OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 519.766

ДВУХУРОВНЕВЫЙ ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР ОТВЕТНЫХ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Д.Ш. Сулейманов (*dvdt.slt@gmail.com*)
Казанский федеральный университет, г.Казань, Республика Татарстан, Россия

В статье описывается лингвистический процессор (ЛП) ответов обучаемого на естественном языке в диалоговом контексте, основанный на двухуровневой модели семантической интерпретации вопросно-ответных текстов. ЛП на входе получает ответ обучаемого на заданный вопрос и на выходе формирует диагностический вектор ситуаций, характеризующий степень правильности ответа, включая его смысловую полноту и корректность.

Ключевые слова: двухуровневая модель ответа, индивидуальная концептуальная грамматика, концептула, семантическая классификация вопросно-ответных текстов.

Введение

В процессе диалогового общения всегда существует контекст, который определяет дополнительную информацию, способствующую правильному пониманию смысла сообщения. В условиях вопросно-ответного диалога такой контекст настолько определенен, что задающий вопрос достаточно четко может априори очертить круг ожидаемых возможных ответов и декодировать ожидаемый смысл из многообразия грамматически правильно построенных фраз в соответствии с этим предварительным знанием. Смысловая типизация вопросов и соответствующая семантическая классификация ответных текстов дают возможность противопоставить каждому типу вопроса ограниченный набор допустимых, т.е. логически правильных, смысловых конструкций (ответных формул). Можно рассматривать совокупность этих формул, соответствующих конкретному типу вопроса, как некоторую грамматику, кодирующую конструкции, передающие правильный смысл ответа в контексте, заданном вопросом. Нами была поставлена и решена задача такой классификации вопросно-ответных текстов, когда форма и соответствующий смысл входного текста напрямую зависят от типа вопроса.

Построение семантического интерпретатора текстов на естественном языке (ЕЯ) в контексте, управляемом вопросом системы к пользователю, имеет свою специфику, выгодно отличающую его от других ЕЯ-диалоговых систем и создающую реальные предпосылки для построения эффективного смыслового интерпретатора [Бухараев Р.Г. и др., 1990].

Введем определения понятий *концептулы и индивидуальных концептуальных грамматик*. *Концептула* - это элементарная смыслообразующая единица семантической структуры текста, отражающая роль лексем в значении вопроса и в определенном их сочетании формирующая смысл ответа в контексте, детерминированном заданным вопросом.

Схемы сочетания концептул, соответствующие правильной передаче ожидаемого смысла ответов определенного класса, будем называть *индивидуальными концептуальными грамматиками (ИКГ)*. Таким образом, каждая ИКГ представляет собой некий семантический синтаксис, отображающий ролевую структуру ответного текста. Использование понятия концептуальной грамматики дает возможность сводить семантический анализ содержания

ответа к анализу соответствия его ролевой структуры некоторой ИКГ, ожидаемой по заданному вопросу.

Семантическая типизация вопросов позволяет разбить множество ответов обучаемого на семантические классы, в каждом из которых требуется раскрытие некоторого смысла, определенного типом вопроса и независимого от формы задания и лексического наполнения вопроса.

В статье раскрываются базовые принципы построения и архитектура лингвистического процессора ответных текстов на естественном языке в диалоговом (вопросно-ответном) контексте. На конкретном примере демонстрируется работа ЛП, который на входе получает ответ обучаемого на заданный вопрос и на выходе формирует диагностический вектор ситуаций, характеризующий степень правильности ответа.

1. Архитектура и принципы построения ЛП вопросно-ответных текстов на ЕЯ

Лингвистический процессор ответов предназначен для анализа ответа обучаемого на естественном языке без дополнительных ограничений на форму и объем ответного текста. ЛП имеет декларативно-процедуральное представление, где декларативную часть составляет модель ответа. В процедуральную часть входят лексический процессор и семантический интерпретатор. Производится двухуровневый анализ ответов: на первом (поверхностном) уровне - лексический, когда осуществляется анализ используемых слов и их канонизация, и на втором (каноническом) - семантический, когда устанавливается соответствие канонического представления ответа ожидаемой семантической схеме. Анализ производится на основе двухуровневой модели ответа. В результате анализа вырабатывается диагностический вектор ситуаций, представляющий собой последовательность кодов, характеризующих типы ошибок в ответе.

Архитектура лингвистического процессора ответных ЕЯ-текстов в контексте, управляемом вопросом, показана на рис. 1.

Обработка ответного текста происходит следующим образом. Ответ обучаемого на конкретный заданный вопрос поступает в лексический процессор (на рисунке - **ЛексП**), который осуществляет полную лексическую обработку текста на основе модели ответа. Модель ответа (**МО**) представляет собой двухуровневую базу знаний, включающую таблицу ролей лексем в оцениваемом ответе на первом (поверхностном) уровне, и комплекс индивидуальных концептуальных грамматик (**ИКГ**), соответствующих ожидаемому классу ответов, на втором (глубинном) уровне. Модель ответа строится и заполняется либо специалистом по предметной области ("инженером по знаниям"), либо самой системой, по задаваемому вопросу на основе информации в базе знаний, когда база знаний включает онтологическую модель предметной области.

Последовательно анализируя каждое входное слово на основе таблицы ролей МО на первом уровне, ЛексП переводит лексемы в соответствующие им роли (концептулы) и в итоге получает каноническое описание смысла ответа (на рисунке - **КО**) в виде последовательности концептул. Те лексемы в ответе, которые, возможно, не будут идентифицированы на основе МО, также могут представлять ценность с точки зрения корректности оценки ответа (например, для дальнейшей проверки их на непротиворечивость с ожидаемым смыслом ответа), поэтому накапливаются в специальных файлах (**СФ**). Вся информация, получаемая в процессе анализа ответа на уровне ЛексП регистрируется в векторе ситуации (**ВС**). Далее, на втором уровне, канонический ответ (КО) поступает в семантический интерпретатор (**СемИ**) и анализируется с привлечением специальных семантических схем – ИКГ, представленных на втором уровне МО. ИКГ реализованы декларативно. Это позволяет изменять (например, дополнять или исправлять, сортировать сочетания концептул по частоте использования их в ответах) и расширять концептуальную грамматику новыми ИКГ без изменения процедуральной части ЛП.

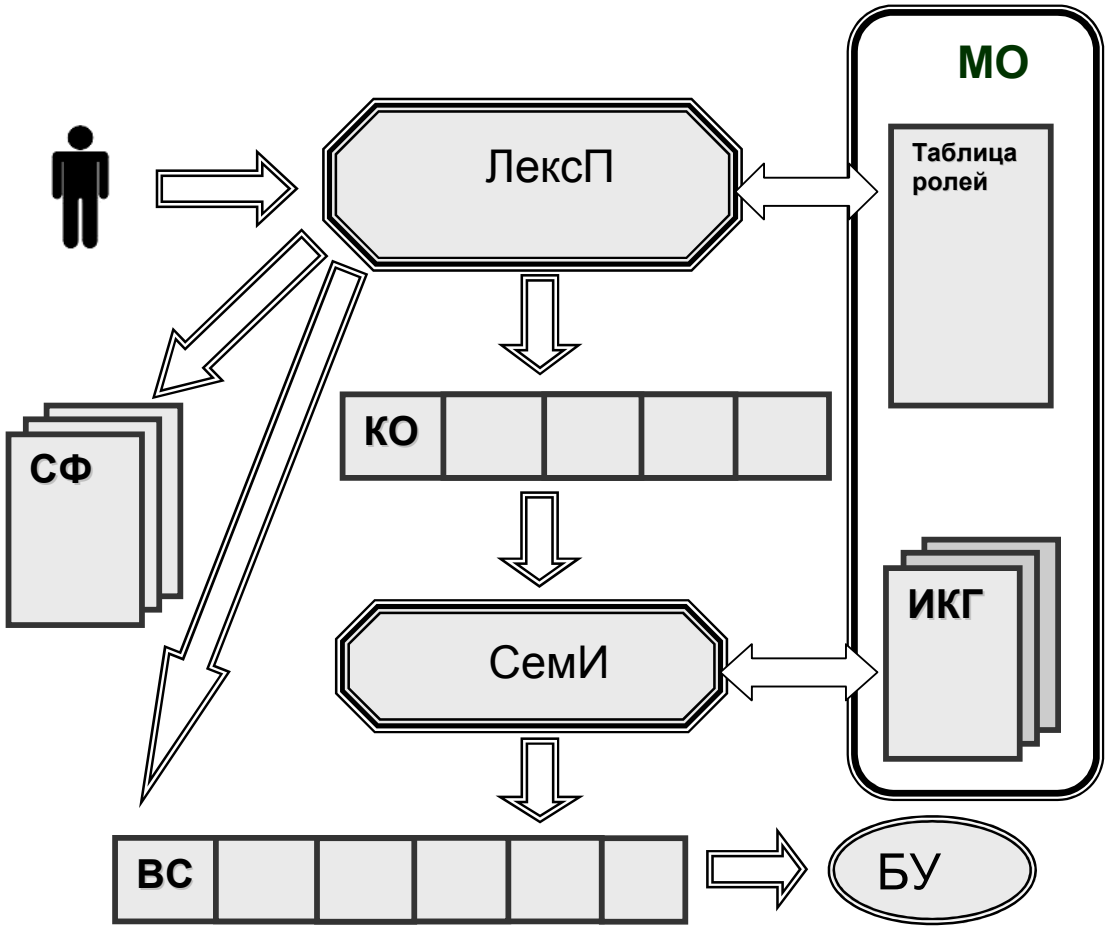

Рисунок 1 - Двухуровневый лингвистический процессор ответных ЕЯ-текстов

Результат формируется в виде дополнения вектора ситуации, частично заполненном на первом уровне. Полный вектор ситуации, как результат анализа ответа двухуровневым лингвистическим процессором, является той информационной базой, на основе которой принимается решение блоком управления (на рисунке - БУ) по дальнейшему управлению процессом обучения.

Далее в статье раскроем более детально ряд утверждений и содержание блоков, приведенных выше и представленных на рисунке 1.

Построение лингвистического процессора базируется на следующих двух методологических принципах и шести принципах реализации прагматически-ориентированной модели.

Методологические принципы:

1. Принцип детерминированности контекста. В силу активности, система «погружает» пользователя в определенный контекст, который определяется заданным вопросом. Соответственно, содержание ответа, его лексикон и даже форма и, отчасти, объем предопределены и пользователь с необходимостью отвечает на вопрос в определенных рамках.
2. Принцип ожидаемости смысла ответа. По заданному вопросу система знает пространство значений вопроса, т.е. ей заранее известен контекст, в котором будет происходить интерпретация ответа и достаточно легко может быть сформирована модель текста, адекватная ожидаемому ответу как по лексике, так по форме изложения и семантической конструкции.

Принципы реализации:

Принцип 1. Выделение системы смыслообразующих единиц – концептул, с целью трансформации проблемы семантического анализа вопросно-ответного текста в проблему синтаксического анализа в условиях использования детерминирующей роли контекста.

Данный принцип согласуется с Постулатом об исходной точке описания [Кибрик и др., 1987]: "исходными объектами лингвистического описания следует считать значения (и предопределяющие их исходные для речемыслительного процесса сущности) и им ставить в соответствие выражающие их языковые формы".

Принцип выделения концептул приводит к необходимости провести типизацию понятий, отношений, грамматических признаков и специальных ролей лексем и установления соответствия между ними и концептулами в управляемом контексте, т.е. в контексте заданного вопроса.

Выделение концептул производится на основе анализа типов лексем и их ролей в вопросно-ответных текстах.

Принцип 2. Семантическая классификация вопросно-ответных текстов на основе типовых отношений: выделение конкретных типов отношений, типов вопросов и классов ответов для реализации детерминирующей роли контекста.

В условиях определенного контекста существует возможность упростить способы кодирования смысловой информации, а, следовательно, и способы ее декодирования. При анализе текста в процессе общения оказывается важным фиксирование контекста и установление зависимости формального выражения смысла (т.е. грамматической конструкции) от этого контекста.

В вопросно-ответном диалоге система функционирует в условиях такого определенного контекста и она способна четко очертить круг ожидаемых возможных ответов, т.е. значений вопроса, и декодировать ожидаемый смысл из многообразия грамматически правильно построенных фраз в соответствии с этим предварительным знанием. Смысловая типизация вопросов и семантическая классификация значений вопроса дают возможность противопоставить каждому типу вопроса ограниченный набор допустимых ответных формул, т.е. логически правильных смысловых конструкций. Можно рассматривать совокупность этих формул, соответствующих конкретному типу вопроса, как некоторую грамматику, кодирующую конструкции, передающие правильный смысл ответа.

Следовательно, при семантическом подходе к типизации вопросов и классификации ответов имеется прямая связь между типом вопроса и классом ответа. Принадлежность ответа к некоторому классу ответов определяется не по его объему и содержанию, и не по форме вопроса, а по типу вопроса системы и по ожидаемому смыслу.

Принцип 3. Разработка индивидуальных концептуальных грамматик (ИКГ) семантических классов, отражающих смысловые конструкции ответов соответствующих классов и в совокупности составляющих концептуальную грамматику (КГ) как схему реализации принципа трансформации семантики в синтаксис, служащей формальной основой для построения семантического интерпретатора, ориентированного на "слушающего".

Правомерность применения этого принципа подтверждается также и Постулатом о мотивированности [Кибрик и др., 1987]: "исторически исходное соотношение между смыслом и грамматической формой мотивировано: устройство грамматической формы отражает тем или иным образом суть смысла".

Сочетания понятий и отношений в текстах, соответствующих определенным семантическим классам, имеют достаточно устойчивые частотные характеристики. Следовательно, при создании системы семантической интерпретации логично ожидать в анализируемом тексте семантические конструкции, имеющие наиболее высокие частотные характеристики для рассматриваемого контекста.

Схемы сочетания концептул, соответствующие правильной передаче ожидаемого смысла, названы нами индивидуальными концептуальными грамматиками (ИКГ).

Принцип 4. Сегментация вопросно-ответных текстов по минимальным смысловым конструкциям для рекурсивного применения правил концептуальной грамматики (базовых смысловых формул).

Принцип 4 обосновывается тем, что любой осмысленный текст допускает актуальное членение на синтагматические группы, линейные или иерархические, а также очевидным утверждением, что любой осмысленный текст полностью «покрывается» линейной или иерархической последовательностью сегментов, отражающих его глубинное каноническое описание.

В проблематике семантического анализа текстов на ЕЯ, особенно для практической реализации разработок, оказывается важной задача членения входного текста на такие части, к которым рекурсивно применимы простые формулы. Сложный текст представляет собой линейную и/или иерархическую последовательность смысловых частей, относящихся к тому или иному семантическому классу ответов. *Сегмент есть часть сложного текста, или полный текст, относящийся к определенному семантическому классу.* Таким образом, сложный текст является линейно и/или иерархически организованным множеством сегментов.

В известных системах понимания ЕЯ практически отсутствуют эффективные механизмы выделения сегментов в анализируемом связном тексте для применения к ним ограниченного набора унифицированных правил анализа. Глубинные причины такого положения лежат в сложности самой проблемы членения входного текста на соответствующие смысловые части. Это посильно только действительно интеллектуальной системе, способной на основе плавающего (уточняющего смысл части текста по месту чтения) контекста выделять смысловые конструкции, рекурсивно идентифицируемые с правилами ИКГ соответствующих классов ответов.

Для реализации принципа сегментации важно ответить на следующие два вопроса: как определить контекст, в рамках которого входной текст должен анализироваться на смысловую корректность, и каким образом выделять в тексте сегменты, чтобы к ним были рекурсивно применимы грамматические формулы.

Ответ на первый вопрос ведет нас, в общем случае, к необходимости подробного анализа проблем плавающего контекста, условий его изменения, сохраняющих смысловую непрерывность восприятия и ряда других методологических, лингвистических, психологических и семиотических проблем. В нашем случае мы сознательно идем на некоторое упрощение ситуации, фиксируя контекст по заданному вопросу. В силу этого входной текст, т.е. ответ обучаемого, однозначно попадает в рассматриваемый контекст и фактически содержит ожидаемый смысл (точнее, должен содержать, иначе текст не является ответом на вопрос или не распознается данным семантическим интерпретатором).

Для применения соответствующих ИКГ требуется определить, к какому семантическому классу ответов относится вводимый текст. В случае вопросно-ответного текста система способна заранее по задаваемому вопросу предопределить семантический класс ожидаемого ответа, тем самым предопределяя и соответствующую ИКГ, применяемую для его смыслового анализа. Семантическая классификация вопросно-ответных текстов производится от простого к сложному. Вначале определяются простые семантические классы ответов, т.е. ответы, в которых раскрывается смысл вида "понятие-отношение-понятие". Затем из таких конструкций строятся более сложные семантические классы, представляющие собой комбинации простых классов, как линейные, так и иерархические, отражающие существование связных текстов из простых, сложносочиненных и сложноподчиненных предложений. Соответственно, сегментация текстов приводит к построению как линейных, так и иерархических представлений, которые рекурсивно распознаются на основе определенных ИКГ.

Принцип 5. Релевантность представления знаний (модели ответа) по смысловой структуре и лексическому наполнению ожидаемому ответному тексту.

Ясно, что наиболее эффективный диалог, т.е. достаточно адекватная и реактивная интерпретация входного текста, будет осуществляться при соблюдении принципа релевантности представления знаний (модели ответа) по смысловой структуре и лексическому наполнению ожидаемому ответному тексту. Это является естественным требованием к системе интерпретации, моделирующей человеко-машинный интерфейс, т.к. в управляемом контексте активный участник диалога всегда имеет возможность заранее построить модель ответа адекватно ожидаемому контексту по своему вопросу.

Принцип 6. Принцип открытости системы, обеспечивающий развитие системы путем накопления новых знаний на основе устойчивых статистических характеристик, в том числе, путем расширения множества обобщенных семантических единиц, введения новых типов вопросов и классов ответов, сортировки и расширения правил концептуальной грамматики как совокупности всех ИКГ, введения новых ИКГ.

2. Семантическая классификация вопросно-ответных текстов

Любая предметная область (ПО) содержательно представляет собой совокупность значимых *понятий* и *отношений* между этими понятиями, изложенной в определенной последовательности. Множество конкретных понятий и отношений по определенным признакам можно разбить на конечное число *типов понятий* и *типов отношений*. Назовем эти типы *семантическими единицами (концептулами)*. Каждое осмысленное предложение ПО можно перевести в текст, составленный из типов понятий и типов отношений, т.е. семантических единиц, без детального учета грамматических признаков лексем, соотнося каждое понятие или отношение с определенным типом.

Полный отказ от элементов классической грамматики ЕЯ оправдан не во всех случаях. В передаче смыла предложения в определенных ситуациях важную роль играют такие грамматические признаки как падежные окончания слов, предлоги и др., и их учет позволяет существенно упростить семантическую интерпретацию ответного текста. Поэтому нами введена дополнительная семантическая единица (концептула) - *грамматическая роль лексем или их частей* для указания соответствующих *грамматических признаков* естественного языка, значимых для более эффективного анализа корректности ответа.

Смысл анализируемого ответного текста зависит также от специфики проблемной области. Этим вызвано введение третьего типа концептул - *специальных ролей лексем* в ответе пользователя.

Таким образом, в исследуемой модели канонический смысл текста определяется сочетанием концептул четырех указанных типов, соответственно, четырьмя группами концептул.

Первая группа концептул - множество концептул, отражающих различные *типы понятий*. Обозначим, K_S = {SS, SS(i), SO, $S_{ОП}$, SA, SP}. Здесь ***SS*** - концептула, отражающая *главное понятие* (первая буква *S* - признак того, что концептула отражает понятие), т.е. понятие/понятия, относительно которого/которых задан вопрос. Сложные тексты могут содержать несколько понятий, связи которых раскрываются в анализируемых предложениях, каждый из которых в процессе анализа определенной части предложения может, в свою очередь, выступать в роли главного понятия. Для их различения в пределах анализируемого текста вводится обозначение: ***SS(i)*** - концептула, отражающая *i-е главное понятие*; ***SO*** - концептула, отражающая *понятие, состоящее в некотором определенном отношении с главным понятием;* $S_{ОП}$ - концептула, отражающая *обобщенное понятие* (ОП). ОП - это *понятие, находящееся по отношению к главному на более высоком уровне в иерархии понятий предметной области* (т.е. интенсионал, например, понятие "человек" есть ОП по отношению к понятию "студент"); ***SA*** - концептула, отражающая *понятие-аргумент;* ***SP*** - концептула, отражающая *понятие-результат*.

Вторая группа концептул - множество концептул, отражающих различные *типы отношений.* Обозначим, K_R ={ R_C , $R_{СОСТ}$,$R_{ВКЛ}$,$R_Д$,$R_{ВРО}$, $R_{ПРО}$, $R_{КЛО}$, $R_{КЧО}$, R_{SO} , R_{OS} , R_A , R_P }. Здесь R_C - это концептула, соответствующая типовому отношению *Состояние,* $R_{СОСТ}$ *– Состав,* $R_{ВКЛ}$ *– Включение,* $R_Д$ *– Действие,* $R_{ВРО}$ *– Временное Отношение,* $R_{ПРО}$ *- Пространственное Отношение,* $R_{КЛО}$ *– Количественное Отношение,* $R_{КЧО}$ *– Качественное Отношение,* R_{SO} - концептула, отражающая отношение ***SS*** к ***SO,*** R_{OS} - концептула, отражающая отношение ***SO*** к ***SS,*** R_A - концептула, отражающая отношение ***SS*** к ***SA,*** R_P - концептула, отражающая отношение ***SS*** к ***SP***.

Третья группа концептул - *Грамматические роли лексем и их частей,* отражает грамматические признаки естественного языка (элементы грамматики, например, суффиксы, союзы, предлоги и др.). Обозначим, K_G = { GP_A , GP_P , ***Gm*** , Gf_1 , Gf_2 }. Здесь ***G*** - признак грамматических ролей; GP_A *- предлог перед SA* (например, для русского языка предлоги *из, от, с* и т.п.)**;** GP_P *- предлог перед* ***SP*** (например, предлоги *в, на, к* и т.п.); ***Gm -****грамматические модификаторы*: *лексемы типа 'чем', 'нежели'* и т.п. после лексемы, выражающей отношение, или *падежные окончания слова* после лексемы, выражающей понятие; Gf_1 *- функциональная лексема, обозначающая признак начала причинной части ответа*, в котором раскрывается

причинно-следственное отношение. Например, лексемы *'потому что', 'так как', 'если'* и т.п.; ***Gf_2*** *- функциональная лексема, обозначающая признак начала следственной части ответа*, в котором раскрывается причинно-следственное отношение. Например, лексемы *'то', 'тогда', 'значит'* и т.п.

Четвертая группа концептул - *специальные роли лексем,* отражающие специфику элементов ответа на конкретный вопрос, т.е. в определенном контексте. Обозначим, K_L= {***LN, LZ, LNE, LI_S , LI_O , LI_A , LI_P , LI_R*** }. Здесь, Здесь ***L*** - признак ролей специальных лексем. ***LN*** - *необязательная лексема*, т.е. лексема, отсутствие или наличие которой в ответе не влияет на смысл ответа; ***LZ*** - *запрещенная лексема*, т.е. лексема, наличие которой в ответе недопустимо (рассматривается как ошибка); ***LNE*** - *неопределенная лексема*, т.е. лексема, не предусмотренная разработчиком курса. ***LI*** - *интервальная лексема*, т.е. лексема, которая накладывает некоторое ограничение на понятие или отношение (указывает область действия, например, *'2К памяти', 'все операторы'* и т.д.). Интервальная лексема при ***SS*** отражается концептулой ***LI_S***. Аналогично записываются другие концептулы для интервальных лексем: ***LI_O*** - при ***SO, LI_A*** - при ***SA, LI_P*** - при ***SP, LI_R*** - при отношениях.

Далее, на основе введенной классификации коцептул, проведем семантическую классификацию вопросно-ответных текстов.

На форму задания вопросов не накладывается специальных ограничений. Ограничения естественным образом исходят из того требования, что вопрос должен быть однозначно понят обучаемым (т.е. по тексту вопроса должно быть понятно, раскрытие какого понятия и смысла требуется в ответе). Так, выделяются следующие *типы вопросов и соответствующие им классы ответов.*

I. Вопросы, требующие явного задания в ответе ключевых понятий (отношения явно заданы в вопросе). Сюда относятся вопросы типа: *'Напишите программу вычисления функции на Паскале', 'Назовите состав компилятора'.*

Этому типу вопросов соответствуют классы ответов, в которых обязательно явно содержатся ключевые понятия. Например, *ответы выборочного типа* (даны несколько ответов, необходимо указать правильный ответ); *ответы типа "ДА - НЕТ"*; *ответы фиксированно-конструируемого типа* (когда дается часть ответа и необходимо дописать недостающие лексемы); *численные ответы* и т.п.

II. Вопросы, требующие раскрытия в ответе типового отношения одного главного понятия.

Это вопросы следующего типа: *'Что выполняется раньше: компиляция или загрузка?', 'Что легче - железо или дерево?' и т.п.*

Можно указать следующие классы ответов, раскрывающие одноименные типовые отношения: *Состав, Включение, Действие, Состояние, Временное отношение, Пространственное отношение, Количественное отношение, Качественное отношение и др.*

III. Вопросы, требующие раскрытия в ответе составного отношения одного главного понятия. Составное отношение может состоять из нескольких простых отношений. Например, таким составным отношением является отношение Функция, которая в ответном тексте одновременно отражает отношение главного понятия и к аргументу, и к результату. К этому типу относятся вопросы типа: *'Какую функцию выполняет компилятор?', 'Назовите предназначение загрузчика', 'Что делает мельница' и т.п.*

Такому типу вопросов соответствуют классы ответов, в которых главное понятие раскрывается через составное отношение. Например, ответ: «Мельница перемалывает зерно в муку» относится к *классу ответов Функция*, в котором отражено отношение главного понятия «мельница» к понятию-аргументу «зерно», а также и к понятию-результату «мука».

IV. Вопросы, требующие раскрытия в ответе произвольной комбинации простых типовых и/или составных отношений одного главного понятия. К данному типу относятся

вопросы: *‘Дайте описание химического вещества К’, ‘Что Вы знаете о кибернетике?’, ‘Дайте определение компилятора’*.

Этим вопросам соответствуют *классы ответов, в которых главное понятие раскрывается через его простое типовое отношение и/или составное отношение*. Можно выделить, например, следующие классы ответов:

1) *Описание* - это *класс ответов, в которых раскрываются произвольные комбинации типовое отношения и/или составного отношение главного понятия с другими понятиями*: S_i *состоит из* S_{i+3} , S_{i+4} , S_{i+5} . *Переводит* S_{i+6} *и* S_{i+7} *и выполняется раньше* S_{i+1} , где S_i , S_{i+7} , S_{i+3} , S_{i+4} , S_{i+5} , S_{i+6} – понятия ПО.

2) *Определение* - *это класс ответов, в которых главное понятие раскрывается через ОП* - обобщающее понятие (т.е. понятие на более высоком уровне в иерархии, интенсионал) и класс Описание. Например, к этому классу можно отнести ответ: *‘Студент - это человек, который обучается в вузе’*.

3) *Причина* - это *класс ответов, в которых раскрывается условие существования некоторых отношений главного понятия с другими понятиями.* Предполагается, что главное понятие следствия и его отношения с другими понятиями заданы в вопросе. Например, рассмотрим текст ответа: *‘Дерево не тонет в воде, потому что удельный вес дерева меньше удельного веса воды’*. Если это ответ на вопрос: *‘Почему дерево не тонет в воде?’*, то ответ относится к классу *Причина*. Здесь *главное понятие* следствия *‘дерево’* и его отношение с объектом *‘вода’* дается в самом вопросе. Часть ответа *‘Потому что удельный вес дерева меньше удельного веса воды’* раскрывает условие существования указанного следствия.

4) *Следствие* - это *класс ответов, в которых раскрывается следствие от существования некоторых отношений главного понятия с другими понятиями.* Тот же пример в этом случае демонстрирует ответ на вопрос: *‘Что следует из того, что удельный вес дерева меньше удельного веса воды?’*. Здесь *главное понятие* причины *‘удельный вес дерева’* и его отношение *‘меньше’* к другому понятию *‘удельный вес воды’* даются в вопросе. В части ответа: *‘Дерево не тонет в воде’* раскрывается следствие от существования указанного условия.

В ответах на вопросы типа I-IV *главное понятие* не меняется в процессе просмотра текста (т.е. предполагается, что ответы содержат информацию только относительно одного *главного понятия*).

V. Вопросы, требующие раскрытия в ответе более чем одного главного понятия. Например, к ним относятся вопросы следующего типа: *‘Расскажите о Казанском университете’, ‘Докажите теорему’ и т.п.*

Этому типу вопросов могут соответствовать ответы, в которых *главное понятие* меняется в процессе просмотра ответа, т.е. роль главного понятия переходит на то понятие, отношения которого с другими понятиями раскрываются далее в ответном тексте. Нами выделены следующие классы ответов, в которых содержатся *главные понятия*, связанные только общим контекстом.

1) *Детализация.* В ответах этого класса происходит *детализация понятий, состоящих в некотором отношении с главным понятием.*

Пример вопроса V типа: *‘Какая связь существует между институтом и заводом?’*. Ответом может быть следующий текст, относящийся к классу *Детализация*: *‘В институте разработана САПР, которая используется для проектирования токарных приспособлений, которые внедряются на заводе’*. В этом ответе три *главных понятия* - *‘институт’*, *‘САПР’*, *‘токарные приспособления’*. Последовательно раскрываются следующие отношения этих понятий с другими понятиями: *разработал* - *‘институт разработал САПР’*, *проектирует* - *‘САПР проектирует токарные приспособления, внедряются* - *‘токарные приспособления внедряются на заводе’*.

Разбиение текстов на семантические классы осуществляется *по типу отношения главного понятия*, раскрываемого в данном ответе, и не зависит ни от конкретной ПО, ни от понятий

данной ПО, ни от конкретного языка общения с системой. Это позволяет строить эффективные предметно-независимые анализаторы, ориентированные на раскрытие определенного типа отношения главного понятия в рамках соответствующего класса ответов.

При семантическом подходе к типизации вопросов и классификации ответов имеется прямая связь между типом вопроса и классом ответа. Принадлежность ответа к некоторому классу ответов определяется не по его объему и содержанию, и не по форме вопроса, а по типу вопроса преподавателя и по ожидаемому смыслу.

3. Индивидуальные концептуальные грамматики. Модель ответа. Описание вектора ситуаций. Сегментация ответных текстов.

Семантическим классам ответов соответствуют присущие им схемы сочетания концептул, передающие характерный (обобщенный) смысл ответов данного класса (значений вопросов). Как было определено выше, схемы сочетания концептул, соответствующие правильной передаче ожидаемого смысла, названы индивидуальными концептуальными грамматиками (ИКГ). Смысл введения ИКГ заключается в сведении семантического анализа текста к синтаксическому анализу его канонического представления в условиях, определенных некоторым контекстом.

В качестве примера рассмотрим ИКГ класса ответов Функция и технологию ее построения.

Пусть задан вопрос типа 3: ‘Какую функцию выполняет компилятор?’ Очевидно, значением данного вопроса (т.е. ответами) может быть множество следующих поверхностных форм:

1) *переводит исходный текст на языке высокого уровня в объектный текст в машинных кодах,*
2) *получает ЯМК из ЯВУ*,
3) *компилятор переводит ЯВУ в ЯМК* .

Здесь отношение *‘переводит’* есть R_A , отношение *‘получает’* - R_P, понятия *‘текст на языке высокого уровня’*, *‘ЯВУ’* - *SA*, *‘текст в машинных кодах’, ‘ЯМК’* - *SP*, предлог *‘из’* - GP_A , предлог *‘в’* - GP_P, понятие *‘компилятор’* есть *главное понятие* - *SS*.

Формализованное представление ответов, соответственно, имеет вид:

1) R_A —> *SA* —> GP_P —> *SP*
2) R_P —> *SP* —> GP_A —> *SA*
3) *SS* —> R_A —> *SA* —> GP_P —> *SP*

Исследуя таким образом всевозможные варианты поверхностных, а далее и глубинных представлений ответов, в которых ожидается раскрытие составного отношения Функция одного главного понятия, мы получаем следующее описание ***ИКГ*** классов ответов ***ФУНКЦИЯ***:

<ИКГ ФУНКЦИЯ> :: = [SS* —>]((R_A **** —> (*** GP_P ***—> SP* —> SA* | SA* —> *** GP_P ***—> SP*) | RP* —> (*** GP_A ***—> SA* —> SP* |SP* —> *** GP_A ***—> SA*)| ((*** GP_P ***—> SP* —> *** R_A **** —> SA* | SA* —> RA* —> *** GP_P ***—> SP*) |(*** GP_A ***—> SA* —> *** R_P **** —> SP* | SP* —> *** R_P **** —> *** GP_A ***—> SA*)))***

Знак ‘|’ обозначает альтернативное вхождение сочетаний концептул. Круглые скобки (,) служат для объединения концептул разных типов. Квадратные скобки [,] означают необязательное вхождение.

Модель ответа строится на основе задаваемого вопроса и представляет собой пару ***<F, G>***. ***G*** обозначает ИКГ класса ответов, соответствующего заданному вопросу. ***F=<L,K>*** - представляет собой информационную структуру, содержащую лексемы, отражающие понятия и отношения и их предполагаемые роли в ответе, где ***L*** - множество лексем, ожидаемых в ответе, а ***K*** - множество концептул. Каждому ***i***-му классу ответов соответствует определенный тип ***F(i)*** со своим набором концептул.

Например, МО для класса *Функция* имеет следующее описание:

ФУНКЦИЯ*:* ***SS=<LM>; SA=<LM>; SP=<LM>;*** R_A***=<LM>;*** R_P ***=<LM>;***

GP_A ***=<LM>;*** GP_P***=<LM>;***LI_S***=<LM>;*** LI_{RA} ***=<LM>;*** LI_{RP} ***=<LM>;***

LI_A ***=<LM>;*** LI_P ***=<LM>; LZ=<LM>; LN=<LM>***. Здесь ***<L>::=<лексема> [,<синоним>, ..., <синоним>]***, ***<LM>::=<L> |... | <L>***.. Для вопроса типа 3: ***"Какую функцию выполняет компилятор?"*** - формируется F(3) по оператору:

ОТВЕТ: КЛАСС = ***ФУНКЦИЯ***;

F: SS=*&комп&, &транс&;* $\boldsymbol{R_A}$*= переводит,преобр&т;****SA****=&ЯВУ&;* $\boldsymbol{R_P}$ *=получает;****SP****=пр*гр &+на+ЯМК,&ЯМК&.*

G: *ИКГ Функция*

Для каждого класса ответов формируется отдельный ***вектор ситуаций*** **(ВС)**. Покажем в качестве примера структуру векторов ситуаций для классов ответов на вопросы типа 2 и типа 3.

ВС для классов ответов на вопросы типа 2 (ВС2) имеет следующее представление: ***КЛАСС = <****Название класса ответов* **> S1 S2 S3 S4 S5 S6 S7.**

Здесь, S1 - это код, характеризующий лексическую полноту ответа. Значением S1 является соотношение количества лексем, использованных в ответе, и лексем, предусмотренных моделью ответа.

S2 - код, указывающий на наличие в ответе запрещенных лексем. Значением S2 является число, характеризующее количество *LZ* в ответе обучаемого.

S3 - код, указывающий на использование в ответе неопределенных лексем, т.е. лексем, непредусмотренных моделью ответа. Значением S3 является количество неопределенных лексем.

S4 - код, характеризующий модальность ответа: а) неуверенность, т.е. присутствие в ответе лексем типа *«возможно», «наверное»* и т.п., улучшающих оценку неверного и принижающих оценку верного ответа; б) категоричность, т.е. присутствие в ответе лексем типа «*конечно», «безусловно», «непременно»* и т.п., усиливающих, подтверждающих правильный или еще более принижающих слабый, неверный ответ; в) нейтральность, т.е. отсутствие в ответе лексем типа а) и б). Таким образом, значением S4 является 0, 1 или 2, соответственно, для случаев а), б) и в).

S5 - код, характеризующий правильность использования интервальных лексем, т.е. лексем-ограничителей, накладывающих определенные ограничения на другие лексемы в ответе. Например, количественные характеристики или слова типа *«не», «нет»* и т.п. Значением S5 является 0 или 1 (верно/неверно).

S6 - код, характеризующий правильность глубинного смысла ответа, т.е. соответствие его канонизированного представления определенной схеме ИКГ. Значением S6 является: а) 0, если канонизированное представление соответствует ИКГ; б) 1, если в ответе отсутствует отношение; в) 2, если канонизированное представление не соответствует ИКГ, т.е. нарушен глубинный смысл.

S7 - код, характеризующий смысловую полноту ответа, т.е. степень соответствия канонизированного представления ответа определенному сочетанию концептул в ИКГ по длине: а) полное соответствие; б) канонизированное представление короче; в) канонизированное представление длиннее. Значением S7 является: 0, для случая (а); 1, для случая (б); 2, для случая (в).

ВС для классов ответов на вопросы типа 3 (**ВС3**) имеет следующий вид (на примере *класса Функция*):

КЛАСС = ФУНКЦИЯ **S1 S2 S3 S4 S5 S6 S7.**

ВС3 отличается от ***ВС2*** содержанием кода S6. Код S6 ***ВС3*** характеризуется следующими значениями: а) 0, если канонизированное представление соответствует ИКГ; б) 1, если в ответе отсутствуют отношения; в) 2, если канонизированное представление не соответствует ИКГ; г) 3, если указано только одно отношения; д) 4, если в ответе отсутствует *SA*; е) 5, если в ответе неверно указан *SA*; ж) 6, если в ответе отсутствует *SP*; з) 7, если в ответе неверно указан *SP*.

Коды S1, ..., S5 и S7 такие же, что и в ***ВС2*.**

В проблематике семантического анализа текстов на ЕЯ, особенно для практической реализации разработок, оказывается важной задача расчленения входного текста на такие части, к которым рекурсивно применимы простые формулы. Сложный ответный текст рассматривается нами как линейная и/или иерархическая последовательность смысловых частей, относящихся к тому или иному семантическому классу ответов. Как определено выше, с***егмент*** *есть часть сложного текста, или полный текст, соотносящийся с определенным семантическим классом.* Следовательно, *сложный текст, с точки зрения структурного образования, является линейно и/или иерархически организованной последовательностью сегментов,* которые рекурсивно распознаются на основе соответствующих ИКГ.

Для реализации принципа сегментации важно ответить на следующие два вопроса: как определить контекст, в рамках которого входной текст должен анализироваться на смысловую корректность, и каким образом выделять в тексте сегменты, чтобы к ним были рекурсивно применимы грамматические формулы ИКГ. В нашем случае, с одной стороны, из-за требований высокой реактивности семантического анализатора в автоматизированной обучающей системе (АОС), с другой стороны, в силу выгодных особенностей проблемной области, позволяющих весьма удачно использовать два введенных выше методологических принципа - «ожидаемости ответа» и «детерминированности контекста», мы сознательно идем на некоторое упрощение ситуации, допуская, что входной текст, т.е. ответ обучаемого, однозначно попадает в рассматриваемый контекст и фактически содержит ожидаемый смысл (вернее, должен содержать, иначе текст не является ответом на вопрос или не распознается нашей системой). Для применения соответствующих ИКГ, требуется определить, к какому семантическому классу ответов относится вводимый текст. В случае вопросно-ответного текста автор курса способен заранее по задаваемому вопросу предопределить семантический класс ожидаемого ответа, тем самым предопределяя и соответствующую цепочку ИКГ, применяемую для его смыслового анализа.

В соответствии с моделью ответа во входном тексте выявляется главное понятие, определяется либо контекст, либо часть контекста, в котором определено это понятие и его взаимосвязи с другими понятиями. Затем выявляются отношения главного понятия с другими понятиями и далее - сами эти понятия. Таким образом выделяется сегмент (параллельно происходит канонизация текста). Этот процесс продолжается до завершения входного текста или пока не встретится признак начала другого сегмента. Новый сегмент определяется по следующим признакам.

Первый признак - поверхностный, признак начала сегмента в тексте. Как правило, обозначается в письменном тексте явно: либо знаком и конкретной функциональной лексемой, либо просто знаком пунктуации. Это символы типа « **,** » - запятая, « **.** » - точка, « — » - тире и т.п. К функциональным лексемам относятся лексемы типа «который», «что», «такой, что» и т.п.

Второй признак - глубинный, содержательно определяющий новый сегмент. Это лексема, отражающая новое отношение, т.е. отношение между понятиями из другого контекста в модели ответа. Это может быть либо новое отношение главного понятия с другими понятиями (линейная структура), либо отношение между другими понятиями (линейная или иерархическая структура). Таким образом, благодаря принципу ожидаемости определенных семантических классов и на основе модели ответа производится сегментация входных текстов и рекурсивно применяются к ним соответствующие цепочки ИКГ. Очевидно, даже для весьма ограниченной ПО нереально предопределить все возможные семантические классы для адекватной сегментации текста и применения к ним соответствующих ИКГ. Всегда будут возможны тексты, которые верны по смыслу, но не поддаются корректной сегментации в рамках данной модели ответа. Однако это не приводит к перестройке базовых концепций, так как система является открытой, знания и обрабатывающие процедуры в ней отделены друг от друга и образование нового семантического класса приводит не к пересмотру и изменению всей совокупности ИКГ, а только к изменению схемы ИКГ или дополнению ее новой ИКГ.

Заключение

Как известно, современные тестирующие программы реализуют, главным образом, выборочный тип ответа и обладают слабым диагностическим арсеналом, особенно в плане автоматизации диагностирования ответов, конструируемых самим обучаемым [Сулейманов, 1999]. Двухуровневый Лингвистический процессор ответов обучаемого, описанный в данной статье, также не является в полной мере той интеллектуальной системой, которая ожидается специалистами как система, способная анализировать и оценивать по смыслу произвольные ответные тексты любой сложности, соответственно, оценивать мыслительные, аналитические способности тестируемого. Однако эта разработка является качественным шагом к интеллектуализации автоматизированного контроля ответа обучаемого за счет возможности ввода обучаемым ответа на заданный вопрос на естественном языке в произвольной форме, без

специальных ограничений, и за счет расширения спектра диагностирования ответа, учитывающего также такие характеристики, как семантическая полнота и корректность. Возможность создания такого ЛП и его эффективность обеспечиваются за счет реализации двух методологических принципов «детерминированности контекста» и «ожидаемости смысла ответа». Очевидно, контекст тестирования, в котором задача ученика - дать ответ на заданный вопрос как можно ближе к тому ответу, который ожидает учитель, чтобы получить хорошую оценку, «заставляет» его отвечать максимально точно, используя те термины, понятия и даже формы определений и фраз, которые дал учитель. Одновременно, задавая вопрос, учитель (система) заранее знает множество значений вопроса (возможные ответы) и может с большой точностью и полнотой сформировать модель ответа, который является ожидаемым по заданному вопросу.

В настоящее время работа над данным проектом продолжается, полностью проработана концептуальная модель двухуровневого лингвистического процессора вопросно-ответных текстов, разработана программная оболочка автоматизации рабочих мест учителя и ученика и реализован прототип ЛП с возможностями контроля ответов на вопросы типов 1-3.

Библиографический список

[Бухараев и др., 1990] Бухараев, Р.Г. Семантический анализ в вопросно-ответных системах / Р.Г. Бухараев, Д.Ш. Сулейманов - Казань: Изд-во Казан. ун-та, 1990.

[Кибрик и др., 1987] Кибрик, А.Е. Моделирование языковой деятельности в интеллектуальных системах / Под ред. А.Е. Кибрика, А.С. Нариньяни; С предисловием А.П. Ершова. -М., Наука. Главная ред. физ.-мат.лит., 1987.

[Сулейманов, 1999] Сулейманов, Д.Ш. Аналитический обзор отечественных и зарубежных работ в области обработки естественного языка в аспекте прагматически-ориентированного подхода /Д.Ш. Сулейманов // Электрон. журнал Казанского госуниверситета “Информационные технологии” [Электронный ресурс]. -1999. – Режим доступа: http://www.kcn.ru/tat_en/science/fttc/ vol000/st.doc (оглавление: contents.htm). - Дата доступа: 8.01.2011.