УДК 001.103:002

МЕТОДОЛОГИЯ ПРОВЕДЕНИЯ ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ЛИНГВИСТИЧЕСКОГО ЭКСПЕРИМЕНТА

И. Е. Воронина

Воронежский государственный университет

Поступила в редакцию 12.04.2011 г.

Аннотация. Рассматривается методология проведения лингвистических исследований, основанная на методе лингвистического эксперимента. Предлагается единообразный подход к проведению лингвистических исследований.
Ключевые слова: формализация естественного языка, компьютерная лингвистика, лингвистический эксперимент, сочетаемость языковых единиц, автоматизация исследований.
Annotation. The methodology of carrying out of the linguistic researches, based on a method of linguistic experiment is considered. The uniform approach to carrying out of linguistic researches is offered.
Key words: Natural language formalization, computer linguistics, linguistic experiment, compatibility of language units, automation of researches.

ВВЕДЕНИЕ

Человеческая деятельность и общение обусловливает необходимость выработки большого количества знаковых систем для передачи информации и организации деятельности. Для понимания содержания сообщения необходим способ трансляции, чтобы источник и получатель одинаково понимают связь между значением и знаком. Если имеются данные о функционировании моделируемой системы или ее аналога за достаточно продолжительное время, то значения некоторых параметров можно определить путем статистической обработки их значений, зафиксированных в процессе наблюдения. Общеизвестные методы математической статистики позволяют не только определить текущее значение параметра, но и оценить достоверность такого определения в зависимости от продолжительности (числа) наблюдений и их совпадения (разброса), определить необходимое число наблюдений для определения значения параметра с заданной точностью.. Иногда значения интересующих параметров удается определить по аналогии со значениями других, схожих с определяемым, значения которого известны. Однако нередки случаи, когда значения параметров моделируемых систем не удается получить названными выше методами. Такая ситуация бывает особенно характерной для систем с высоким уровнем неопределенности и не имеющих достаточной предыстории функционирования.

Проблемы формализации слабоструктурированных областей столь неоднозначны и нетривиальны, что попытка их решения вынуждает во многих случаях полагаться на интуицию исследователя. В условиях отсутствия возможности применения стандартных математических и алгоритмических методов решения задачи, любой подход, реализованный в виде набора инструментальных средств для проведения исследований, позволяющий получить приближение к правильному результату, имеет значение. При этом необходимо оценивать полученные результаты для дальнейшего анализа материала. Приложив определенные усилия, связанные с активным использованием средств автоматизации, требующиеся оценки можно получить эмпирическим путем. Оценки можно искусственно создавать, изучая последствия их применения.

Д. А. Поспелов, определяя основные направления работ по искусственному интеллекту, наряду с проблемой представления знаний, проблемой планирования и организации поведения, особо выделил проблему общения. Естественный язык, по его мнению, – это не только средство общения пользователя с информационной системой, но и моделирующая система,

средствами которой можно описать окружающий мир во всем его многообразии [1, 2]. Известны комплексы проблем под общим названием «проблема отладки информационного обеспечения» [3], связанные с трудоемкостью ввода громадного объема знаний и данных при создании промышленных информационных систем. При использовании естественного языка ввод и модификация информационного обеспечения становятся доступными не только специалистам в области автоматизации. То же самое касается и диалогового описания процедур планирования, выполнения основных функций взаимодействия пользователя с ЭВМ, понимания и корректировки поведения системы.

Сегодняшние актуальные проблемы, связанные с лингвистической средой, – это, в первую очередь, информационный поиск в среде Интернет, то есть создание систем интеллектуального поиска и обработки информации.

МОДЕЛИРОВАНИЕ ЛИНГВИСТИЧЕСКИХ ПРОЦЕССОВ

Традиционно, любая формализация подразумевает наличие совокупности правил, позволяющих строить описание объекта на декларативном или функциональном уровне. По сути дела, эти правила позволяют ответить на вопрос «как можно» (построить, описать, сделать и т. д.). Предлагается подход к формализации, основанный на системе правил «как нельзя». Правила вида «как нельзя» разбиваются на группы. Каждая группа правил определяет фильтр. Каждый фильтр – это подсистема запретов на наличие семантического отношения, связывающего структурные единицы. Вся проблема в том, что ни один из фильтров нельзя определить однозначно и сразу. Именно по этой причине весь разрабатываемый инструментарий должен быть ориентирован на использование опыта и интуиции исследователя, подкрепляемых использованием математических оценок для принятия решения.

Рассчитывая на то, что между объектами предметной области всегда предполагается наличие семантического отношения, можно определить общий вид правила:

СЕМАНТИЧЕСКОЕ ОТНОШЕНИЕ (<объект1 > И < объект2 >) ИМЕЕТ МЕСТО, или

СЕМАНТИЧЕСКОЕ ОТНОШЕНИЕ (<объект1 > И < объект2 >) ИМЕЕТ МЕСТО С ОПРЕДЕЛЕННОЙ ДОЛЕЙ УВЕРЕННОСТИ (SemR ($Comp_1$ –> $Comp_2$), или, когда подразумевается степень уверенности x_i, SemR ($Comp_1 \xrightarrow{x_i} Comp_2$)).

Частный случай правила, имеющего важное значение в лингвистических исследованиях:

СОЧЕТАЕМОСТЬ(<объект1 > И < объект2 >) ИМЕЕТ МЕСТО С ОПРЕДЕЛЕННОЙ ДОЛЕЙ УВЕРЕННОСТИ ($Comp_1$ –> $Comp_2$, или, когда подразумевается степень уверенности x_i, $Comp_1 \xrightarrow{x_i} Comp_2$).

Таким образом, семантическое отношение может трактоваться как метрическая лингвистическая шкала с нечеткими квантификаторами, определяющими субъективную оценку наличия этого отношения. Исследователь сам может установить пороговое значение, сравнение с которым поможет отсеять часть претендентов на сочетаемость, оставив материал для размышления и изучения. Принятие решения будет заключаться в формулировке правила сочетаемости (фильтра).

Но и отвергнутый материал может быть подвергнут исследованию. поскольку ранжирование правил по степени значимости позволит получить картину, которая может косвенно быть полезна при принятии решения

Нельзя не отметить, что без средств автоматизации исследования часто могут быть выполнены лишь на весьма ограниченном материале (а некоторые действия и вовсе невозможны) в силу значительной трудоемкости процесса. Но сами по себе средства автоматизации не могут быть эффективны, если еще на этапе их проектирования, а затем и на этапе создания прототипа системы не происходит глубокое погружение в предметную область и отсутствует тесное взаимодействие с исследователем. Роль личности исследователя (эксперта) трудно переоценить: во многих случаях именно он является генератором идей, вдохновителем и критиком процесса и результатов разработки. Без серьезной совместной работы невозможно создание адекватного инструмента, позволяющего решать задачи в соответствующей предметной области.

Модель проведения исследований, основанная на эволюционном подходе, предполагает выявление новых знаний в виде правил путем наблюдения и эксперимента, программное подтверждение этих правил и пополнение системы. Цели, которые ставит перед собой человек, далеко не всегда могут быть достигнуты только за

счет его собственных возможностей или внешних средств, имеющихся у него в данный момент. Такое стечение обстоятельств называют проблемной ситуацией. Примером такой ситуации, требующей разработки методологии и создания программного инструментария, является случай, когда обычные способы сбора и переработки информации не обеспечивают необходимой полноты и быстроты ее обработки, что значительно снижает качество принимаемых решений. Это хорошо иллюстрируют исследования сочетаемости структурных единиц в лингвистике. Что касается прикладных научных исследований в области формализации естественного языка, наличие такой проблемной ситуации является предпосылкой для разработки новых подходов и на их базе инструментальных средств для их проведения.

При проведении лингвистических исследований представляется разумным использовать тот факт, что язык – открытая система закрытых подсистем. Каждая подсистема конечна, следовательно, ее можно моделировать, а затем устанавливать определенные отношения между подсистемами. Основа моделирования лингвистических процессов – порождение. Это связано, в первую очередь, с трудностью формализации ЕЯ: для выявления формализованных правил приходится осуществлять анализ через синтез языкового материала, а затем подтверждать правила путем порождения информации уже на основании выявленных ранее правил.

В качестве основного подхода к вопросам проведения исследований предлагается метод лингвистического эксперимента, предложенный И. А. Бодуэном де Куртенэ в XIX в. [3], и развитый его учеником Л. В. Щербой в XX в. [4]. Согласно этому методу, правила сочетаемости языковых единиц диагностируются посредством обращения к отрицательному материалу, который анализируется и служит базой для выявления, формализации и подтверждения правил порождения слов. Достоинства отрицательного материала трудно переоценить, поскольку он позволяет ставить, а затем и решать вопросы, которые не возникают и не могут возникнуть при, по существу, автоматическом пользовании положительным материалом родного языка. Синтез отрицательного материала осуществляется из реально существующих морфем, занимающих позиции, отвлеченные от реально существующих слов

Технологическая цепочка проведения исследований [6] на каждом этапе должна быть подкреплена методами и средствами для автоматизации исследований. Цель автоматизации – повышение эффективности проведения исследований для каждого звена технологической цепочки в целом через повышение эффективности выполнения отдельных операций за счет устранения нетворческой, рутинной работы и обеспечения быстроты, полноты и качества переработки информации. Кроме того, эффективность является мерой степени достижения цели и предполагает наличие целей более высокого уровня, для которых «продукция» такой подсистемы является реальным вкладом в общий исследовательский процесс.

Полагая главной целью исследований (в идеале) полную формализацию процесса, составляющего суть звена технологической цепочки, в качестве подцелей можно рассматривать поэтапное приближение к достижению цели путем пополнения системы фильтров и диагностирования процесса фильтрации. Здесь огромную роль играет устранение рутинной, нетворческой работы по накоплению информации (например, генерации лингвистического материала по заданным начальным установкам).

В нашем случае можно рассматривать информацию как ресурс исследовательского процесса, поэтому эффективность представляет собой комплексное понятие. Оно включает в себя такие слагаемые, как:

– качество информации и ее соответствие потребностям процесса исследований;

– трудоемкость обработки информации, отражающую уровень производительности труда исследователя;

– доступность системы для пользователя.

Идеология автоматизированного решения задач определяет последовательность этапов решения, зависящую от возможностей формализации задачи. Как правило, пользователь формулирует задачу в терминах той предметной области, в которой он является специалистом. При этом он может иметь слабое представление о возможностях и путях реализации поставленной им задачи на современных средствах вычислительной техники. Для эффективного применения средств автоматизации должна быть построена формальная схема решения задачи. Соответствие между формальной пос-

тановкой задачи и исходной, заданной пользователем, достигается в ходе взаимодействия с исследователем

Следует затронуть вопрос о динамике модели подсистемы ЕЯ. На стадии проведения исследований модель не является отображением «один к одному». Она скорее является основой инструмента изучения. Если полагать, что адекватная модель – это та модель, с помощью которой достигается поставленная цель, то введенное понятие адекватности не полностью совпадает с требованиями полноты, точности и правильности. В этом случае адекватность означает, что эти требования выполнены не вообще, а лишь в той мере, которая достаточна для достижения цели. Цель – это ответ на вопрос, по каким правилам организована подсистема определенного языкового уровня. Для достижения цели необходимо выявить эти правила. Наличие правил позволит исключить из рассмотрения так называемый отрицательный материал. Отсечение отрицательного материала происходит с помощью фильтрации. Процесс фильтрации осуществляется с помощью системы фильтров, которые имеют вид правил, реализующих запреты на определенные сочетания структурных составляющих. Набор фильтров не закрыт. Он предполагает добавление новых фильтров в их систему, которая имеет иерархическую структуру. Программа предоставляет средства, необходимые для получения и накопления исследовательского материала и его анализа, для того чтобы сформулировать правила нового, более высокого уровня для их проверки и подтверждения. Поэтому модель является расширяемой. Она организована таким образом, что подразумевает свое расширение в процессе эксплуатации (этот процесс можно назвать эволюцией модели) добавлением новых компонентов в виде фильтра очередного уровня. Конечным этапом эволюции является полнота набора фильтров, что и приведет к главной цели модели – будут сформулированы и программно реализованы правила построения определенной подсистемы.

В нашем случае определение исходной системы, сбор и обработка данных очевидным образом должны быть автоматизированы. Определение исходной системы и сбор данных в этом случае представлены моделью системы. Обработка данных определяет поведение модели. Лицо, принимающее решение (ЛПР), занимается интерпретацией данных, именно его решение влияет на пополнение модели (добавление новых признаков, новых правил, определяющих запреты на порождение определенных цепочек). Но ЛПР (исследователю в нашем случае) необходим инструмент для обеспечения обоснования и поддержки принимаемых решений, при этом объектом управления является модель процессов, представляющих один из языковых уровней. Наш объект относится к так называемым нетрадиционным объектам управления [2] и обладает перечнем свойств, характерным для таких объектов.

Диагностика процесса функционирования модели выступит в качестве инструмента, обеспечивающего принятие решения. Концептуальная схема проведения исследований ЕЯ приведена на рис. 1.

Под объектом в данном случае понимается объект моделирования, то есть языковые про-

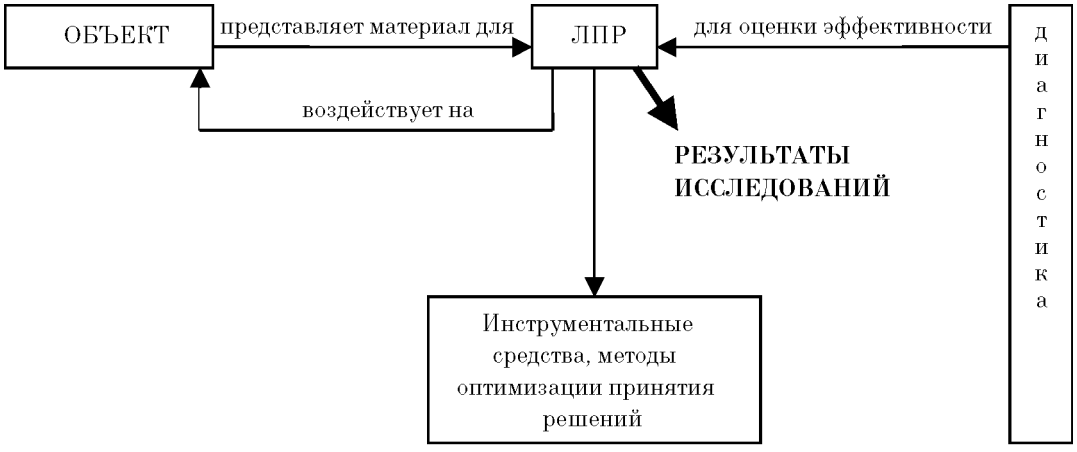

Рис. 1. Концептуальная схема проведения исследований

цессы определенного уровня иерархии. Моделирование объекта происходит на основании стартовой информации, делающей возможной саму попытку начальной формализации и, следовательно, автоматизации. Поскольку речь идет о порождающей модели, результатом ее функционирования будет сгенерированный материал, подлежащий наблюдению и изучению.

Использование диагностики уместно на этапе создания фильтра очередного уровня. При включении его в существующую иерархию (или просто набор) фильтров можно продиагностировать наличие отклонения модели от реальной системы (подсистемы) и получить необходимые количественные оценки.

Диагностический инструментарий не обязательно должен порождать количественные оценки с абсолютной степенью точности, поскольку имеет своей целью скорее обозначение тенденций, подтверждение или опровержение правильности хода исследовательского процесса.

Подразумевается использование представленной схемы на каждом шаге технологической цепочки. Таким образом, предлагается единообразный подход к последовательному полному или частичному решению проблем формализации. Под частичным понимается решение, достаточное для реализации определенной задачи или класса задач и не претендующее на законченность или всеобщность.

Для реализации исследований необходимо создать математическое обеспечение процесса диагностики и принятия решения. Предполагается разделить математическое обеспечение на две категории:

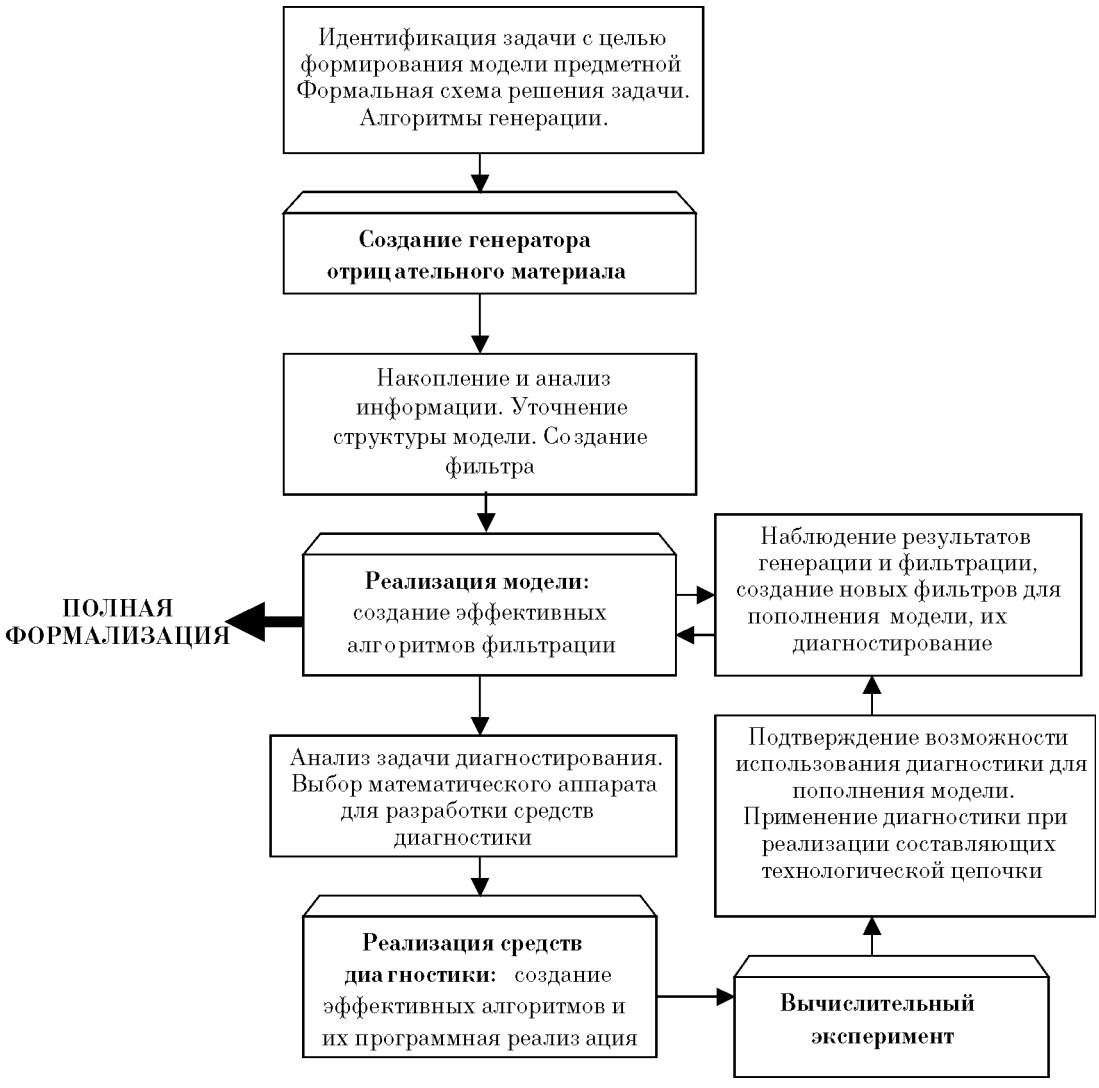

Рис. 2. Этапы проведения автоматизации исследований лингвистических объектов

1) имеющее чисто диагностическую направленность, функциональная нагрузка которого заключается в определении степени приближения аналога (модели) лингвистического объекта к реальности;

2) математическое обеспечение для поддержки принятия решения.

ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

В [7] представлены математические оценки второй категории в наиболее часто встречающейся ситуации формализации сочетаемости языковых единиц определенного уровня. Рассматривается сочетаемость слов. Прежде чем будут окончательно сформулированы правила сочетаемости, которые затем подвергнутся диагностике, необходимо их выявить. Этот процесс весьма нетривиален, далеко не прост и не обязательно успешен. Поэтому используются дополнительные механизмы, которые позволяли бы стимулировать принятие решения, говоря проще «подталкивать» процесс. Инструментальные средства, реализующие математическое обеспечение, оценивание достаточно просты и рассчитаны на то, чтобы пользователями были не обязательно только высококвалифицированные специалисты предметной области. На начальном этапе исследований оценивание исходного материала может быть весьма рутинной задачей, не требующей серьезной квалификации, что предполагает использование вспомогательного персонала.

В целом, жизнеспособность подхода проиллюстрирована на примере решения словообразовательных задач [8], выявления семантической близости ключевых слов [9].

СПИСОК ЛИТЕРАТУРЫ

1. *Поспелов Д. А.* Логико-лингвистические модели в системах управления / Д. А. Поспелов. – М. : Энергоиздат, 1981. – 232 с

2. *Поспелов Д. А.* Ситуационное управление. Теория и практика / Д. А. Поспелов. – М. : Наука, 1986. – 288 с

3. *Попов Э. В.* Общение с ЭВМ на естественном языке / Э. В. Попов. – М. : Наука, 1982. – 260

4. *Бодуэн де Куртенэ И. А.* Избранные труды по общему языкознанию / И. А. Бодуэн де Куртенэ. – М. : Изд-во АН СССР, 1963.– Т. 2. – 392 с.

5. *Щерба Л. В.* Избранные работы по русскому языку / Л. Б. Щерба. – М. : Учпедгиз, 1957. – 188 с.

6. *Воронина И. Е.* Компьютерное моделирование в лингвистике / И. Е. Воронина // Материалы семинаров научно-образовательного центра «Волновые процессы в неоднородных и нелинейных средах» ; отв. ред. А. С. Сидоркин. – Воронеж : Воронежский государственный университет, 2004. – С. 7–25.

7. *Воронина И. Е.* Оценки сочетаемости структурных единиц в задачах формализации естественного языка / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2006. – № 1. – С. 51–57.

8. *Воронина И. Е.* Задачи словообразования как составная часть проведения исследований в области естественно-языкового общения / И. Е. Воронина // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2006. – № 2. – С. 135–141.

9. *Воронина И.Е.* Алгоритмы определения семантической близости ключевых слов по их окружению в тексте / И.В. Попова, И.Е. Воронина, А.А. Кретов // Вестн. Воронеж. гос. ун-та. Серия Системный анализ и информационные технологии. – 2010. – № 1. – С. 148–153.

Воронина Ирина Евгеньевна – кандидат технических наук, доцент кафедры программного обеспечения и администрирования информационных систем факультета ПММ

Voronina Irina Ye. – Associated Professor of Software & Information System Administering Chair, Department of Applied Mathematics, Computer Science & Mechanics, Voronezh State University