УДК: 519.688

М о з г о в о й М. В. Простая вопросно-ответная система на основе семантического анализатора русского языка.

В работе описываются идеи построения вопросно-ответной системы, иллюстрирующей возможное применение семантического анализатора В. А. Тузова. Обсуждается алгоритм выявления фраз, являющихся ответами на задаваемые пользователем вопросы. Анализируются особенности строения предложений русского языка применительно к задаче построения вопросно-ответной системы.

Библиогр. 11 назв. Табл. 4.

519.688 Вестник СПбГУ. Сер. 10, 2005, вып. 2

М. В. Мозговой

ПРОСТАЯ ВОПРОСНО-ОТВЕТНАЯ СИСТЕМА НА ОСНОВЕ СЕМАНТИЧЕСКОГО АНАЛИЗАТОРА РУССКОГО ЯЗЫКА

1. Введение. Вопросно-ответная (диалоговая) система — это программа, которая умеет обрабатывать введенный пользователем вопрос на естественном (русском, английском и т. д.) языке и печатать осмысленный ответ. Некоторые диалоговые системы пытаются создать атмосферу полноценного «общения» между человеком и компьютером на естественном языке. Разумеется, на практике компьютеру до настоящей «интеллектуальности» все еще очень далеко, но существуют задачи, в которых она и не требуется. Например, для того чтобы ответить на вопрос, заданный к некоторому тексту, требуется лишь грамотно «разложить по полочкам» вопрос и определить, какое именно предложение текста содержит в себе ответ.

Описываемая здесь вопросно-ответная система является довольно простым примером использования *семантического анализатора*, разработанного проф. В. А. Тузовым [1]. В настоящее время постановка задачи звучит так. Дается некоторый текстовый файл, состоящий из предложений на русском языке. Пользователь вводит запрос, по форме являющийся обычным вопросительным предложением. Система анализирует текстовый файл (предложение за предложением), находит фразу, содержащую в себе ответ, и выдает ее часть, собственно отвечающую на вопрос, пользователю. Если ответов найдено несколько, все они печатаются.

2. Краткий обзор существующих разработок. Авторы большинства создаваемых в настоящее время вопросно-ответных систем естественным образом ориентируются на английский язык. Однако любая серьезная система должна каким-либо образом анализировать структуру запроса, опираясь на знания о языке, на котором он сформулирован. Поэтому произвести объективное сравнение систем, рассчитанных на разные языки, практически

невозможно. Тем не менее можно изучить принципы работы любой данной системы и сделать вывод о глубине производимого ею анализа предложений запроса и входного текста.

Для английского языка уже существуют синтаксические анализаторы (например, CMU Link Parser [2]) и справочные системы по словам языка (WordNet [3]). В частности, WordNet является очень полезным обобщенным инструментом, и в настоящее время ведется его адаптация для русского языка [4].

Системы CMU Link Parser и WordNet используются, например, в вопросно-ответной системе, описанной в работе [5], совместно с поисковой машиной общего назначения Managing Gigabytes.

Программа QA-LaSIE [6], разработанная в Шеффилдском университете специально для TREC QA Track 9, интересна прежде всего тем, что ее вывод пользователю представляет собой не целые предложения и абзацы, а «точные» (по мнению системы) ответы на поставленный вопрос. QA-LaSIE умеет выполнять частичный семантический анализ, производя «квазилогическую форму», стоящую, по словам авторов, на полпути между предложением и его полным семантическим описанием.

Одной из наиболее развитых разработок является созданная в Далласе система Lasso [7]. Введенный запрос анализируется, определяются его тип («what-who», «what-when», «how long», «how rich» и т.п.), запрашиваемая сущность и тип ответа («DATE», «LOCATION», «PERSON», «NUMBER», ...). Алгоритм определения приведенных элементов основывается на последовательном применении восьми различных «эвристик». Для нахождения ответа система просматривает документы в поисках вхождения интересующих типов сущностей. Типы сущностей для слов из входных документов определяются при помощи «лексико-семантического анализатора», основанного на словарях Gazetteer и WordNet. Найденные ответы ранжируются по релевантности при помощи специальной оценочной функции.

К сожалению, системы подобного уровня для русского языка в настоящее время находятся лишь в стадии разработок. Существуют коммерческие разработки вроде TextAnalyst фирмы Microsystems Ltd, основанные на статистических методах, но они, по сути дела, производят поиск слов, входящих в запрос, и релевантность результатов часто оказывается низкой.

Вместе с тем есть объективные предпосылки для развития

отечественных систем. В работе [8] приводится описание морфологического анализатора русского языка на основе «обобщенного анализатора» — системы AGFL. Статья [9] описывает подход «частичного синтаксического разбора», включающего в себя выделение ключевых понятий и смысловых связей между ними.

Синтаксический анализ русских предложений умеет выполнять продукт DictaScope фирмы Dictum Software. Фирма ведет также разработку диалоговой системы Dictum, но она еще не существует даже в ознакомительной версии.

3. Семантический анализатор В. А. Тузова. Как уже упоминалось выше, описываемая здесь вопросно-ответная система основана на семантическом анализаторе В. А. Тузова, поэтому понять принципы ее работы без краткого знакомства с ним невозможно.

Вообразим некоторый «семантический язык», являющийся, с одной стороны, совершенно формальным (как язык программирования), а с другой — пригодным для описания чего бы то ни было не в меньшей степени, чем любой естественный язык (вопрос художественной красоты языка, разумеется, не рассматривается). Преобразование предложений естественного (в нашем случае русского) языка в конструкции языка семантического и есть задача семантического анализатора.

Если не центральным, то заведомо самым трудоемким элементом анализатора является *семантический словарь*. В принципе, это обычный толковый словарь, но каждая его «статья» описана строго формально, что делает ее пригодной для анализа компьютером.

Многие слова, такие как «табурет», «аббат» или «трасса», через другие не определяются (в том смысле, что табурет не является производным от чего-либо). Такие слова считаются *базовыми* и в словаре не описываются. Для каждого базового слова указывается лишь его *класс* (см. табл. 1).

Небазовые слова определяются через базовые путем применения к ним так называемых *базисных функций* (см. табл. 2). Множество базовых слов состоит из десятков тысяч элементов, базисных же функций меньше 50.

Аргументы *z1*, *z2*, ... в функциях («аргументы» описываемого слова, см. табл. 3) формируют контекст, в котором функция имеет смысл. Например, у слова «жечь»(точнее, у одного из его

Таблица 1. **Примеры базовых слов**

Слово	Класс
Иванов	Существительное / Физический объект / Живой / Человек / Индивид / ФИО / Фамилия
Протон	Существительное / Физический объект / Природа / Микромир
Радость	Существительное / Психика / Душа / Чувства / Довольство / Радость

Таблица 2. **Некоторые базисные функции**

Название	Описание
Caus(x, y)	x делает так, чтобы y (x каузирует y)
Copul(x, y)	x есть y
Lab(x, y)	x подвергается действию y
Ne(x)	отрицание x

определений) два аргумента. Некий объект *z1* может жечь другой объект *z2*. Конкретные значения этих аргументов будут известны лишь во время анализа реального предложения.

Таблица 3. **Примеры определения слов**

Слово	Определение
жечь	Caus(*z1*, Lab(*z2*, ОГОНЬ))
аморфный	NeHab(*z1*, ФОРМА)
мулатка	Copul(МУЛАТ, ЖЕНЩИНА)

Получается, что формулу слова «жечь» можно расшифровать следующим образом: «*z1* делает так, что *z2* подвергается действию огня». Аналогично, «аморфный» — это «*z1*, не имеющий формы», а «мулатка» — это «мулат, который является женщиной».

Помимо определения смысла тех или иных слов, анализатор должен выявить структуру самого предложения. Предложения являются суперпозициями слов, их составляющих [10]. Следовательно, более формально цель работы семантического анализатора можно выразить так: для каждого предложения вывести его представление в виде суперпозиции слов, для каждого слова вывести его семантическое описание.

Рассмотрим упрощенную (без разбора отдельных слов) распечатку с результатами работы семантического анализатора:

```
Президент заинтересован в улучшении отношений с союзниками.
.....................................................
@Крат заинтересован<X002.002>
  (@Им Президент<X001.001>,
   @вПред в<X003.101>
    (@Пред улучшении<X004.001>
      (@Род отношений<X005.003>
        (@сТв с<X006.028>
          (@Тв союзниками<X007.001>)
  )))).
```

4. От семантического анализатора к вопросно-ответной системе. При описании различных вопросно-ответных систем обычно упоминают так называемые «уровни понимания» [11]. На начальном уровне можно рассматривать предложения независимо друг от друга и не производить никакого дополнительного анализа слов. Рассмотрим текст:

Министр энергетики РФ Игорь Юсуфов в среду направил телефонограмму в компанию "Транснефть"и ЦДУ ТЭК. Он распорядился, чтобы нефтепроизводители получили право экспортировать нефть исключительно исходя из заявленных объемов нефтедобычи.

Даже самая простая вопросно-ответная система должна быть способна ответить на вопросы типа «Кто направил телефонограмму в ЦДУ ТЭК?» или «Куда Юсуфов направил телефонограмму?» А вот вопросы вроде «Кто распорядился, чтобы нефтепроизводители получили право экспортировать нефть?» или «В какой день недели Юсуфов отправил телефонограмму?» вполне могут оказаться для нее непосильными. На первый из них система легко ответит «он», но для определения того, что «он» в данном случае это Игорь Юсуфов, требуется проделать дополнительную работу. Корректная обработка второго вопроса требует от системы понимания того, что среда — это день недели.

На нынешнем этапе наша система занимается лишь «жонглированием» словами входного текста и задаваемого вопроса, не пытаясь создавать какую бы то ни было базу знаний. Тем не менее она способна грамотно ответить на простые вопросы к тексту, что уже немаловажно.

На данный момент система обрабатывает так называемые специальные вопросы (т.е. вопросы, задаваемые к тому или иному члену предложения) с вопросительными словами *кто*, *кого*, *кому*, *кем*, *о ком*, *что*, *чего*, *чему*, *чем*, *о чем*, а также *где*, *куда* и *откуда*.

5. Синтаксис и семантика. Получение ответов на описанные выше вопросы требуют корректного определения падежных форм слов предложения и построения дерева зависимости одних слов от других. Реально это означает уровень *синтаксического* или даже *морфологического* анализа. Тем не менее три вопросительных слова (из рассматриваемых) требуют уже базового семантического разбора предложения. Речь идет о словах *где*, *куда* и *откуда*. В принципе можно сказать, что вопросу *где* соответствует предложный падеж, вопросу *куда* — винительный, а вопросу *откуда* — родительный. Однако на вопросы *где прошла жизнь?*, *куда пошел Сергей?* и *откуда пришел Иван?* нельзя ответить сочетаниями *в трудах*, *на риск* и *из вежливости* соответственно. Чтобы выявить, к примеру, принципиальное отличие в значениях сочетаний *из вежливости* и *из ресторана*, необходим семантический анализ. В настоящее время анализатор В. А. Тузова автоматически определяет возможность применения вопросительных слов *где*, *куда* и *откуда* (а также *как* и *почему*) к данной конструкции предложения.

Другим, менее ярким примером использования семантической информации в вопросно-ответной системе является проверка (с помощью семантического словаря) на одушевленность. В частности, слово *собака* не может быть ответом на вопрос *что лежит на диване?*

6. Вопросно-ответная система: практические аспекты. Функционирование системы основывается на некоторых предположениях, которые мы рассмотрим подробнее.

Центральное слово (т.е. корень дерева) вопросительного предложения является глаголом. Здесь надо отметить, что вопросительность предложений никак не влияет на древовидную структуру получаемой формальной записи. Примеры: *куда собираются экспортировать газ?* => собираются(куда, экспортировать(газ)); *На кого окажет давление Путин?* => окажет_давление(на(кого), Путин)*. Разумеется, некоторые вопросительные предложения в русском языке имеют другую форму (*кто дома?* => дома(кто); *кому весело?* => весело(кому)), но подобные случаи мы пока не рассматриваем.

Вопросительное слово является первым аргументом центрального слова вопросительного предложения. Примеры из п. 5 наглядно подтверждают это предположение. Специальной обработки требует только случай предложного падежа. Скобочная форма типичного вопроса с конструкцией *о ком/о чем* выглядит так: центр_глагол(о(ком), арг_2, ..., арг_n).

Центральное слово предложения, содержащего ответ, совпадает с

*Для экономии места будем использовать скобочную форму для описания древовидных структур. В записи $f(a_1, \ldots, a_n)$ элементы $a_1, \ldots, a_n$ являются потомками узла f.

центральным словом вопроса. Один (реже несколько) аргумент центрального слова этого предложения является собственно ответом на вопрос. В случае «падежных» вопросов определить требуемый аргумент можно по падежу: падеж корневого элемента аргумента-ответа совпадает с падежом вопросительного слова. Если же вопрос требует семантического анализа, потребуется произвести поиск элемента, к которому этот вопрос может быть применен (как уже было отмечено, семантический анализатор предоставляет такую информацию). Пример: *Президент заявил о необходимости модернизировать транспортную инфраструктуру Дальнего Востока.* Скобочная форма этого предложения выглядит так: заявил (Президент, о(необходимости (модернизировать (инфраструктуру (транспортную, Дальнего_Востока))))). Рассмотрим вопрос: *кто заявил о необходимости модернизировать инфраструктуру Дальнего Востока?* Он будет преобразован в конструкцию заявил (Кто, о (необходимости (модернизировать (инфраструктуру (Дальнего_Востока))))). Поскольку центральное слово нашего повествовательного предложения совпадает с центральным словом вопроса, оно становится «интересным», т.е. в нем имеет смысл поискать ответ. Вопросительное слово *кто* находится в именительном падеже (эту информацию предоставляет семантический анализатор), следовательно нас интересуют аргументы центрального слова повествовательного предложения, находящиеся в именительном падеже. В данном случае таким аргументом будет лишь слово *Президент*. Это означает, что если тестируемое предложение действительно содержит ответ, то им будет «Президент».

Недостаточно лишь найти в тестируемом предложении требуемую падежную форму. Задавая вопрос вроде *кто заявил о необходимости модернизировать инфраструктуру Дальнего Востока?*, мы должны прежде всего убедиться, что тестируемое предложение действительно содержит информацию о некотором заявлении, связанном с модернизацией инфраструктуры Дальнего Востока. Самым простым и при этом вполне пригодным на нынешнем этапе работы представляется следующий алгоритм:

```
исключить из дерева вопросительного предложения первый аргумент
исключить из дерева тестируемой фразы аргумент
                                        с потенциальным ответом
проверить полученные деревья на соответствие
```

Осталось определить, что такое «соответствие». Интуитивно понятно, что фраза *Президент заявил о необходимости модернизировать транспортную инфраструктуру Дальнего Востока* «соответствует» вопросам *кто заявил о необходимости модернизировать инфраструк-*

туру Дальнего Востока? и *кто заявил о необходимости модернизировать инфраструктуру?*, но не соответствует вопросу *кто заявил о необходимости модернизировать инфраструктуру Поволжья?* Заметим прежде всего, что в тестируемой фразе обязательно должно содержаться *не меньше* информации о факте, чем в вопросе. Поэтому ответ на вопрос *кто заявил о необходимости модернизировать инфраструктуру?* может содержаться в предложении, сообщающем о том, что некоторое лицо заявило о модернизации инфраструктуры плюс еще что-то (например, инфраструктура транспортная, а относится заявление к Дальнему Востоку). Если же в вопросе ясно сказано, что речь идет о транспортной инфраструктуре, мы должны отобрать лишь те предложения, в которых фигурирует именно транспортная инфраструктура — и никакая иная.

Если не вдаваться в технические подробности, алгоритм проверки соответствия модифицированного дерева тестируемой фразы (MTT) модифицированному дереву вопросительного предложения (MQT) работает следующим образом. Сначала проверяется необходимое условие соответствия: каждое слово MQT должно быть найдено в MTT. Затем производится проверка структуры деревьев: если слово A является предком слова B в MQT, это же отношение должно наблюдаться и в MTT.

Определив соответствие вопроса и текущего повествовательного предложения, можно вернуть пользователю ответ. Поскольку ответом является некоторое поддерево предложения, его еще придется «развернуть» в корректную фразу.

Пару слов следует сказать еще об одной проблеме. Алгоритм поиска в MTT слов из MQT не может ограничиться лексикографическим сравнением слов: в некоторых случаях приходится вносить поправку на род и число. Пример: кто *пошел* в школу? — Маша *пошла* в школу. Дети *пошли* в школу. Естественно, обе фразы содержат ответ на поставленный вопрос, но глагол *пойти* в первом случае стоит в женском роде, а во втором — во множественном числе, что не соответствует мужскому роду глагола в вопросительном предложении. Реально на данный момент это затруднение решается при помощи словаря словоформ.

В завершение раздела уместно привести примеры ответов системы (см. табл. 4) на задаваемые пользователем вопросы. Входным текстом будет небольшой фрагмент рассказа А. П. Чехова «Поцелуй».

Загремел рояль; грустный вальс из залы полетел в настежь открытые окна, и все почему-то вспомнили, что за окнами теперь весна, майский вечер. Все почувствовали, что в воздухе пахнет молодой

листвой тополя, розами и сиренью.

Таблица 4. **Вопросы и ответы**

Вопрос	Ответ
Что загремело?	Рояль
Что полетело в окна?	Грустный вальс
Куда полетел вальс?	Из залы
Чем пахнет в воздухе?	Молодой листвой тополя, розами и сиренью

7. Выводы и планы на будущее. Рассматриваемые в настоящей работе типы вопросительных предложений русского языка имеют ярко выраженную структуру, корректно определяемую семантическим анализатором. Поскольку специальные вопросы задаются к тому или иному члену предложения, знание падежных форм и базовый семантический анализ позволяют определить поддерево тестируемой фразы, являющееся ответом. При этом оставшаяся часть фразы должна «соответствовать» вопросу.

В качестве возможных планов на будущее отметим обработку других типов вопросительных предложений. Другое интересное направление — использование не просто семантических описаний, но некоторого рода иерархии действий. К примеру, глаголы «побежать» и «поплестись» означают, в частности, «направиться». Поэтому конструкции с каждым из них могут быть интересны в качестве ответа на общий вопрос *Вася направился в школу?* Интересное поле для деятельности открывает исследование синонимов. Хотя теоретически предложение *Иван купил автомобиль* не имеет никакого отношения к вопросу *кто купил машину?*, на практике оно может заинтересовать пользователя именно в качестве ответа. Еще больший вызов представляет собой анализ местоимений (т.е. алгоритм связи местоимения с неким объектом), но он требует куда более серьезной исследовательской и практической работы.

Mozgovoy M. V. Simple Question-Answering System, Based on Russian Semantic Analyzer.

The work describes ideas of development of question-answering system, which illustrates possible application of V. A. Tuzov's semantic analyzer. The algorithm of actual answer phrases extraction is discussed. The paper also contains an analysis of the characteristics of Russian sentences structure (in conjunction with the problem of question-answering system creation).

Литература.

1. *Тузов В. А.* Компьютерная лингвистика: опыт построения компьютерных словарей (в печати).

2. *Grinberg D., Lafferty J., Sleator D.* A robust parsing algorithm for link grammars // Proc. of the Fourth Intern. Workshop on Parsing Technologies, Prague, September 1995.

3. *Fellbaum C.* WordNet: An Electronic Lexical Database. Cambridge, The MIT Press, 1998, 423 pp.

4. *Азарова И. В., Митрофанова О. А., Синопальникова А.А.* Компьютерный тезаурус русского языка типа WordNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Междунар. конференции «Диалог», 2003. С. 43–50.

5. *Cooper R. J., Rüger S. M.* A Simple Question Answering System // Imperial College of Science, Technology and Medicine, London, England, 2000.

6. *Scott S., Gaizauskas R.* QA-LaSIE: A Natural Language Question Answering System // Proc. of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, 2001. P. 172-182.

7. *Moldovan D., Harabagiu S., Pasca M., et al.* Lasso: A Tool for Surfing the Answer Net // TREC-8. 1999. P. 175-183.

8. *Азарова И. В.* Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL // Труды Междунар. конференции «Диалог» 2003. С. 51-55.

9. *Ермаков А. Е.* Этапы лингвистического анализа текста в программных продуктах RCO // Русский язык: исторические судьбы и современность. II Междунар. конгресс исследователей русского языка. Труды и материалы. М., МГУ, 2004.

10. *Тузов В. А.* Синтаксическая структура русского языка. //Вестн. С.-Петерб. ун-та. Сер.1: Математика, механика, астрономия. 1997. Вып.1 (№1).

11. *Корхов А. В.* Метод построения вопросно-ответной системы с использованием математической формализации русского языка // Труды XXXII научной конференции факультета ПМ-ПУ СПбГУ. СПб., 2001.