КОМПЬЮТЕРНАЯ ВОПРОСНО-ОТВЕТНАЯ СИСТЕМА СМЫСЛОВОГО АНАЛИЗА ТЕКСТА

Богатов Н.М., Родоманов Р.Р.

Кубанский государственный университет, г.Краснодар

Создана вопросно-ответная система (ВОС), способная обмениваться информацией между человеком и компьютерной диалоговой системой. В экзаменуемом вопросно-ответном диалоге система активна (задает вопросы), а пользователь пассивен (отвечает). Анализатор ответов в ВОС проверяет правильность ответа обучаемого на соответствие с ожидаемым. В случае неточного или неправильного ответа человека, ВОС задает наводящий вопрос, который указывает на пропущенную истину или неправильно написанное слово.

Диалог между ВОС и человеком происходит на естественном языке. На естественном языке передается информация о различных явлениях, свойствах, событиях и т.д. Одна из задач ВОС – анализ ответов обучаемого [1]. Базовые принципы построения семантического интерпретатора вопросно-ответных текстов на естественном языке обсуждались в работе [2].

Смысл текста, при анализе ответа экзаменуемого, система «понимает» сопоставляя содержимое ответа и фрагменты семантической сети. Узлы семантической сети представлены множеством часто встречаемых понятий текста, слов и устойчивых словосочетаний.

Для автоматизации анализа ответов создана программа лингвистического анализа русскоязычных текстов [3], предназначенная для определения смысла предложения по смысловому значению слов.

Среди множества предложений естественного языка лишь весьма не многие являются «правильными», то есть адекватно отражают синтаксическое строение предложения. Одну и туже мысль можно выразить разными словами. Работа ВОС основывается на анализе поступивших фраз по смыслу.

В результате анализа ответов обучаемого необходимо получить набор параметров, характеризующих степень правильности ответа.

Целью работы является создание компьютерной вопросно-ответной системы, способной выполнять смысловую проверку текста на естественном языке и оценивать знания экзаменуемого студента.

Семантическое представление вопросно-ответной системы

Для передачи и приема знаний между ВОС и человеком служит естественный язык. На естественном языке передается информация о различных явлениях, свойствах, событиях и т.д.

Знания, записанные в семантическую сеть, подразделяются на классы. Классом знаний в семантической сети является группа объектов, явлений объединенных общностью признаков. Примером могут служить следующие слова: объект, свойство и т.д. Понятие объект, разделяется на два подкласса, одушевленный и неодушевленный. К подклассу одушевленные относятся все живые существа, птицы, животные, насекомые и т.д. К подклассу неодушевленные относятся предметы, явления действительности и т.д.

В качестве основания для такой классификации выступает принцип взаимозаменяемости слов или словосочетаний, т.е. нескольких слов, объединенных подчинительной связью. При этом замена одного слова или словосочетания другим не приводит к превращению осмысленного текста к бессмысленному тексту. В качестве примера рассмотрим ветвь – *вещество, металл, железо.* В предложении «*корпус прибора выполнен из железа*», возможно, заменить слово «*железо*», словом «*металл*», при этом смысл предложения останется неизменным.

Класс в семантической сети возникает при поступлении новых знаний, которые по своим признакам не могут относиться к существующим классам.

Процесс общение и понимание ВОС

Понимание некоторого языкового сообщения ВОС связано, прежде всего, с его распознаванием, или декодированием. В нашей системе распознавание происходит следующим образом. Система обработки текста выполняет морфологический и синтаксический анализ текста. Первоначально полученное сообщение расчленяется на составляющие его элементы, из которых по заранее принятым правилам строится описание данной последовательности. Далее происходит семантический анализ текста, в процессе которого результат описания сравнивается в семантической сети с эталонами и в случае совпадения их с анализируемыми знаковыми последовательностями происходит ассоциативная активация соответствующих областей семантической сети.

Понимание нельзя свести к простому запоминанию или умению пересказать полученное сообщение. Можно сохранить в памяти какое либо сообщение, более или менее точно воспроизвести его, тем не менее, упуская главное – смысл.

Ручной режим пополнения базы знаний

Пополнение базы знаний в ручном режиме происходит в следующей последовательности:

составляется предложение несущее необходимую информацию,

создается дерево составляющих,

дерево составляющих переносится в семантическую сеть.

В качестве примера рассмотрим ответ на вопрос: Что называется дифференциальной нелинейностью ЦАП?

Смысл ответа содержится в предложении: Дифференциальная нелинейность – это отклонение действительных ступеней квантования от их среднего значения.

При ответе студентом предложение может быть создано с помощью других слов, возможно изменение последовательности слов в предложении, но главное, чтобы оставался смысл. Поэтому создаваемое предложение, с помощью которого загружаются знания в семантическую сеть, должно содержать только основные слова, отображаемые смысловую форму предложения.

Дерево составляющих для предложения: ((Дифференциальная нелинейность) – (это (отклонение (действительных (ступеней квантования))) (от (их (среднего значения))))), приведено на рис.1.

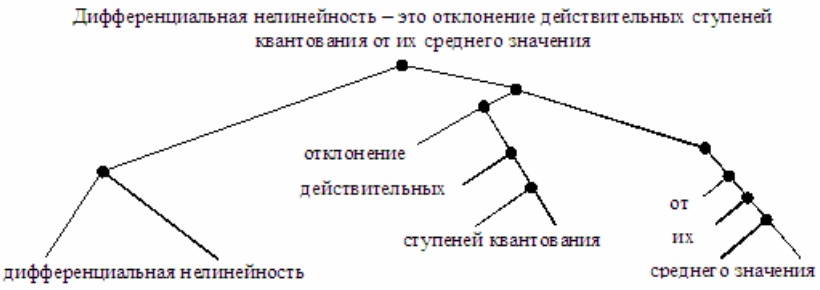


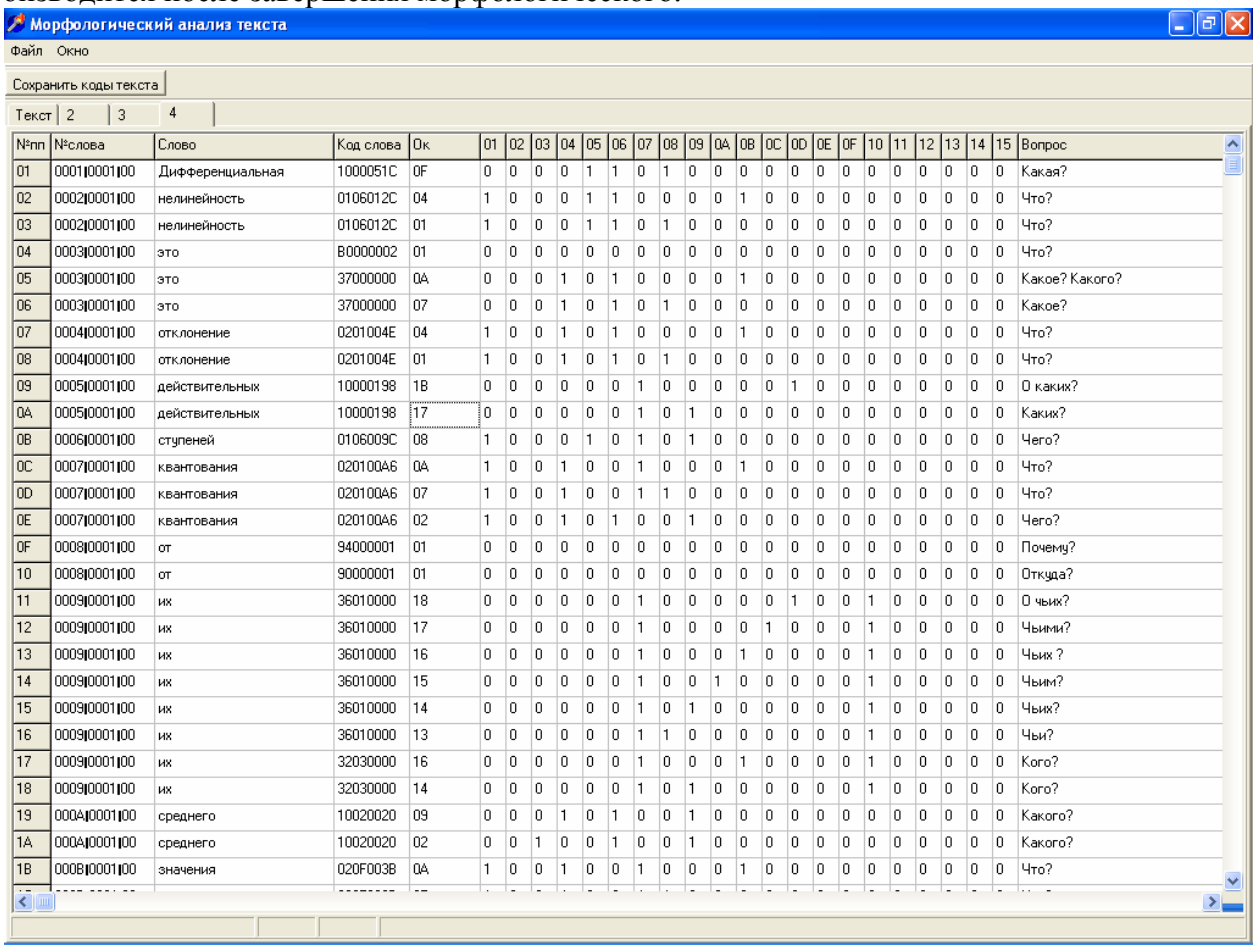
Рис.1. Пример дерева составляющих

Построенное дерево словосочетаний указывает в предложении словосочетания разных «уровней», но не предоставляет при этом никаких различий среди словосочетаний одного уровня. «Главенствование» одного словосочетания над другим, необходимо определять интуитивно.

Этот способ достаточно надежен, так как знания заносятся человеком, но трудоемок. При большом объеме загружаемых данных человек устает и возможно проявление человеческих факторов.

Автоматический режим пополнения базы знаний

В автоматическом режиме пополнение базы знаний начинается с морфологического анализа предложения. В задачу морфологического анализа входит идентификация слов предложения с морфологическим словарем системы. В процессе морфологического анализа определяется часть речи, категория, форма слова и вопрос. Результатом морфологического анализа становится таблица, приведенная на рис. 2. В таблице отображаются все возможные виды слов с различными окончаниями. Информация о категории и форме слова содержится в колонках 01h – 15h. Эти данные необходимы для синтаксического анализа, который производится после завершения морфологического.



№пп	№слова	Слово	Код слова	Ок	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F	10	11	12	13	14	15	Вопрос
01	0001|0001|00	Дифференциальная	1000051C	0F	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Какая?
02	0002|0001|00	нелинейность	0106012C	04	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Что?
03	0002|0001|00	нелинейность	0106012C	01	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Что?
04	0003|0001|00	это	B0000002	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Что?
05	0003|0001|00	это	37000000	0A	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Какое? Какого?
06	0003|0001|00	это	37000000	07	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Какое?
07	0004|0001|00	отклонение	0201004E	04	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Что?
08	0004|0001|00	отклонение	0201004E	01	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Что?
09	0005|0001|00	действительных	10000198	1B	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	О каких?
0A	0005|0001|00	действительных	10000198	17	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Каких?
0B	0006|0001|00	ступеней	0106009C	08	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Чего?
0C	0007|0001|00	квантования	020100A6	0A	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Что?
0D	0007|0001|00	квантования	020100A6	07	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Что?
0E	0007|0001|00	квантования	020100A6	02	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Чего?
0F	0008|0001|00	от	94000001	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Почему?
10	0008|0001|00	от	90000001	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Откуда?
11	0009|0001|00	их	36010000	18	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	О чьих?
12	0009|0001|00	их	36010000	17	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	Чьими?
13	0009|0001|00	их	36010000	16	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	Чьих ?
14	0009|0001|00	их	36010000	15	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	Чьим?
15	0009|0001|00	их	36010000	14	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	Чьих?
16	0009|0001|00	их	36010000	13	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	Чьи?
17	0009|0001|00	их	32030000	16	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	Кого?
18	0009|0001|00	их	32030000	14	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	Кого?
19	000A|0001|00	среднего	10020020	09	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Какого?
1A	000A|0001|00	среднего	10020020	02	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	Какого?
1B	000B|0001|00	значения	020F003B	0A	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	Что?

Рис.2. Таблица результата морфологического анализа

Задача синтаксического анализа – определить структуру входного предложения в соответствии с грамматикой русского языка и произвести коррекцию результатов морфологического анализа. Коррекция результатов морфологического анализа производится в соответствии со структурой предложения. При нескольких значениях слов определяется одно, наиболее сочетаемое со значениями соседних слов. В синтаксическом словаре находятся таблицы определяющие правила связи слов в предложении. Информация о синтаксическом строении предложения представляет собой набор сведений о «главенствовании» одних слов над другими. Словосочетание возникает на основе подчинительной связи. Одним из способов изображения синтаксической структуры предложения является дерево подчинения. Дерево подчинения создается на основе таблицы классов слов.

После завершения синтаксического анализа создается окно, показанное на рис. 3.

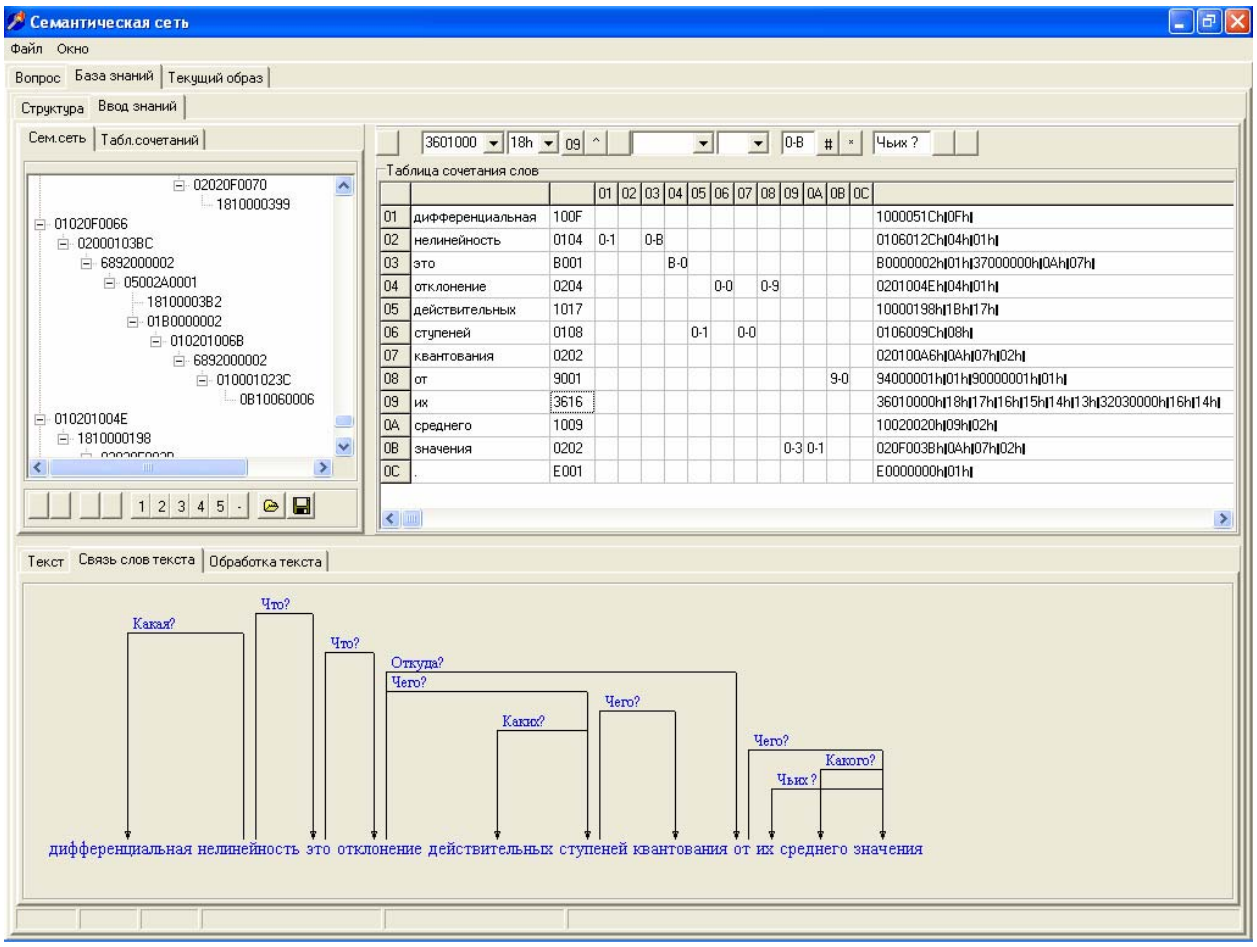


Рис.3. Окно ввода знаний в семантическую сеть

В процессе морфологического и синтаксического анализа при вводе новых текстов, возможно определение неопознанных слов и словосочетаний.

В задачу семантического анализа текста проводимую оператором, входит «обучение» вопросно-ответной системы, т.е. ввод слов и словосочетаний которые не были определены в автоматическом режиме. Словосочетания вводятся в таблицу сочетания слов расположенную в окне ввода знаний. По оси строк и столбцов таблицы располагаются все слова входящие в предложение. Поочередно происходит анализ возможности сочетания всех слов предложения. Если возможность сочетания определена, в ячейке пересечения устанавливается код сочетания. Код сочетания определяется программно. Код сочетания состоит из двух символов, первый символ определяет код класса первого слова сочетания, а второй символ определяет код второго слова сочетания. Например, сочетание слов «*дифференциальная нелинейность*» соответствует коду «0-1» - сочетание слов нулевого и первого классов, т.е. «существительное – прилагательное».

Для удобства и контроля вводимых знаний автоматически строится дерево подчинения слов, расположенное в нижней части окна ввода знаний. Связь слов в предложении довольно легко контролировать с помощью вопросов, которые определяются для каждого слова в процессе морфологического анализа.

Ввод новых слов производится в словарь, который является основной базой данных вопросно-ответной системы из списка, который автоматически создается при морфологическом анализе текста.

Данные, поступающие в словарь сочетания слов, используются при дальнейшем синтаксическом анализе текста, как при вводе знаний, так и при анализе ответов.

Загрузка знаний, полученных в процессе семантического анализа, происходит при помощи окна иерархического дерева, позволяющего создать, удалить или изменить узел или ветвь семантической сети.

Заключение

Создана система, выполняющая смысловую обработку ответа, позволяющая вводить ответ на естественном языке. Вводимый текст ответа может быть построен с помощью различных предложений, отражающих смысл вопроса. Использование предлагаемой системы при самообучении человека, позволяет производить самоконтроль уровня знаний, уменьшить влияние человеческого фактора при тестировании.

Пополнение базы знаний системы может происходить в ручном или автоматическом режиме.

ВОС используется при аттестации студентов физико-технического факультета КубГУ. Все ответы студентов после тестирования сохраняются в базе данных, что позволяет в дальнейшем анализировать, как индивидуальные знания каждого студента, так и всей группы. Это помогает выявлять «слабые моменты» в изложенном материале. В процессе работы планируется при дальнейших тестированиях индивидуально каждому студенту повторно задавать вопросы, на которые были получены неправильные ответы, оповещать преподавателя о пробелах в знаниях студентов, указывая ему на фрагменты неусвоенного студентами материала.

Литература

1. *Богатов Н.М., Родоманов Р.Р.* Автоматизация вопросно-ответного диалога в обучающей системе. Современные наукоёмкие технологии. 2006. № 4. С. 23-25.
2. *Сулейманов Дж.Ш.* Исследование базовых принципов построения семантического интерпретатора вопросно-ответных текстов на естественном языке в АОС. Educational Technology & Society 4(3) (2001) 178-192.
3. *Родоманов Р.Р., Богатов Н.М.* Программа лингвистического анализа русскоязычных текстов «ПЛАРТ». Свидетельство об официальной регистрации программы для ЭВМ №2005612382 12.09.2005.