XXIX Неделя науки СПбГТУ. Материалы межвузовской научной конференции. Ч.V: С.75-76, 2001.

УДК 002.5

Р.А. Конев (5 курс, каф. САиУ), Б.И. Морозов, к.т.н., доц.

СИСТЕМА ОТВЕТА НА ВОПРОСЫ ПРИ ОБЩЕНИИ С ЭВМ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Для представления данных и решения задач в компьютерах применяются специальные языки. Язык можно определить как набор символов, используемых для представления данных (семантика), и правил, предназначенных для обработки этих символов (синтаксис) и решения задач. Пользователи вынуждены учиться формулировать задачи на понятных для компьютера языках. Точно так же компьютер выдает полученное решение на этих языках, если только оно предварительно не переведено на язык пользователя. Если правила перевода выражены в виде совокупности знаний (символов и процедур), то логично предположить, что могут быть разработаны средства, позволяющие компьютеру понимать постановку задачи на естественном языке, а затем на естественном же языке выдавать ее решение. Важность разработки методов и средств взаимодействия с компьютером на естественном языке обусловлена необходимостью их включения в вопросно-ответные информационно-поисковые системы ориентированные на неподготовленного пользователя.

Применение известных подходов для создания универсальной системы автоматической обработки естественно-языковых конструкций затруднено по целому ряду причин: во-первых, составление емких словарей, во-вторых, необходимо совершить весьма трудоемкую операцию формирования тезауруса и разработать систему правил для определения синтаксиса и семантики предложений. В силу огромных объемов словарей скорость работы такой системы будет невелика. Кроме того, ни один словарь не может претендовать на полноту, из-за морфологического многообразия русского языка.

Предлагаемый метод автоматической обработки вопросов, предполагает последовательность действий обратную традиционной схеме автоматической обработки текста. В информационной вопросно-ответной системе в качестве входной информации выступают предложения несущие определенный смысл, который можно заранее определить исходя из предметной области применения системы. Например, в информационной системе дистанционного обучения входная информация – вопросы учеников, в системе автоматизированного управления – информация о текущем состоянии объектов и т.п. Таким образом, семантическая информация входных предложений для той или иной информационной системы заранее известна. Из каждого предложения при автоматической обработке вопросов необходимо выделить смысловые слова, отражающие суть вопроса и являющиеся ключом при поиске необходимой информации для ответа. Для этого необходимо проследить закономерности построения вопросов, т.е. установить порядок следования членов предложения и возможные варианты частей речи, выступающих в роли этих членов предложения. В результате этого мы определим синтаксическую информацию предложений.

Таким образом, после детального анализа принципов конструирования вопросов можно получить семантическую и синтаксическую информацию о каждой форме вопроса. И на основании этой информации провести морфологический анализ слов вопроса, чтобы определить смысловые слова.

В качестве примера реализации предлагаемого метода был разработан модуль автоматического ответа на вопросы учеников в системе дистанционного обучения. Для этого были проанализированы всевозможные конструкции вопросов, с целью

выявления закономерностей их построения. В результате анализа было выявлено, что в вопросах учащихся словам, отражающим основной смысл фразы, предшествуют определенные союзы (наречия в роли союзов) или указательные прилагательные. Слова, отражающие суть запроса – будем считать смысловыми словами, а слова им предшествующие – ключевыми. Таким образом, для поиска смысловых слов достаточно в предложении найти ключевое слово. Множество ключевых слов для системы дистанционного обучения включает в себя такие слова и словосочетания как: *о, об, про, что такое, кто, где, как* и т.п.

Для создания системы автоматической обработки естественно-языковых вопросов необходимо составить базу ключевых слов, состав которой будет зависеть от области применения системы. В процессе работы система будет производить морфологический анализ поступающих вопросов для нахождения ключевых слов. В качестве словаря для морфологического анализа будет использоваться база ключевых слов.

Таким образом, при использовании предлагаемого метода морфологический анализ поступившего вопроса будет выполняться лишь частично, а этапы синтаксического и семантического анализов пропускаются. Такой подход к анализу вопросов позволит значительно сократить время обработки предложений, т.к. при этом не будут выполняться самые сложные и продолжительные этапы анализа вопросов. При использовании этого метода отсутствует необходимость в составлении огромных словарей и тезауруса. Кроме того, такой метод анализа вопросов позволит успешно обрабатывать вопросы, в которых встречаются термины на разных языках, специальные сокращения, специальные символы (знаки), а также математические и химические формулы.

В дальнейшем этап поиска ответа осуществляется по смысловым словам, полученным в результате автоматической обработки вопросов.

Для повышения релевантности ответов, помимо случаев непосредственного совпадения слов поискового массива со смысловыми словами, необходимо предусмотреть вариацию окончаний, суффиксов и чередование букв в корне слов. Очевидно, что отличие регистра в написании слов поискового массива и ключевых слов поиска не должно влиять на успешный исход процедуры поиска.

Помимо этих традиционных элементов процедуры установления семантического соответствия крайне важно учитывать, что в поисковом массиве словосочетание, являющееся параметром поиска, иногда может быть разорвано. То есть, при нахождении в тексте одного слова из заданного словосочетания необходимо продолжить поиск остальной части словосочетания в поисковом массиве. Это позволит повысить полноту поиска и уменьшить время работы поисковой системы. Факт возможного разрыва словосочетаний не учитывается ни в одной современной информационно-поисковой системе.