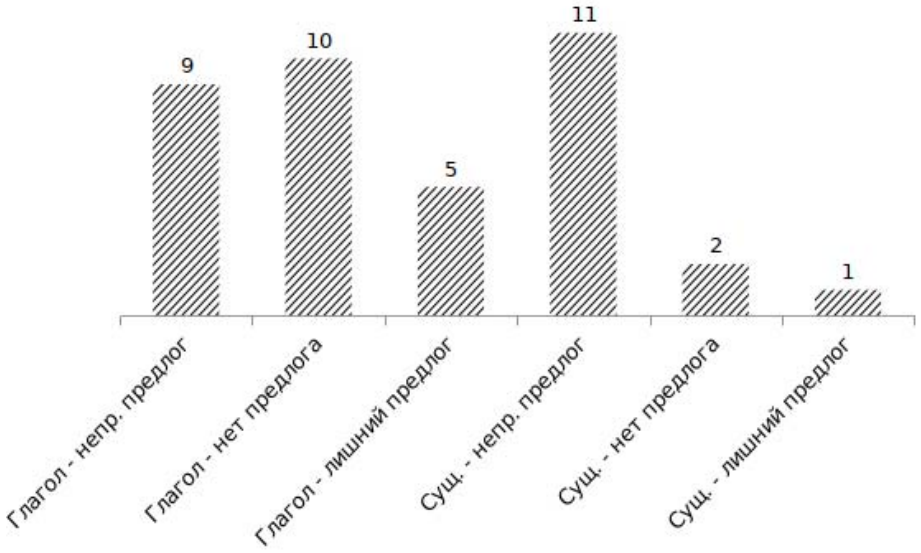

Рис. 3. Распределение типов ошибок в предложных глаголах и существительных

ПАРАЛЛЕЛЬНЫЙ КОРПУС ОШИБОК ПЕРЕВОДА: СОЕДИНЯЯ ПЕРЕВОДОВЕДЕНИЕ И КОРПУСНУЮ ЛИНГВИСТИКУ

Второй из представленных в этой статье (исторически первый) корпус ошибок – совместный проект кафедры перевода Тюменского государственного университета и факультета филологии НИУ ВШЭ. Он создаётся из оригиналов текстов и соответствующих им англо-русских и русско-английских переводов, выполненных студентами переводческих специальностей российских вузов и содержащих переводческие ошибки. Основная цель корпуса – предоставить исследователям и преподавателям перевода обширный и репрезентативный материал для изучения переводческих ошибок и вариативности перевода.

За последние годы идея корпуса воплотилась в полноценный продукт, доступный конечному пользователю. Корпус содержит более 2200 текстов (около 270 оригиналов и около 2000 переводов) и более 1 300 000 словоупотреблений (740 тыс. на русском и 565 тыс. на английском языке). Новые тексты продолжают добавляться, их источниками служат студенческие переводы, собираемые в восьми российских университетах:

1. Тюменский государственный университет, кафедра перевода и переводоведения, специалитет «Перевод и переводоведение», 3–5-й курсы; Центр лингвистического образования, направление подготовки «Лингвистика»; бакалавриат, сокращенная образовательная программа по профилю «Перевод и переводоведение» (вечерняя форма обучения), дополнительная квалификация «Переводчик в сфере профессиональной коммуникации».
2. Московский государственный университет имени М.В. Ломоносова, Факультет иностранных языков и регионоведения, кафедра английского языка для естественных факультетов, дополнительная квалификация «Переводчик в сфере профессиональной коммуникации».
3. Удмуртский государственный университет, Институт иностранных языков и литературы (Ижевск), 5-й курс, очное отделение; 2-й курс магистратуры, специальность «Перевод и переводоведение».
4. Московский авиационный институт, 4-й курс дневного отделения, специальность «Перевод и переводоведение».
5. Пермский государственный национальный исследовательский университет, дополнительная квалификация «Переводчик в сфере профессиональной коммуникации».
6. Нижегородский государственный лингвистический университет, специальность «Перевод и переводоведение».
7. Казанский государственный университет, факультет дополнительного образования, специальность «Переводчик в сфере профессиональной коммуникации».
8. Национальный исследовательский университет «Высшая школа экономики», студия перевода СМИ.

Было произведено выравнивание (alignment) текстов корпуса по предложениям. Каждому предложению оригинала соответствует одно или несколько предложений перевода (имеются соотношения видов one-to-one, one-to-many и many-to-one). Выравнивание производилось автоматически при помощи библиотеки Hunalign [5] и интерфейса LF Aligner. Однако алгоритмы автоматического выравнивания несовершенны, и около 30% пар пришлось корректировать вручную. В настоящее время тексты конвертированы в параллельный корпус в формате TMX (Translation Memory Exchange, подвид XML), практически не содержащий ошибок выравнивания. Таким образом, при поиске можно видеть, какие переводы связаны с теми или иными оригиналами и – наоборот. При этом одному оригиналу в корпусе часто соответствует несколько переводов (до нескольких десятков), что позволяет исследовать вариативность перевода.

По адресу http://rus-ltc.org расположен веб-интерфейс поиска по корпусу. Пользователь вводит интересующее его слово (или последовательность слов) оригинала или перевода и получает список предложений, содержащих запрос, и их эквивалентов на втором языке. Ранее мы планировали использовать для поиска модифицированный TEC Browser из Translational English Corpus [6], но выяснилось, что он не приспособлен для работы с параллельными корпусами. Поэтому был написан собственный интерфейс поиска по двуязычному TMX-файлу. В настоящее время он активно совершенствуется, но в целом уже работоспособен.

Например, по запросу *his mother* система выводит несколько студенческих вариантов перевода одной английской фразы:

The widow's son rushed into the room, found his mother on the floor and saw the computer screen which read…

Ее сын, примчавшийся в комнату и нашедший свою мать лежащей на полу, посмотрел на монитор, где было написано следующее…

Сын вдовы вбежал в комнату, увидел мать на полу и прочитал на экране компьютера…

Сын вдовы нашел ее на полу, потом он увидел почтовое сообщение на экране монитора…

В комнату вбежал сын вдовы, обнаружив свою мать на полу, он посмотрел на экран компьютера и прочитал…

Уже в этом кратком примере можно видеть, как по-разному непрофессиональные переводчики справляются с проблемой асимметрии использования притяжательных местоимений в русском и английском языках: имеются варианты от простой кальки («свою мать») до использования анафоры («её»).

Все тексты в корпусе дополнительно содержат метаданные о переводчике, самом тексте и ситуации перевода. Так, при помощи интерфейса поиска можно легко узнать, что все переводы, приведенные выше, выполнены в 2010 г. студентками 3-го курса ТюмГУ в качестве домашнего задания. Для некоторых переводов указана также оценка. При поиске в переводах результаты можно ограничивать по любому из этих параметров, чтобы исследовать связь между ними и ошибками или вариативностью. Существует и экспериментальный интерфейс поиска с учётом русской морфологии – http://dev.rus-ltc.org/search.

Результаты поиска можно скачать в удобном текстовом формате для последующей обработки. Также важным нам представляется и наличие возможности для любого пользователя скачать весь корпус целиком – как в виде параллельного корпуса в формате TMX, так и в виде архива с исходными текстовыми файлами, снабжёнными метаданными. Единственное требование – это соблюдение условий лицензии Creative Commons Attribution-ShareAlike: пользователь может использовать эти данные в любых целях при наличии ссылки на авторов корпуса (RusLTC Team) и распространении производных работ под аналогичной свободной лицензией.

Разумеется, сам по себе параллельный англо-русский и русско-английский корпус не является чем-то новым в переводоведении и корпусной лингвистике. Новизна и особенность нашего корпуса – именно в разметке его по переводческим ошибкам. В настоящее время практически уже создана наша собственная тестовая классификация ошибок [7] и идёт процесс разметки переводов в системе brat (посмотреть на разметку можно здесь: http://dev.rus-ltc.org/brat/#/rusltc/). По её завершении мы будем иметь некоторый подкорпус, внутри которого можно будет искать по типам ошибок. Нечто подобное представляет из себя MeLLANGE Learner Translator Corpus [8], но в нём нет текстов на русском языке, и он мал по объёму – всего 429 переводов (около 150 тысяч словоупотреблений), причём размечены по ошибкам лишь половина из них.

Мы планируем разметить вручную схожий объём, а затем попытаться «экстраполировать» эту разметку на оставшийся корпус, используя многослойную разметку по частям речи и членам предложения. Таким образом можно, например, в полуавтоматическом режиме находить ошибки, связанные с порядком слов.

ЗАКЛЮЧЕНИЕ

После создания корпусов REALEC и RusLTC и их практического использования можно сделать следующие выводы: корпус текстов с ошибками, допускаемыми в иностранном языке (в нашем случае – английском), действительно предлагает огромные возможности для исследования языковых ошибок.

Использование корпусов возможно в следующих областях:

1. Проведение научных исследований в сфере SLA – выявление причин допускаемых ошибок и анализ частотности совершения ошибок определенного типа.

2. Оценка работы студента в соответствии с коэффициентом, вычисленным на основе классификации критичности ошибок (путем сложения коэффициентов, присвоенных ошибкам, и деления на количество ошибок). При таком подходе работы могут автоматически проверяться по двум параметрам – грамматичности работы и её понятности для читающего.

3. Разработка лингводидактических материалов для изучающих иностранный язык – тренажёров, основанных на корпусных данных о совершаемых ошибках.

4. Формирование стратегии преподавания иностранного языка.

Вместе с тем, параллельный корпус RusLTC уже сейчас представляет собой открытый ресурс, позволяющий производить исследования в области типичных переводческих ошибок и вариативности перевода. Завершение разметки по ошибкам позволит ещё более расширить сферу его применения.

В целом два выше описанных корпуса дополняют друг друга, представляя разные аспекты использования и создания learner corpora для русского языка.

СПИСОК ЛИТЕРАТУРЫ

1. Ishikawa S. A New horizon in learner corpus studies: The aim of the ICNALE Project // Corpora and language technologies in teaching, learning and research / eds. G. Weir, S. Ishikawa, & K. Poonpon. – Glasgow, UK: University of Strathclyde Press. – P. 3–11.
2. Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J. Brat: a Web-based Tool for NLP-Assisted Text Annotation // Proceedings of the Demonstrations Session at EACL 2012.
3. Richards J. A non-contrastive approach to error analysis // English Language Teaching 25. – 1971. – P. 204–219.
4. Corder S.P. The significance of learners' errors // International Review of Applied Linguistics 5. – 1967. – P. 161–169.
5. Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel corpora for medium density languages // Proceedings of the RANLP. – 2005. – P. 590–596.
6. Baker M. The role of corpora in investigating the linguistic behaviour of professional translators // International Journal of Corpus Linguistics. – 1999. – Vol. 4, № 2. – P. 281–298.
7. Куниловская М.А. Классификация переводческих ошибок и их электронная разметка в brat // Проблемы теории, практики и дидактики перевода: Сб. науч. тр. Сер. "Язык. Культура. Коммуникация". Вып. 16. Т. 1. – Н. Новгород: Нижегородский государственный лингвистический университет им. Н.А. Добролюбова, 2013. – С. 59–71.
8. Kübler N. A Comparable Learner Translator Corpus: creation and use // Proceedings of the Comparable Corpora Workshop of the LREC Conference, May 31, 2008. – Marrakech, Maroc, 2008. – P. 73–78.

Материал поступил в редакцию 05.11.14.

Сведения об авторах

КУТУЗОВ Андрей Борисович – кандидат филологических наук, научный сотрудник Национального исследовательского университета – Высшая Школа Экономики, Москва
e-mail: akutuzov@hse.ru

КУЗЬМЕНКО Елизавета Алексеевна – стажер лингвистической лаборатории по корпусным технологиям Национального исследовательского университета – Высшая Школа Экономики
e-mail: eakuzmenko_2@edu.hse.ru

УДК 81’322.2 : 811.161.1

М. Г. Тагабилева

О некоторых нестандартных случаях реализации модели образования композитов со значением *nomina agentis* в русском языке

В современном русском языке существует несколько продуктивных моделей образования сложных слов со значением имени деятеля. Одна из самых продуктивных среди них – модель с нулевым суффиксом: «основа1 +(соединительная гласная) + основа2 + нулевой суффикс», где основа2 – обязательно глагольная. Существуют, однако, целые классы лексем с той же семантикой, образованные, на первый взгляд, по данной словообразовательной модели, но демонстрирующие ряд особенностей, которые не свойственны обычным подобным производным. Работа посвящена подробному описанию и анализу возможных причин возникновения подобных случаев.

Ключевые слова: *морфология, словообразование, композиты, имя деятеля*

В современном русском языке существует несколько продуктивных моделей образования сложных слов со значением имени деятеля, в том числе подробно описанная в наших предыдущих работах модель с нулевым суффиксом: «основа1 + (соединительная гласная) + основа2 + нулевой суффикс» (ср. *чародей, женолюб, зверолов, конокрад*), где основа2 – обязательно глагольная (ср. *-вод* от *водить*, *-терп-* от *терпеть*, *-люб-* от *любить* и т.д.). Основа1 может быть субстантивной, местоименной, наречной, нумеративной или адъективной (субстантивированным прилагательным) [1, 2]. Как практически любая словообразовательная модель (за исключением, пожалуй, модели словосложения для образования слов типа *диван-кровать*), описанная модель накладывает целый ряд ограничений на участвующие в ней основы. Так, по нашим наблюдениям, композиты с нулевым суффиксом образуются только от бесприставочных односложных основ глаголов несовершенного вида. Другим ограничением является ограничение на кластер согласных на конце второй (глагольной) основы: композиты с нулевым суффиксом редко образуются от глагольных корней, заканчивающихся на сочетание согласных. Как верно замечено в [3], наиболее благоприятным является сочетание сонорного с глухим, менее приемлемым, но все же возможным является сочетание фрикативного со смычным. Помимо этого, можно утверждать, что в целом рассматриваемая модель явно тяготеет к трехсложности: стандартная структура производных включает в себя односложную первую часть, односложную вторую часть и соединительный гласный, всего три слога (ср. *же-но-люб, кра-е-вед, до-мо-сед* и многие другие) (подробно об этих и других подобных ограничениях см. [1, 2]).

Существуют, однако, целые классы лексем с той же семантикой, на первый взгляд образованные по стандартным моделям, но демонстрирующие ряд особенностей, которые не свойственны обычным подобным производным. В настоящей работе мы хотели бы остановиться именно на таких случаях.

Первый класс таких «нестандартных» производных образован лексемами, в которых, вопреки норме, глагольная основа занимает первое место, а субстантивная, соответственно, второе. По нашим данным, в современном русском языке существует всего восемь композитов с обратным порядком основ: *вертопрах, ломонос, лежебок, лизоблюд, чинопёр, щелкопёр,* а также *скалозуб* и *любомудр*. По нашему мнению, эти композиты стоит отличать от композитов типа *сорвиголова*, в которых глагольная основа (а точнее, что важно, не просто основа, а глагол в форме повелительного наклонения) регулярно занимает первую позицию. Такие композиты, по-видимому, образованы по иной словообразовательной модели и заслуживают отдельного рассмотрения.

Хотя попытки исследования композитов с обратным порядком основ ранее предпринимались (см., например, [3–5]), объяснения причин перестановок в них ни в одной из известных нам работ не приведено.

Обратимся к материалу.

Лексемы *скалозуб* и *любомудр*, по-видимому, являются искусственными образованиями и должны быть исключены из рассмотрения. Слово *любомудр* является калькой греческого слова *философ* и нарушает сразу несколько ограничений, накладываемых на композиты, образованные по модели с нулевым суффиксом: лексема образуется с **продуктивным** глагольным корнем в начале (ср. *правдолюб*, но **мудролюб*) и является нарушением упомянутого выше сильного морфонологического ограничения на

кластер согласных на конце подобных композитов. Слово *скалозуб*, в свою очередь, явно является авторским производным: по данным Национального корпуса русского языка, лексема не встречается в письменных текстах ранее даты публикации пьесы Грибоедова «Горе от ума», в которой, как хорошо известно, Скалозуб – фамилия одного из центральных персонажей и является явной аллюзией на уже существующее слово *зубоскал* (впервые употребленное, по данным НКРЯ, в 1769 г.). Впоследствии, как это часто бывает с именами известных персонажей классических произведений, фамилия становится именем нарицательным.

Слова *щелкопёр* и *вертопрах,* вероятно, образовались, как верно отмечено в [3], под влиянием ограничения на сочетание согласных на конце подобных композитов: *щелкопёр – лк, вертопрах – рт.* Причины появления слова *лежебок* видятся также в первую очередь морфонологическими (как показано в [2], композиты с нулевым суффиксом плохо образуются от основ на звонкий заднеязычный). Однако стоит признать, что в данном случае все же очень вероятно влияние вышеупомянутой словообразовательной модели с первым глагольным компонентом-императивом типа *сорвиголова* – ср. существование более употребительного варианта *лежебока*: «Аномальный сразу в нескольких аспектах композит *лежебока*, вероятно, следует выводить из **лежибок* с первым компонентом – императивом (ср. *сорвиголова*)» [6, с. 164].

Отдельную проблему составляют слова *лизоблюд, ломонос* и *чинопёр*, где не обнаруживается явных морфонологических причин мены порядка корней.

Слово *лизоблюд* появляется в русском языке только в XVIII в., о чем свидетельствуют данные словарей («Словарь русского языка XVIII века» и «Словарь русского языка XI–XVII вв.»), в то время как его аналог со стандартным порядком корней *блюдолиз* зафиксирован уже в древнерусском языке (см. И.И. Срезневский «Материалы для словаря древнерусского языка по письменным памятникам»). Изначальное буквальное значение слова *блюдолиз (=тот, кто вылизывает блюдо, собирая остатки (чужой) еды)*, имеющее отрицательную коннотацию, постепенно вымывается. Так, этот процесс можно проследить по следующим примерам из НКРЯ:

(1) После обеда садится он опять на прежнее место, где засыпает или забавляется рассказами нескольких блюдолизов, привлеченных в его дом приятным запахом его кухни. [И.А. Крылов. Почта Духов, или Ученая, нравственная и критическая переписка арабского философа Маликульмулька с водяными, воздушными и подземными духами (1789)]

(2) Вот она, как ни бестолкова, как ни привыкла себе барничать да нашего брата похуже своего блюдолиза считать, однако смирилась, шелковая стала, хоть вокруг пальца обмотай. [В.И. Даль. Хлебное дельце (1857)]

(3) ...кроме глотания ножей, ничего изобрести не можешь, потому что и в этом искусстве ты не пошел дальше Апфельбаума, и в этом искусстве ты еще не научился давать представления без помощи стола, накрытого сукном, под которым сидит душка Разбитной, сей нелицемерный холоп и блюдолиз всех Чебылкиных, Зубатовых и Удар-Ерыгиных, и подает тебе, по востребованию, жареных голубей. [М.Е. Салтыков-Щедрин. Сатиры в прозе (1859–1862)]

(4) Злые языки рассказывали, что Рузаев был со всесильным графом когда-то в одном корпусе, даже дружил с ним, но потом, посчитав его выскочкой и блюдолизом, презрительно порвал их дружескую связь. [Михаил Шишкин. Всех ожидает одна ночь (1993-2003)]

Можно заметить, что в примере (1) сосуществуют одновременно два компонента значения – историческое прямое, буквальное и явная отрицательная оценка, в примерах (2) и (3) исторически первое значение также еще прослеживается за счет контекста «еды», но в примере (4) этот смысл не прослеживается уже совсем и слово *блюдолиз* обозначает просто «подлиза, прихлебатель».

На наш взгляд, далее имеет место весьма интересный процесс: при вымывании значения и утрате ассоциативных связей с мотивирующими основами (*блюд-* от *блюдо*) происходит процесс переосмысления словообразовательных связей и основа *-блюд* начинает ассоциироваться уже с обсценным корнем *блюд-/бляд-* (ср. *ублюдок*) и восприниматься как основной носитель значения – глагольный корень. В результате этого новое «семантическое ядро» композита занимает обычную для него позицию в конце композита, а основа *лиз-* становится, соответственно, первой.

Переосмысление словообразовательных связей, по нашему мнению, является причиной перестановки и в лексеме *чинопёр*.

Слово *чинопёр*, изначально произошедшее от «чинить перья» т.е. «заниматься мелкой никчемной работой», по свидетельствам словарей, имеет скорее переносное значение «мелкий чиновник, писарь; чинодрал» (С.А. Кузнецов «Большой толковый словарь русского языка»), т.е. является частичным синонимом слова *щелкопёр*, чье первое значение, по свидетельству словаря Даля, «писец, писарь в суде, приказный, чиновник по письмоводству, пустой похвальбишка и обирала»:

«У Н.Г. Помяловского в отрывках из незаконченного романа ”Брат и сестра“ встречаем образование чинопёр. Речь идет об отставном титулярном советнике, который ”три раза срывал по 300 руб. сер. за то, что били его морду, а морду его, ей богу, и даром можно бить. Эта шельма, уволенная по прошению, обыкновенно подбирал человек шесть забулдыг, в их присутствии раздражал кого-нибудь, незнакомого господина, тот бил его по морде, начиналось дело, и титулярный получал следуемый по закону гонорарий. Наконец, гражданская палата обратила внимание на то обстоятельство, что титулярного что-то очень часто бьют и запретила ему впредь подавать просьбы“» [5, с. 4].

Представляется, что перестановка здесь происходит одновременно под влиянием отчасти синонимичной однокоренной лексемы *щелкопёр* и ассоциации с корнем *чин-* (ср. *чиновник, чинодрал*): «Со словом чинопёр можно сравнить чинодрал. Ведь в слове щелкопёр вторая часть -пер соотносится с

именной основой пер-о (ср. щелкать пером)» [там же]. О наличии тесной ассоциативной связи между словами *чинопёр* и *чинодрал* свидетельствует и следующий пример:

(5) – Это, батюшка, не то что у нас какой-нибудь чинодрал или чинопёр, безжизненнейшая, мертвая душа, строчит какие-то бессмысленнейшие бумаги и не задумается расказнить всякого, кто усомнится в живом значениии списанного бумажного листа. [Г.И. Успенский. Кой про что (1885)]

Таким образом, глагольная основа теряет связь с мотивирующим глаголом и переосмысляется как именная, в результате чего и «становится» в начало.

Слово *ломонос*, имеющее, по данным толковых словарей, в современном русском языке только значение «травянистое или кустарниковое вьющееся растение сем. лютиковых» (МАС), изначально употреблялось как эпитет к слову *мороз* (ср. *мороз-ломонос*), что обозначало «очень сильный мороз». Хотя подобное толкование отсутствует в толковых словарях, о наличии этого значения свидетельствуют некоторые народные пословицы, названия народных праздников, а также следующая цитата:

(6) Твой образ будет, знаю наперед, в жару и при ***морозе-ломоносе*** *не уменьшаться, но наоборот в неповторимой перспективе Росси*. (И. Бродский «Похороны Бобо»)

Существуют различные трактовки происхождения названия растения, однако ни одна из них не зафиксирована ни в толковых, ни в этимологических словарях. Большинство встреченных нами объяснений опубликовано не в лингвистических источниках, а в журналах по цветоводству, что не делает эти версии достаточно надежными, например: «Одни считают, что можно сломать себе нос, запутавшись в зарослях, которые образует клематис. Другие склонны связывать название растения с неприятным запахом выкопанных корней. Но все-таки это и другие названия клематиса в русском языке закрепились благодаря сходству частей растения: ломонос – за загнутый («сломанный») нос у семян…» (http://vesti.lv/society/theme/city/19339-v-lomonosovskij-moroz-lomonos-ukril-svoj-nos.html).

Однако поскольку изначально слово *ломонос*, по-видимому, появилось как эпитет к слову мороз, логично предположить, что именно этот стандартный контекст и повлиял на нестандартный порядок основ в сложении: легко заметить, что большинство народных примет и пословиц строится на созвучии (рифме), что верно в нашем случае.

Таким образом, можно утверждать, что лексемы с нестандартным порядком основ в словообразовательной модели с нулевым суффиксом появляются по нескольким причинам. Во-первых, мена порядка корней может происходить под влиянием морфонологических ограничений, накладываемых в общем случае на исследуемую модель. Во-вторых, немаловажную роль играют процесс переосмысления словообразовательных связей и аналогия (будь то аналогия с созвучным синонимом, как в случае *чинопёра*, или простое созвучие в частотном контексте, как в случае слова *ломонос*). В-третьих, можно предположить, что влияние оказала сама возможность образования и существования в русском языке композитов с первым глагольным корнем, реализующаяся в словообразовательной модели типа «*сорвиголова*», частично синонимичной исследуемой нами модели.

Словообразовательная модель с нулевым суффиксом является достаточно продуктивной в современном русском языке. Как можно заключить из полученных нами в результате исследования Национального корпуса русского языка и текстов сети Интернет данных, эта модель особенно продуктивна при образовании слов с ярко выраженной отрицательной коннотацией, в том числе обсценной лексики, ср., например, *пустоболт, мудозвон* и другие.

Так как основным носителем значения обычно является глагольный компонент, отрицательная оценка также зачастую является частью глагольного значения. Однако существует целый класс случаев, когда отрицательная оценка принадлежит целиком первому корню. А поскольку коннотация на самом деле и составляет основную часть значения, смысловой акцент смещается с глагольного корня на первый (субстантивный) и именно последний становится ядром композита.

В ходе исследования нами было обнаружено три субстантивных корня, обладающих подобными свойствами (не исключено, однако, что существуют и другие корни с похожими свойствами): *дубо-, дуро-,* а также обсценный *мудо-*. Похожим образом ведет себя и корень *пусто-*, который в данном случае можно трактовать как адвербиальный (ср. *пустохвал* от «хвалиться попусту») или как субстантивный (ср. *пустоболт* от «болтать пустое»). Число производных с такими первыми основами выше, чем (в среднем) с другими, а значение производного мало зависит от значения второго корня.

Так как, по нашей гипотезе, семантическим ядром в подобных сложениях является не глагольный, а субстантивный корень, будем условно называть все производные с одинаковой первой основой словообразовательным гнездом с этой мотивирующей основой.

В словообразовательное гнездо с мотивирующей основой *дуро-* входят следующие обнаруженные нами производные: *дурошлёп, дуролом, дуропляс, дурохлоп, дуросвет, дуроплюй*. Несмотря на различную семантику второй основы, значения всех перечисленных производных очень близки, и, как представляется, во многих контекстах они могут быть взаимозаменяемыми:

(7) Пытаясь самостоятельно оказаться в игре, я убеждал себя: Марине слишком тяжело рассказывать мне о последних часах царя Димитрия, о гибели верных ему поляков, о своем бегстве из Кремля, вот она и пишет мне только об этой, нынешней, жизни: «… когда же ты, дурошлёп, вышлешь новую тельняшку? [Андрей Дмитриев. Закрытая книга (1999)]

(8) Очерк писал один наш дуропляс, который ко всем этим сложным задачам Автандила Автандиловича сумел присобачить собственную линию, делая вид, что он расхваливает недозволенного художника, как будто можно хвалить что-то недозволенное. [Фазиль Искандер. Сандро из Чегема (Книга 2) (1989)]

(9) – Эй, ты! Дуроплюй! Косой притабанил и выхватил махалку из воды. [Б.С. Житков. Черная махалка (1925)]

(10) На такие приказы Григорий Иванович с высокой колокольни плевал, говорить по-немецки отказался (он всегда за линией фронта играл роль дуролома полицая). [Анатолий Азольский. Диверсант // «Новый Мир», 2002]

Как видно из примеров, значения производных действительно очень близки и не зависят от семантики глагольного корня: по сути, все эти слова являются эвфемизмами обсценных аналогов.

Подобная картина наблюдается и у производных гнезда *дубо-*, в котором корень имеет менее очевидную отрицательную оценку (а именно обладает такой коннотацией лишь в одном, непрямом, значении). Данное гнездо составляют следующие произ-водные: *дуболом, дуботолк/дуботол, дуботряс, дубонос, дуборос:*

(11) Помню, приезжал некий театральный критик из Москвы, Соня принимала его по первому разряду, хотя за глаза, смеясь, называла дуболомом, надеялась, что он похвалит её в газете, и действительно он упомянул её в рецензии одной фразой: "Спорно, но интересно толкование этого образа актрисой Вишневской". [Анатолий Рыбаков. Тяжелый песок (1975–1977)]

(12) – Генерал повернулся к пустым дверям и крикнул в пространство: – Дуботряс березовый! [Павел Крусанов. Укус ангела // «Октябрь», 1999]

(13) – Тебе имя Докука – больше всех галдел… – Тебе – Дубонос… – Тебе – Дурло… За что? [Михаил Успенский. Там, где нас нет (1995)]

(14) Ты чего к месту прирос, Титка? Вставай, дуборос, кланяйся за угощение… поехали! Послышался беспорядочный треск сдвигаемой мебели, шарканье ног и беспомощные женские вздохи. [Л.М. Леонов. Русский лес (1950–1953)]

Корень *пуст-* обладает, по данным Национального корпуса русского языка, наибольшим из всех перечисленных корней количеством производных: *пустослов, пустопляс, пустобай, пустобрёх, пустозвон, пустохват, пустошлёп, пустоплёт, пустохвал*. Отличительной особенностью данного корня является то, что он явно тяготеет к сочетанию с глаголами речевой деятельности (ср. **словить, брехать, болтать, баять, хвалиться, звонить* (в значении «болтать»), *плести*), а также то, что глагольный корень в данном случае, внося пусть и небольшой вклад в значение, все же может ограничивать сферу употребления того или иного производного, в связи с чем не все такие композиты являются взаимозаменяемыми. Рассмотрим некоторые примеры:

(15) Егор, как говорили в Пажени, весь выдался в Мирона, покойного отца своего: такой же пустоболт, сквернослов и курильщик, только подобрей характером. [И.А. Бунин. Веселый двор (1911)]

(16) А если подобному речетворцу, а вернее, пустобреху сообщат что-нибудь немаловажное, всех касающееся, он немедленно отмахнется: «Мне все это-до лампочки!» [Л.А. Кассиль. Дело вкуса (1964)]

(17) И преспокойно жила дальше, пока Эд не высказался в кулуарах о её обожаемом Ч. Она была готова выцарапать пустозвону глаза, но заодно он крепенько засел в её мыслях, совсем как в прежние времена. [Дарья Симонова. Превосходство (2002)]

(18) На приступках мужчина сидел – пустобай, заворотничок, красновеснушчатый и красноглазый; зевай-раззевайский пускал он на драный сапог. [Андрей Белый. Москва. Часть 1. Московский чудак (1926)]

(19) Министры – английский Кейт и прусский Гольц – неотрывными взглядами следили за пустоплётом-царьком, изредка перебрасываясь между собою сухими короткими фразами и двусмысленно подмигивая один другому. [В.Я. Шишков. Емельян Пугачев. Книга первая. Ч. 1–2 (1934–1939)]

(20) – Ой, старые вы пустохваты, пропаду на вас нету, – отстав от Андрея, вдруг вцепилась в разговор Клавка, будто ожгли ее. [Валентин Распутин. Прощание с Матёрой (1976)]

Из приведенных выше примеров видно, что явно взаимозаменяемыми являются производные со вторым компонентом – корнем глагола речевой деятельности, сложнее ситуация с другими производными:

(21) И прежде всего литература поможет, которая что угодно исказит, как это сделало, например, с французской революцией то вреднейшее на земле племя, что называется поэтами, в котором на одного истинного святого всегда приходится десять тысяч пустосвятов, выродков и шарлатанов. [И.А. Бунин. Окаянные дни (1925)]

(22) И Харькову и Степанову в сущности нечего было делать и оба они, по натуре пустоплясы и бездельники, изощрялись в своей взаимной вражде и не давали мне покоя своими взаимными доносами и кляузами. [Г.А. Соломон (Исецкий). Среди красных вождей (1930)]

Значение композитов *пустосвят* и *пустопляс* в данных примерах неочевидны, но нет никаких указаний на то, что они как-либо связаны с речевой деятельностью, поэтому замена их в данном контексте на одно из производных из группы «пустоболт» изменила бы смысл высказывания. Однако, возможно, под влиянием большого количества композитов со значением «болтун» в исследуемом словообразовательном гнезде, некоторые производные с более широким значением могут принимать на себя и данное конкретное значение:

(23) Раздражал его, собственно. Скворцов – болтун, пустопляс. Смеётся, зуб стальной. [И. Грекова. На испытаниях (1967)]

Нельзя не отметить очевидную фонетическую близость этой группы корней, всегда содержащих корневое «у» (ср. *дУро-, дУбо-, пУсто-, мУдо-*). Возможно также, что аномальный корень с неясным значением *долбо-*, встречающийся у нескольких производных с отрицательной оценкой (ср. обсценное *долбозвон, долбодуй*), изначально является переосмыслением корня *дубо-*. «Народная этимология» данного корня (ассоциация с глаголом *долбить*) в данном случае явно вторична и ошибочна, так как иначе пришлось бы признать существование в русском языке уникальных композитов со значением имени деятеля с двумя глагольными основами. О связи же производных с корнями *долбо-* и *дубо-* свидетельствует не только их явная фонетическая бли-

зость и общий (основной) компонент значения – отрицательная оценка.

Таким образом, можно сказать, что среди исследуемых композитов со значением имени деятеля с нулевым суффиксом выделяется особая группа лексем, смысловым ядром которых является не глагольный (второй) корень, а первый (субстантивный или иной) корень. Значение же глагольного корня в таких производных вымывается (хотя и не в одинаковой степени, ср. описанный выше «вклад» глагольной основы в значение в гнездах *дуро-* и *дубо-,* с одной стороны, и *пусто-*, с другой). Такие производные имеют в первую очередь значение отрицательной оценки, которая привносится в них первым (субстантивным) корнем.

СПИСОК ЛИТЕРАТУРЫ

1. Tagabileva M. Composites denoting nomina agentis in the Russian language: distinguishing competing models. Wiener Slawistischer Almanach. Sonderband 85 (2013). – P. 196–208..
2. Тагабилева М.Г. О некоторых моделях образования сложных слов со значением nomina agentis в русском языке // ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН. – 2014. – Т. X, Ч. 1. – С. 854–865.
3. Сичинава Д.В. Ограничения на вторую глагольную основу в русских сложных словах с нулевым суффиксом: Курсовая работа (рукопись). – М., 1999.
4. Виноградов В.В. Из истории русских слов и выражений // Вопросы стилистики. – М., 1966.
5. Виноградов В.В. Об экспрессивных изменениях значений и форм слов // Советское славяноведение. – 1968. – № 4.
6. Иткин И.Б. Русская морфонология. – М.: Гнозис, 2007.

Материал поступил в редакцию 22.09.14.

Сведения об авторе

ТАГАБИЛЕВА Мария Геннатуловна – преподаватель Национального исследовательского университета «Высшая школа экономики», Москва; аспирант МГУ им. М.В.Ломоносова
e-mail: geratagabileva@gmail.com

Центр (Отдел) научно-информационного обслуживания (ЦНИО) ВИНИТИ РАН

предлагает услуги по предоставлению информационно-аналитических обзоров

ВИНИТИ РАН осуществляет подготовку информационно-аналитических обзоров по инновационным и приоритетным направлениям научных исследований в области точных, естественных и технических наук. Обзоры готовятся ведущими специалистами ВИНИТИ, работающими в определенных областях науки и техники. Аналитические материалы содержат результаты анализа и обобщения информации по актуальным научным проблемам, а в некоторых случаях – и прогностические выводы. Основой для составления обзоров служит отечественная и зарубежная научно-техническая литература, доступная ВИНИТИ РАН: фонд НТЛ, включающий более 2 млн отечественных и иностранных журналов, книг, депонированных рукописей, авторефератов диссертаций и другой научной литературы, ретроспектива – с 1987 года. Имеется доступ к базам данных и Интернет-ресурсам: БД ВИНИТИ (разработка ВИНИТИ), БД SCOPUS, БД зарубежных патентов и другим. Кроме того, ВИНИТИ доступны зарубежные электронные платформы ряда ведущих научных издательств, выпускающих основную часть академических рецензируемых журналов, в полнотекстовом варианте.

Основные тематические направления предлагаемых обзоров:

- Науки о жизни;
- Физико-математические науки;
- Химия и науки о материалах;
- Индустрия наносистем и материалов;
- Науки о Земле;
- Рациональное природопользование;
- Информационно-телекоммуникационные системы;
- Энергетика, энергоэффективность, энергосбережение;
- Транспортные, авиационные и космические системы;
- Производственные технологии.

Предлагается подготовка и заказ информационно-аналитических обзоров и материалов по тематике заказчика. Такие обзоры могут относиться к упомянутым выше тематическим направлениям, но могут иметь и междисциплинарный характер. В этом случае обзоры отражают актуальную научную информацию и научные достижения, происходящие на стыке наук.

Более подробная информация о приобретении, заказе и цене обзоров представлена на сайте ВИНИТИ www.viniti.ru

Приобретение и заказ обзоров от юридических лиц проводится на договорной основе. Форма договора для последующего оформления представлена на сайте ВИНИТИ.

Оформление договоров и других необходимых документов производится Центром научно-информационного обслуживания ВИНИТИ (ЦНИО). Возможен прием заказов от физических лиц, оплата производится на расчетный счет или в кассу ВИНИТИ РАН.

Выполненные в ВИНИТИ обзоры предоставляются заказчикам в печатном виде либо в электронном варианте после оплаты заказа.

Обращаться в ЦНИО ВИНИТИ:

- адрес: 125190, Россия, г. Москва, ул.Усиевича, 20.

- телефоны: 8(499) 155 -42 -43, 8(499) 155 -42 -17

- **эл. почта cnio@viniti.ru, fdk@viniti.ru.**

- **факс 8(499) 930 -60 -00 (для ЦНИО).**