

# Семантический анализ естественно-языковых текстов в вопросно-ответном режиме

Дж.Ш. Сулейманов, Мад.М. Аюпов

Казанский государственный университет, Академия наук Татарстана

E-mail: [dvdt@telecet.ru](mailto:dvdt@telecet.ru)

## Abstract

*The semantic analysis of the texts within of the question-answer situation is described. In the report the basic principals of the NL-processor's development are considered. Two important methodological principals: "expecting" of the meaning of the question and "determinacy" of the context allowed to implement the efficiency semantic analyzer.*

## 1. Введение

В последние годы активизировались теоретические и прикладные работы по развитию лингвистической стороны проблемы человеко-машинного диалога, а именно, исследование диалога как способа общения и вида текста [1].

Диалог человека с машиной означает интерактивный обмен посланиями между пользователем и диалоговой системой в соответствии с условленным языком диалога и формой диалога для достижения определенной задачи.

Диалоговое взаимодействие пользователя с автоматизированной системой протекает в одном из следующих режимов: 1) активна система, когда на вопросы системы отвечает пользователь, 2) активен пользователь, когда на запрос пользователя определенным образом реагирует система, и наконец, 3) двухсторонне активный диалог. Примером диалоговой модели, наиболее естественно моделирующей вопросно-ответную ситуацию, является вопросно-ответный диалог в автоматизированной обучающей системе (АОС). Здесь имеются следующие особенности, способствующие разработке эффективных прагматически-ориентированных лингвистических моделей.

### 1.1. Особенность входного текста

В АОС текст на естественном языке (ЕЯ) - это множество значений заданного вопроса. Вопрос

накладывает определенные ограничения на форму ответа и его содержание. Ожидаемый объем ответа ограничивается требуемой степенью подробности по заданному вопросу. Сводится к минимуму неоднозначность лексем.

### 1.2. Особенность формальной основы анализа

При контроле ответа обучаемого в АОС для получения эффективных алгоритмов анализа ЕЯ-текста могут быть использованы упрощенные лингвистические модели, ориентированные на информированного (т.е. знакомого с контекстом) "слушающего".

### 1.3. Особенность выходной информации

В результате анализа ответов обучаемого необходимо получить набор параметров, характеризующих степень правильности ответа (диагностику), с целью управления учебным процессом.

В данной работе рассматривается один из подходов к семантическому анализу ответов обучаемого на основе индивидуальных концептуальных грамматик, представляющих собой формальные семантические конструкции ожидаемых значений заданного вопроса.

В условиях вопросно-ответного диалога контекст ответа настолько определен, 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 предварительным знанием. Семантическая классификация вопросов и ответов позволяет заранее противопоставить каждому типу вопроса ограниченный набор допустимых смысловых конструкций, т.е. ответных формул. Можно рассматривать совокупность этих формул, соответствующих конкретному типу вопроса, как некоторую грамматику, кодирующую конструкции, передающие правильный смысл ответа. Нами поставлена и решена задача проведения такой

классификации вопросно-ответных текстов, когда форма и соответствующий смысл входного текста напрямую зависят от типа вопроса.

## 2. Основные понятия и принципы построения

Концептула- это элементарная смыслообразующая единица семантической структуры текста, отражающая роль лексем в значении вопроса и в определенном их сочетании формирующая смысл текста в контексте, детерминированном заданным вопросом.

Схемы сочетания концептул, соответствующие правильной передаче ожидаемого смысла, названы индивидуальными концептуальными грамматиками (ИКГ). Использование понятия концептуальной грамматики дает возможность сводить выявление семантического содержания ответа к анализу его грамматического соответствия некоторой ИКГ [2]. Построение лингвистического процессора базируется на следующих принципах.

### Методологические принципы:

1. Принцип *детерминированности контекста*. В силу активности, система «погружает» пользователя в 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 системе 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.

*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.

*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".

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

### Выделение типов понятий и соответствующих концептул.

SS - *концептула*, отражающая *главное понятие*, т.е. понятие, относительно которого задан вопрос. Главное понятие названо *Фактором*.

Сложные тексты могут содержать несколько понятий, связи которых раскрываются в ответных предложениях, каждый из которых в процессе анализа 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.

SA - концептула, отражающая понятие-аргумент.

SP - концептула, отражающая понятие-результат.

Обозначим через  $K_S$  множество концептул, отражающих различные типы понятий, т.е.,  $K_S = \{SS, SS(i), SO, SA, SP\}$ .

### Выделение типов отношений и соответствующих концептул.

Введем понятие Типового Отношения (ТО). *ТО - это семейство отношений, отражающих однотипный смысл и раскрывающих определенный признак понятий предметной области (ПО)*. Например, отношения, выраженные лексемами типа 'играет', 'спит', 'плавает', объединяются в семейство ТО СОСТОЯНИЕ по признаку: выражать состояние понятия.

Аналогично определяются ТО ДЕЙСТВИЕ, СОСТАВ, ВКЛЮЧЕНИЕ, ВРО (*ВРеменное Отношение*), ПРО (*Пространственное Отношение*), КЛО (*КоЛичественное Отношение*), КЧО (*КаЧественное Отношение*). Введем понятие *Составного Отношения (СО) Функция*. *СО Функция - это устойчивая комбинация двух ТО ДЕЙСТВИЕ:*

действия, направленного на аргумент (т.е. отношение SS к SA) и действия, направленного на результат (т.е. отношение SS к SP). В следующем тексте раскрывается СО ФУНКЦИЯ понятия  $S_1$ : ‘ $S_1$  переводит  $S_2$  в  $S_3$ ’.

Введем формальные обозначения для концептуал, соответствующих типов отношений.

$R_C$  - это концептула, соответствующая ТО СОСТОЯНИЕ;

$R_{СОСТ}$  - ТО СОСТАВ;

$R_{ВКЛ}$  - ТО ВКЛЮЧЕНИЕ;

$R_D$  - ТО ДЕЙСТВИЕ;

$R_{ВРО}$  - ТО ВРО;

$R_{ПРО}$  - ТО ПРО;

$R_{КЛО}$  - ТО КЛО;

$R_{КЧО}$  - ТО КЧО;

$R_{SO}$  - 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. Через  $K_R$  обозначим множество концептуал, отражающих различные типы отношений, т.е.

$K_R = \{ R_C, R_{СОСТ}, R_{ВКЛ}, R_D, R_{ВРО}, R_{ПРО}, R_{КЛО}, R_{КЧО}, R_{SO}, R_{OS}, R_A, R_P \}$ . Здесь  $R_{SO}, R_{OS}$  принимают значения из следующего множества:  $\{ R_C, R_{СОСТ}, R_{ВКЛ}, R_D, R_{ВРО}, R_{ПРО}, R_{КЛО}, R_{КЧО} \}$ .

Таким образом, концептулы первой группы (обозначим  $K_1$ ) включают следующее множество прагматических ролей:  $K_1 = K_S \cup K_R$ .

**Грамматические роли лексем и частей** отражают грамматические признаки естественного языка (элементы грамматики, например, суффиксы, союзы, предлоги и др.). Выделяются следующие грамматические признаки (для русского языка) и соответствующие концептулы:

1. Предлог перед SA (например, предлоги из, от, с и т.п.) отражается концептулой  $GPA$ .
2. Предлог перед SP (например, предлоги в, на, к и т.п.) отражается концептулой  $GP_P$ .
3. Грамматические модификаторы: либо лексемы типа ‘чем’, ‘нежели’ и т.п. после лексемы, выражающей отношение, либо падежные окончания слова после лексемы, выражающей понятие - отражаются концептулой  $Gm$ .
4. Функциональная лексема, обозначающая признак начала причинной части ответа, в котором раскрывается причинно-следственное отношение. Например, лексемы ‘потому что’, ‘так как’, ‘если’ и т.п. отражаются концептулой  $Gf_1$ .
5. Функциональная лексема, обозначающая признак начала следственной части ответа, в котором раскрывается причинно-следственное

отношение. Например, лексемы ‘то’, ‘тогда’, ‘значит’ и т.п. отражаются концептулой  $Gf_2$ .

Таким образом, концептулы второй группы (обозначим,  $K_2$ ) включают следующее множество грамматически ролей:

$K_2 = \{ GP_A, GP_P, Gm, Gf_1, Gf_2 \}$ .

**Специальная роль лексем** отражает специфику элементов ответа на конкретный вопрос в заданной предметной области, т.е. в определенном контексте.

Выделяются следующие специальные лексемы и отражающие их концептулы:

1. *Необязательная лексема*, т.е. лексема, отсутствие или наличие которой в ответе не влияет на смысл ответа. Отражается концептулой  $LN$ .
2. *Запрещенная лексема*, т.е. лексема, наличие которой в ответе недопустимо (рассматривается как ошибка). Отражается концептулой  $LZ$ .
3. *Неопределенная лексема*, т.е. лексема, не предусмотренная АВТОРОм курса. Отражается концептулой  $LNE$ .
4. *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.

Таким образом, концептулы третьей группы (обозначим  $K_3$ ) включают следующее множество специальных ролей:  $K_3 = \{ LN, LZ, LNE, LI_S, LI_O, LI_A, LI_P, LI_R \}$ .

Введем понятие Ключевого Параметра (КП). **КП** - это определенная последовательность символов, предусмотренная АВТОРОм и обязательная в ответе.

### 3. Семантическая классификация вопросно-ответных текстов

В целях сокращения пространства поиска ответа на вопрос и унифицированного применения индивидуальных концептуальных грамматик осуществляется семантическая классификация вопросно-ответных текстов по сложности раскрываемых отношений.

**I. Вопросы, требующие явного задания в ответе КП** (отношения явно заданы в вопросе). *Например*: ‘Напишите программу вычисления функции на Паскале’. Этому типу вопросов соответствуют классы ответов, в которых обязательно явно содержатся **КП**. Например, ответы выборочного типа;

ответы типа "ДА - НЕТ"; ответы фиксированно-конструируемого типа; численные ответы и т.п.

**II. Вопросы, требующие раскрытия в ответе ТО одного ФАКТОРА.** Например: 'Что выполняется раньше: компиляция или загрузка?'. Выделяются следующие классы ответов, раскрывающие одноименные ТО: СОСТАВ, ВКЛЮЧЕНИЕ, ДЕЙСТВИЕ, СОСТОЯНИЕ, ВРО, ПРО, КЛО, КЧО.

**III. Вопросы, требующие раскрытия в ответе СО одного ФАКТОРА.** Например: 'Что делает компилятор?'. Такому типу вопросов соответствуют классы ответов, в которых ФАКТОР раскрывается через СО. Например, выделен класс ответов ФУНКЦИЯ, в котором ФАКТОР раскрывается через его СО ФУНКЦИЯ:  $S_i$  переводит  $S_{i+6}$  в  $S_{i+7}$ .

**IV. Вопросы, требующие раскрытия в ответе произвольной комбинации ТО и/или СО одного ФАКТОРА.** Например: 'Дайте описание компилятора'. Этим вопросам соответствуют классы ответов, в которых ФАКТОР раскрывается через его ТО и/или СО.

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

**V. Вопросы, требующие раскрытия в ответе более чем одного ФАКТОРА.** Например: 'Расскажите о компиляторе'. Этому типу вопросов соответствуют ответы, в которых ФАКТОР меняется в процессе просмотра ответа.

#### 4. Индивидуальные концептуальные грамматики

Семантическим классам ответов соответствуют присущие им схемы сочетания концептуал, передающие глубинный смысл ответов данного класса. Схемы сочетания концептуал, соответствующие правильной передаче ожидаемого смысла, названы индивидуальными концептуальными грамматиками (ИКГ).

ИКГ классов ответов на вопросы типа 1 (обозначим, ИКГ1) имеет следующее формализованное представление:  $\langle ИКГ1 \rangle ::= SS^*$ . Здесь и далее в ИКГ символ \* («звездочка») - признак повтора.

Ответы семантических классов на вопросы типа II определяются схемой понятие—отношение—понятие. В этих ответах значащими являются понятия и ТО между ними.

Классы ответов, раскрывающих ТО понятий,

имеют идентичные ИКГ (обозначим ИКГ2):

$$\langle ИКГ2 \rangle ::= SS^* \rightarrow R_{SO} \rightarrow SO^* | R_{SO} \rightarrow SS^* | SS^* \rightarrow R_{SO} \rightarrow SO^* | SO^* \rightarrow R_{OS} \rightarrow SS^* | R_{OS} \rightarrow SO^* | SO^* \rightarrow R_{OS} \rightarrow (R_{SO} \rightarrow Gm)^* \rightarrow SO^* | R_{SO} \rightarrow (SO \rightarrow Gm)^* | (R_{SO} \rightarrow Gm)^* \rightarrow SO^* | R_{OS} \rightarrow (SS \rightarrow Gm)^*$$

ИКГ классов ответов ФУНКЦИЯ имеет следующее описание:

$$\langle ИКГ ФУНКЦИЯ \rangle ::= [SS^* \rightarrow ](R_A^* \rightarrow (GP_P \rightarrow SP^* \rightarrow SA^* | SA^* \rightarrow GP_P \rightarrow SP^*) | RP^* \rightarrow (GP_A \rightarrow SA^* \rightarrow SP^* | SP^* \rightarrow GP_A \rightarrow SA^*)) | ((GP_P \rightarrow SP^* \rightarrow R_A^* \rightarrow SA^* | SA^* \rightarrow RA^* \rightarrow GP_P \rightarrow SP^*) | (GP_A \rightarrow SA^* \rightarrow R_P^* \rightarrow SP^* | SP^* \rightarrow R_P^* \rightarrow GP_A \rightarrow SA^*))$$

ИКГ классов ответов, соответствующих вопросам типа IY-V, являются произвольной комбинацией ИКГ классов ответов на вопросы типа II и III.

$\langle ИКГ4 \rangle ::= \langle T2 \rangle^* | \langle T3 \rangle^* | \langle T2 \rangle^* \rightarrow \langle ИКГ4 \rangle^* | \langle T3 \rangle^* \rightarrow \langle ИКГ4 \rangle$ , здесь T2 обозначает ИКГ классов ответов на вопросы типа II; T3 - ИКГ классов ответов на вопросы типа III. Для каждого класса ответов на вопросы Y типа разрабатывается ИКГ, присущая только данному классу. Рассмотрим ответ, относящийся к классу ДЕТАЛИЗАЦИЯ: 'Si переводит  $S_{i+1}$  в  $S_{i+2}$ , который находится на  $S_{i+3}$ , который выполняется раньше  $S_{i+4}$ , который больше, чем  $S_{i+6}$ , который стирается'.

Канонизированное представление ответа имеет следующий вид:

$$SS(1) \rightarrow SS(1)ДЕТ \rightarrow R_A \rightarrow SA \rightarrow GP_P \rightarrow SP \rightarrow SS(2)ДЕТ \rightarrow R_{SO} \rightarrow SO \rightarrow SS(3)ДЕТ \rightarrow R_{SO} \rightarrow SO \rightarrow SS(4)ДЕТ \rightarrow R_{SO} \rightarrow Gm \rightarrow SO \rightarrow SS(5)ДЕТ \rightarrow R_{SO}$$

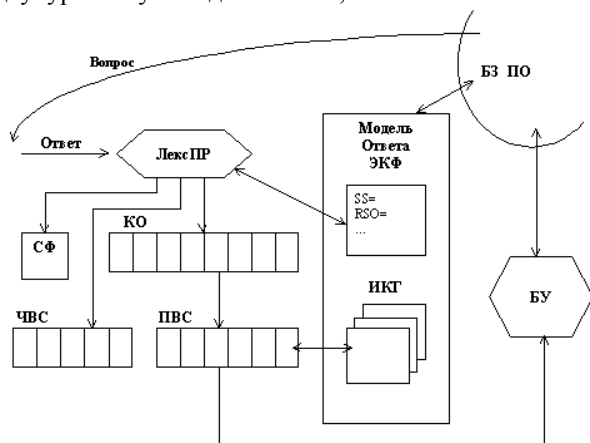
Здесь концептула SS(1) есть ФАКТОР1 - отражает понятие  $S_i$ ,  $S_{i+1}$  есть SA, ФАКТОР2 SS(2)=SP отражает понятие  $S_{i+3}$ , ФАКТОР3 SS(3)=SO -  $S_{i+3}$ , ФАКТОР4 SS(4)=SO - понятие  $S_{i+4}$ , ФАКТОР5 SS(5)=SO - понятие  $S_{i+5}$ . КО 'переводит' относится к СО ФУНКЦИЯ -  $R_A$ , 'находится на', 'раньше', 'больше', 'стирается' относятся соответственно к ТО ПРО, ВРО, КЛО и СОСТАВ. Все ТО обозначаются концептулой  $R_{SO}$ . Предлог 'в' есть  $GP_P$ .

Таким образом, ИКГ класса ДЕТАЛИЗАЦИЯ (обозначим ИКГ5) имеет следующее описание:  $\langle ИКГ5 \rangle ::= ([SS(i)^* \rightarrow ] SS(i)ДЕТ \rightarrow \langle G(i)^* \rangle^*)$ .

Здесь SS(i) - концептула, отражающая i-й ФАКТОР; SS(i)ДЕТ обозначение начала детализации i-го ФАКТОРА; G(i) - часть ИКГ класса ДЕТАЛИЗАЦИЯ с постоянным SS(i), т.е. грамматика G(i) есть ИКГ4 и раскрывает только i-й ФАКТОР.

#### 5. Функционирование семантического анализатора

Система интерпретации ЕЯ-текстов в контексте, управляемом системой, включает лексический процессор, семантический интерпретатор и двухуровневую модель ответа, показанные на Рис. 1.

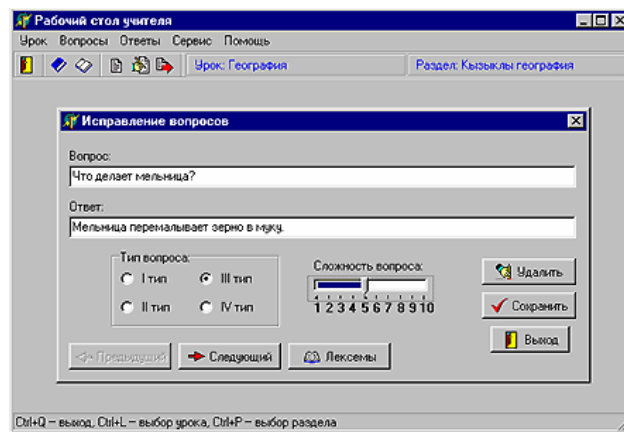


**Рис. 1. Крупноблочная схема вопросно-ответного лингвопроцессора.**

Интерпретация ответного текста происходит следующим образом. Ответ поступает в *лексический процессор (ЛексП)* и на основе экземпляра *фрейма (ЭКФ) модели ответа (МО)* переводится в *канонизированное представление ответа (КО)* в виде последовательности концептуал. Часть информации на лексическом уровне может представлять интерес для дальнейшего разбора (например, для проверки на непротиворечивость с ожидаемой частью ответа), поэтому накапливается в *специальных файлах (СФ)*. Одновременно формируется *частичный вектор ситуации (ЧВС)*, отражающий промежуточную диагностику ответа. Далее канонический текст интерпретируется с привлечением *ИКГ*. Результат формируется в виде некоторого *полного вектора ситуации (ПВС)*, по которому в *блоке управления (БУ)* принимается управляющее действие системы на основе соответствующих опций, заполненных предварительно преподавателем и содержащихся в базе знаний предметной области (*БЗ ПО*).

Разработана интерфейсная оболочка, обеспечивающая удобное взаимодействие при эксплуатации семантического анализатора для преподавателя (при подготовке базы вопросов, модели ответов и других опций) и обучаемого (при

ответе на вопросы системы). На рис. 2 показан фрагмент автоматизированного рабочего места преподавателя. Программа реализована на языке Delphi и обеспечивает семантический анализ ответов и интерфейс на русском, татарском и английском языках.



**Рис. 2. Фрагмент интерфейса АРМ преподавателя.**

## 6. Заключение

В статье предложен подход к разработке семантического анализатора естественно-языковых текстов в диалоговых обучающих системах в условиях детерминированного контекста, определяемого заданным вопросом. На ряде иллюстрированных примеров изложены особенности и преимущества анализа вопросно-ответных текстов в ситуации «ожидаемого текста» и «детерминированного контекста». В настоящее время разрабатывается развитая версия семантического анализатора в условиях двуязычных вопросно-ответных текстов на татарском и русском языках.

## 7. Литература

- [1] Андрусенко Т.Б. Лингвистические структуры в компьютерных учебных средах. -Киев: Наукова Думка, 1994. -160 с.
- [2] Бухараев Р.Г., Сулейманов Д.Ш. Семантический анализ в вопросно-ответных системах. - Казань: Изд-во Казан. ун-та. - 1990. -124 с.