УДК 510.62

А. Ф. ОСЫКА, канд. техн. наук

ОБЗОР ИССЛЕДОВАНИЙ ПО АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ЕСТЕСТВЕННОГО ЯЗЫКА В США

Проблема автоматической обработки естественного языка (АОЕЯ) привлекает внимание ученых самых различных специальностей: философов, психологов, лингвистов, математиков, специалистов по искусственному интеллекту и др. Относительно данной проблемы существуют различные мнения, начиная от невозможности автоматизировать понимание естественного языка (ЕЯ) [1] и кончая необходимостью исследовать и моделировать на ЭВМ довольно тонкие процессы понимания ЕЯ человеком [2, 3]. Около 30 % публикаций по искусственному интеллекту, интерактивным системам, психологии понимания, лингвистике посвящены проблемам АОЕЯ.

Рассмотрим работы по автоматическому анализу и синтезу письменного языка. Большое количество таких работ в США, разнообразие их тематики и методов изучения не позволяет остановиться сколько-нибудь подробно даже на самых значительных исследованиях. Данный обзор представляет собой попытку ответить на вопрос: что делается в США в области АОЕЯ? Обзор можно условно разделить на три части: описание систем АОЕЯ и их элементов в соответствии с некоторыми признаками (назначение, подход к анализу ЕЯ и т. д.); краткая характеристика проблем (лингвистических, психологических и т. п.), решаемых в той или иной системе АОЕЯ; схематическое описание устройства и возможностей некоторых систем АОЕЯ.

Назначение системы АОЕЯ — важный признак, определяющий ее многие особенности. Наиболее широко ведутся работы по таким направлениям: создание интерфейса базы данных (БД), использующего ЕЯ [4—6]; разработка вопросно-ответных систем на основе ЕЯ и программ,понимающих ЕЯ [3, 7, 8]; машинный перевод (МП) [9—11]; моделирование процессов овладения ЕЯ [12—14]; использования ЕЯ в качестве программного языка [15—17]. Многие исследователи не ставят своей целью создание законченной системы АОЕЯ, а решают некоторые частные вопросы этой проблемы: построение алгоритмов синтаксического анализа, синтеза текстов и др. О некоторых из них будет сказано при описании соответствующих компонентов систем АОЕЯ.

Существенным признаком системы является вид анализа, играющего ведущую роль при переходе от предложения на ЕЯ к представлению его значения на внутреннем языке системы. Такой переход обычно осуществляют в два этапа. Сначала определяют синтаксическую структуру предложения, затем на основе результата синтаксического анализа строят семантическое представление предложения, т. е. эксплицитную запись его значения на внутреннем языке системы. Эта схема используется в системах АОЕЯ, возможности которых в теоретическом плане не ограничены тематикой обрабатываемых текстов и структурой допустимых предложений [18—21]. Но практическая реализация подобных систем всегда требует введения указанных ограничений. Суть семантического анализа при таком подходе: на графе синтаксической структуры предложения отыскиваются подграфы определенной конфигурации с заданным типом вершин. Эти подграфы заменяются выражениями на языке семантического представления.

Случаи отклонения от традиционной схемы более многочисленны, чем случаи, ее подтверждающие. Например, в некоторых первых вопросно-ответных системах при анализе вопросов основное внимание уделялось выявлению их синтаксической структуры. Переход к семантическому представлению выполнялся с помощью сравнительно простой процедуры, дополняющей синтаксический анализ [22, 23]. Для выполнения синтаксического анализа предложений наиболее часто применяется грамматика Расширенных Сетей Переходов (Augmented Transition Networks) и некоторые виды грамматик Н. Хомского, для которых построено большое количество алгоритмов, использующих различные стратегии анализа [24—26].

Стала популярной идея непосредственного перехода от предложения к его семантическому эквиваленту, минуя этап синтаксического анализа. Предложены различные методы построения такого семантического анализатора. Например, в программах, разработанных в Йельском университете, главной опорой для анализа сообщения при преобразовании его в сеть Концептуальных Зависимостей служит слово [7, 27, 28]. Оно дает информацию о концептах семантического уровня и предсказывает отношения между этими концептами путем предсказания появления следующих слов в предложении. Безусловно, указанный подход не исключает необходимости проверки соответствия грамматических признаков предсказанных слов. Но она имеет вспомогательный характер, так как не ставит целью описать синтаксическую структуру всего предложения.

В отдельных случаях для построения семантических анализаторов, работающих непосредственно с элементами предложения, используются некоторые подобия формальных грамматик, в которых вместо синтаксических нетерминальных символов вводятся символы семантического языка. Например, в системе

автоматизированного обучения SOPHIE, которая дает сведения о неисправностях в электронных цепях, при анализе запроса на английском языке и построении дерева его семантической структуры ведется поиск таких составляющих: НЕИСПРАВНОСТЬ, ИНСТРУМЕНТ, ТИП СОЕДИНЕНИЯ, ТЕРМИНАЛ и т. д. [29]. Вопросно-ответная система RENDEZVOUS при анализе запросов по поводу доставки грузов морским транспортом применяет правила переписывания предложений. При этом обнаруженные слова или словосочетания, выражающие некоторые концепты, заменяются символами семантического языка [30]. В интерфейсе БД PLANES анализ запроса на английском языке ведется с помощью другого вида семантической грамматики — Расширенной Сети Переходов, дуги которой помечены семантическими категориями [31].

В некоторых вопросно-ответных системах с ограниченной тематикой запросов на ЕЯ для анализа предложений используется метод шаблонов (templates). Разновидностью метода является выделение ключевых слов в предложении с минимальным контекстом слева и справа. Выделенная часть предложения приводится к некоторому стандартному виду и сравнивается со словосочетаниями, которые хранятся в словаре [32]. В вопросно-ответной системе LADDER предложение запроса сравнивается с шаблоном для целого предложения. Шаблон представляет собой фразу определенного типа, в которой сохранены только служебные слова, а на месте полнозначных помещено описание семантики слов, удовлетворяющих данному шаблону [4].

Важная характеристика систем АОЕЯ — использование ею в процессе работы (или неиспользование) знаний о внешнем мире, т. е. хранящейся в памяти ЭВМ модели объектов внешнего мира и отношений между ними. Подобная модель внешнего мира применяется во многих системах АОЕЯ для формирования выходной информации, соответствующей тексту на входе: запросу на ЕЯ, вопросу, сообщению и т. п. Например, для лингвистического интерфейса БД такой моделью служит сама БД [30, 31]. В системах, понимающих связные рассказы, в вопросно-ответных системах в качестве модели внешнего мира может служить схематическое описание некоторой стандартной ситуации, называемое фреймом [33, 34]; описание привычной последовательности событий, называемое вслед за Р. Шенком скриптом [2, 7]; формальное описание некоторой области знания (методы решения алгебраических уравнений, функционирование и неисправности электронных цепей и т. д.) [29, 35].

Вместе с тем наличие готовой модели внешнего мира не является обязательным условием функционирования системы АОЕЯ. Например, системы МП могут обходиться без нее [9], а вопросно-ответные системы могут формировать локальную

модель (необходимую информационную базу) на основе контекста поступившего сообщения [23].

Имеется и другой аспект применения моделей внешнего мира в системах АОЕЯ. Такая модель может использоваться для формирования выходной информации и в процессе лингвистического анализа входной информации на ЕЯ. Например, представленные в БД отношения между данными могут использоваться в качестве семантического компонента при лингвистическом анализе запроса [31, 36]. Знания о внешнем мире существенно используются в таких семантических анализаторах, которые «понимают» входное сообщение на ЕЯ путем соотнесения его составляющих с элементами модели стандартной ситуации — фрейма или скрипта [8, 34].

Существенный признак системы АОЕЯ — формальный аппарат, применяемый для описания значения предложения или целого сообщения. Для представления семантики используются довольно разнообразные средства: аппарат теории множеств, как обычных [6], так и размытых [37]; формулы исчисления предикатов первого порядка [21]; семантические сети [38]; системы математических уравнений, соответствующих некоторой области знаний [35]; специальные формальные языки, которые сходны с языками программирования и удобны для описания элементарных концептов ЕЯ и отношений между ними [18, 39, 40]; набор семантических конституент (переменная плюс ее значение в анализируемом тексте) в качестве значения запроса на ЕЯ в специализированной БД [4, 31].

Способ представления семантики текста на ЕЯ зависит от назначения системы АОЕЯ, тематики обрабатываемых предложений, набора операций, выполняемых с поступившей информацией, и т. д. Например, если при переводе запроса с ЕЯ на внутренний язык системы можно не учитывать смысл глаголов в предложениях, то в качестве семантического языка может быть выбран аппарат теории множеств [6]. Для передачи некоторой неопределенности многих языковых понятий иногда удобно использовать аппарат размытых множеств [37].

Часто для ответа на запрос (например, в БД) необходима не просто выдача готовой информации, хранимой в эксплицитном виде в памяти системы, а поиск и установление новых связей между объектами в БД. В этом случае для представления значения запроса, как и свойств объектов в БД, может применяться аппарат исчисления предикатов. При присоединении формул, получаемых в результате анализа запроса, к содержимому БД возникает возможность вывода новых формул, которые характеризуют свойства объектов, указанных в запросе [21].

Одним из распространенных способов формального представления значения текстов на ЕЯ являются семантические сети [38]. В числе первых семантические сети начал использо-

вать Р. Квиллиан, который применил их для описания семантики английских слов [41]. Существует довольно много различных концепций построения семантических сетей, авторы которых стремятся как можно полнее передать значение отдельного слова, предложения или целого сообщения, приблизить формальное понимание текста к его пониманию человеком. Первые семантические сети представляли значение предложения в виде дерева, вершиной которого было нерасчлененное значение глагола. Ветви этого дерева, помеченные символами глубинных падежей Ч. Филлмора, вели к словам, передающим в анализируемом предложении значения соответствующих глубинных падежей [42]. В дальнейшем работа по увеличению выразительных возможностей семантических сетей шла по пути введения новых видов отношений между концептами семантического уровня, разграничения типов вершин, введения кванторных понятий, кодирование механизма дедуктивного вывода в графе семантической сети и т. п. [43, 44].

В качестве примера рассмотрим несколько подробнее средства и возможности языка Концептуальных Зависимостей (КЗ) — особого вида семантических сетей, используемого в Йельском Университете для передачи значения текстов на ЕЯ [3, 7, 8]. Его разработчики стремились сделать представление значения на языке КЗ независимым от ЕЯ. В языке КЗ имеются аналоги предложения концептуального уровня, 6 членов предложения, 6 частей речи, введены 15 типов синтаксической связи между концептуальными частями речи. Значение глаголов приравнивается к одному из 11 элементарных действий или описывается с помощью концептуального глагола-связки «DO» плюс изменение состояния объекта. Концептуальные глаголы имеют 10 времен, а существительные — 4 падежа. Запись значения предложения на языке КЗ представляет собой особый вид графа, вершинами которого являются концепты, а дугами — отношения между концептами.

В результате анализа текста получается формальная запись его значения, которое представляет собой сумму непосредственных значений предложений. В отличие от этого человек понимает сообщение на ЕЯ гораздо глубже, используя свое знание мира, память. Для преодоления этого недостатка был предложен ряд механизмов, которые дополняют и объединяют информацию, полученную непосредственно при анализе отдельных предложений, позволяют понимать текст на сверхфразовом уровне [2, 3, 45]. Прежде всего это механизм ситуационного анализа, называемый скриптом. Скрипт применяется к записи КЗ, полученной при анализе текста. Он распознает некоторую стандартную последовательность действий (посещение ресторана, начало поездки на автомобиле и т. п.) и добавляет всю недостающую информацию из стандартного описания подобной ситуации, хранимого в памяти ЭВМ.

В тех случаях, когда описываемая последовательность событий не является стандартной, т. е. не подходит ни под один скрипт, может оказаться полезным механизм объединения событий с помощью причинной связи. Недостающее звено в такой цепи событий может быть дополнено исходя из свойств действующих лиц и объектов, участвующих в событиях. При отсутствии непосредственной причинной связи между событиями пониманию их последовательности может помочь установление целей участников данной ситуации [2, 3]. Например, испытываемое «чувство голода» предполагает наличие цели «принять пищу», «отсутствие денег» вызывает цель «добыть деньги», «усталость» — цель «отдых» и т. п. Достижение некоторой стандартной цели реализуется с помощью одного из возможных стандартных планов. План — это обычная последовательность событий, которая ведет к реализации цели. Каждый план предсказывает появление описания некоторых событий в тексте. Считается реализованным тот план (из нескольких возможных для данной цели), который нашел хотя бы частичное подтверждение в анализируемом тексте. В своей деятельности люди руководствуются, как правило, не одной целью, а несколькими. Эти цели могут взаимодействовать между собой, быть конфликтными, становиться преобладающими в зависимости от ситуации. Для учета подобной множественности целей вводится понятие темы — набора целей, которые встречаются обычно вместе вследствие свойств одного или нескольких действующих лиц рассказа [2].

Проиллюстрируем использование механизмов скриптов, целей и планов на примере анализа такого текста: «С. не видел отца и мать уже около года. При первой возможности он купил билет на самолет. Вещи были уложены заранее». Простая сумма значений предложений мало что дает для понимания текста, как его понимает человек. Эта сумма значений не позволяет ответить на такие вопросы: «Куда куплен билет на самолет?», «С кем предстоит встреча?», «Почему вещи были уложены?», «Кто купил билет на самолет?» и т. д. Более глубокому пониманию смысла этого текста (внеязыковой ситуации, передаваемой текстом) может способствовать понимание цели С.— «сильное желание увидеть родителей», вызванное долгой разлукой. Для достижения этой цели может быть использовано несколько планов: «поездка поездом», «поездка самолетом», «приглашение в гости» и т. п. План «поездка самолетом» получает подтверждение в анализируемом тексте. Этот план автоматически дополняет информацию о значениях отдельных предложений описанием адреса родителей, приготовлений к отъезду, приобретения билета в авиакассе (что может быть сделано ссылкой на специальный скрипт «посещение авиакассы») и т. д. Дополнительная информация, извлекаемая из памяти ЭВМ, и

позволяет ответить на вопросы относительно текста, подобные перечисленным выше.

Авторы языка КЗ и дополняющих его механизмов ситуационного «понимания» текстов отмечают, что предложенные ими средства не могут передать значение любого текста. Эти средства предназначены для анализа текстов, описывающих в основном деятельность человека. Этим, возможно, объясняется небольшое число значений концептуальных глаголов и различных видов синтаксических связей языка КЗ. В деятельности людей (животных) удобно выделять некоторые стандартные ситуации, устанавливать наличие целей и планов. Но эти средства неэффективны для представления значения текстов, описывающих пейзажи, окружающую среду, чувства и эмоции человека, события, не связанные с деятельностью человека, и т. д.

Одна из важнейших проблем в области АОЕЯ — формализация свойственных человеку способностей умозаключения, извлечения дополнительной информации, понимания подразумеваемого в процессе восприятия текста исходя из буквального значения предложений. Для выполнения подобных операций в процессе анализа исходного текста и формирования выходной информации, как правило, используется формальная модель соответствующей области внешнего мира. Чем больше свойств объектов и отношений между ними отражено в модели, тем больше информации дает такая модель для понимания текста. В различных системах АОЕЯ для получения дополнительной информации относительно анализируемого текста используются такие средства представления значений о внешнем мире: БД, семантические сети, фреймы, скрипты, цели, планы, темы и т. д. [2, 3, 33, 36, 43, 46].

Механизмы извлечения такой дополнительной информации различны. Они зависят от способов представления семантики анализируемого текста. Например, для систем, использующих фреймы, скрипты и связанные с ними механизмы, это просто считывание готовой информации, хранимой в памяти системы, при выполнении некоторого условия [2, 8]. Для системы АОЕЯ, в которых значение текста передается с помощью формул исчисления предикатов, получение дополнительной информации означает необходимость вывода новых тождественно истинных формул [21].

При обсуждении вопроса о получении системами АОЕЯ дополнительной информации относительно смысла текста было указано на двойную роль скриптов и фреймов. С одной стороны — это аппарат анализа и записи значения исходного текста, а с другой — средство представления знаний о внешнем мире в памяти системы. Такая двойственность объясняется тем, что схематически фрейм или скрипт можно представить как некоторый список вопросов, характеризующих некоторую ситуацию. Ответы на вопросы ищутся в тексте. В этом плане фрейм или

скрипт выступает как аппарат анализа текста и представления его значения. Вместе с тем в памяти системы имеются стандартные ответы на большинство из этих вопросов. Эти данные используются при формальном описании ситуации, когда на вопрос не находится ответа в тексте. С этой точки зрения фрейм или скрипт является средством представления знаний о внешнем мире.

Кратко охарактеризуем некоторые другие проблемы АОЕЯ, на решение которых сконцентрированы усилия исследователей. Часто препятствия для автоматизации обработки текстов возникают в связи с анафорой. Под анафорой понимается средство сокращения и связи текста с помощью введения местоимений, именных групп и имен собственных, значение которых было заранее задано или будет раскрыто в последующем тексте [47, 48]. Референтом анафорической ссылки может быть объект (один или несколько) или событие (одно или несколько). Для «понимания» значения анафорических ссылок в системах АОЕЯ используется несколько подходов: эвристические приемы (например, референт местоимения ищется в анализируемом или предыдущем предложении среди именных групп, возможность использования именной группы в качестве значения местоимения оценивается числом и т. д.) [35, 49]; синтаксические методы поиска референта местоимения по дереву синтаксической структуры предложения [50]; семантические методы, использующие, как правило, аппарат ситуационного анализа контекста — фреймы, скрипты и т. д. [2, 34, 47].

Следует отметить, что все эти разнообразные приемы не обеспечивают полное решение проблемы анафоры.

При автоматизации обработки текстов серьезные проблемы могут быть вызваны неоднозначностью анализа именных групп, в которых в роли определений выступают существительные (например, cupper engine gas pipe). Для правильного анализа таких словосочетаний используется проверка синтаксических связей между словами или семантических связей между концептами в предыдущем контексте или в постоянной памяти системы (например, в БД) [36, 51].

Эллиптические предложения являются источником особых трудностей при автоматической обработке текстов на ЕЯ. В системах АОЕЯ используется несколько методов анализа таких предложений. Суть одного из них: производится семантический анализ эллиптической фразы, поступившей в составе запроса в БД, затем отыскивается ближайшее предыдущее предложение, семантическая структура которого максимально совпадает со структурой эллиптической фразы. Недостающие компоненты для эллиптической фразы берутся из найденного таким образом предложения [29, 31].

Семантический анализ фраз ЕЯ, содержащих кванторные слова (все, некоторые, каждый, несколько и т. д.), вызывает

определенные трудности, привлекающие внимание исследователей. Изучается процесс анализа таких фраз с привлечением информантов [52]. Предложено несколько конкретных способов автоматического анализа подобных предложений, основанных, как правило, на использовании аппарата исчисления предикатов или теории множеств [6, 21, 53].

Для автоматизации общения на ЕЯ человека и ЭВМ большое значение имеет понимание общих правил построения диалогов между людьми, закономерности их протекания, изучение их различных аспектов: контроль диалога и формирование направляющих и уточняющих вопросов [54], понимание целей отдельных высказываний и целей диалогов [55] и т. д. По мере протекания диалога объекты и действия, которые находятся в центре внимания говорящих, меняются. Автоматическое определение такого центра внимания на каждом этапе диалога или повествования — важная задача, решение которой затрудняется наличием в текстах сравнений, объяснений на примерах, вводных предложений и даже некоторых «многозначительных» слов. Поясним изложенное на примере такого текста: «Сегодня я ходил в кино, в ресторан, а теперь начну заниматься, так как завтра у меня экзамен. Начало экзамена в 9 утра». Основное внимание в этом тексте уделяется занятиям и экзамену. О них, скорее всего, пойдет речь дальше. Но без учета центра внимания семантический анализатор может начать извлекать из памяти системы информацию о посещении кинотеатра со всеми имеющимися подробностями, затем — о посещении ресторана. Однако эти подробности не требуются для достаточно полного «понимания» изложенной ситуации. Это будет напрасной тратой времени и памяти ЭВМ. И то, и другое — критические факторы работы любой системы АОЕЯ. Проблема определения центра внимания в диалоге или повествовании — объект интенсивных исследований [7, 56].

В процессе общения между собой люди далеко не всегда строят идеальные в грамматическом отношении предложения, которые могут иметь морфологические, синтаксические и другие ошибки. Тем не менее такие предложения могут достаточно хорошо передавать мысль говорящего. Не возникает необходимости в уточняющих вопросах. Системы АОЕЯ, предназначенные для обработки только грамматически правильных предложений, рассчитаны на идеальные, нереальные условия работы. Возможность их практического использования невелика. Поэтому в настоящее время уделяется большое внимание разработке методов и систем АОЕЯ, которые позволяют правильно анализировать тексты, содержащие восстановимые по контексту ошибки. Например, идея одного из таких методов состоит в сравнении семантики отрезков текста с иерархическими структурами контекстов. Несовпадение некоторого элемента текста с частью иерархической структуры позволяет предположить

наличие ошибки, выбрать наиболее совпадающую структуру и, используя ее, исправить ошибку в исходном тексте [57]. В лингвистическом интерфейсе PLANES к БД о полетах и профилактике самолетов при анализе исходного запроса слова объединяются в группы, характеризующие тип вопроса, тип самолета, период времени и т. п., прежде всего по семантическому признаку. При объединении слов в группы могут игнорироваться некоторые синтаксические погрешности текста [31].

Предыдущая часть обзора была посвящена проблемам анализа текстов на ЕЯ. Синтез текстов, исходя из некоторого представления их значения, также является очень важной задачей. Практически в каждой системе АОЕЯ предусматривается возможность синтеза текста для вывода промежуточной или результирующей информации [5, 8, 49]. Следует отметить, что «системы, которые не ставят своей задачей моделировать использование языка человеком для представления информации, могут применять довольно жесткие методы синтеза текста или даже «законсервированные» ответы» [28]. Примером системы с изощренным анализатором текстов ЕЯ и слабыми возможностями для их синтеза является лингвистический интерфейс LUNAR [18].

Вместе с тем блоки синтеза многих систем АОЕЯ позволяют формировать предложения довольно сложной синтаксической структуры, разнообразные по значению [28, 58, 59]. Например, программа перехода от значения, представленного на языке КЗ, позволяет синтезировать тексты независимо от их тематики, а также операций, выполняемых с исходным текстом: перефраза, реферирование, нахождение дополнительных данных и т. п. [28]. Программа синтеза текстов, разработанная в Массачусетском технологическом институте, формирует английские предложения в три этапа. На первом этапе семантическое представление сообщения разбивается на группы и для каждой группы генерируется набор ядерных предложений. На втором — исходя из тематических и синтаксических предпосылок выбирается набор трансформаций, которые следует выполнить с каждым набором предложений. На третьем этапе выполняются предписанные трансформации, объединяются измененные ядерные предложения каждого набора в отдельное предложение, вводятся местоимения, получается окончательная цепочка английских слов [57]. Проводятся исследования по улучшению стиля синтезированных текстов [60], а также некоторых других аспектов автоматического формирования текстов на ЕЯ.

Кратко опишем некоторые известные системы АОЕЯ: их назначение, выполняемые операции с текстом, используемые средства обработки текстов и т. п.

Автоматическое понимание связных текстов на ЕЯ является привлекательной задачей, имеющей практическое применение. В качестве примера рассмотрим две программы, разработан-

ные в Йельском университете. Программа SAM предназначена для реферирования газетных сообщений о некоторых типах сообщений: землетрясениях, транспортных происшествиях, наводнениях, встречах глав правительств и т. п. Обработка исходного текста производится поэтапно. Вначале текст переводится на язык КЗ. Затем запись на языке КЗ анализируется с помощью скрипта, схематически описывающего соответствующую ситуацию. Например, скрипт «транспортное происшествие» содержит такие переменные: вид транспорта, место происшествия, сколько человек убито, кто виновен. Каждый скрипт содержит от 40 до 100 и более шаблонов (вопросов), которые применяются к представлению на языке КЗ для определения значения соответствующих переменных в скрипте. Когда значения этих переменных определены и тем самым получена сокращенная запись содержания сообщения, выполняется синтез текста на английском, русском или испанском языках. Полученный текст демонстрирует «понимание» программой исходного текста. Данная программа может также отвечать на вопросы по тексту. В ней имеются средства для установления подразумеваемых связей между событиями описываемого эпизода, а также для получения стандартной для подобных эпизодов информации [2, 8].

Анализ исходного текста программой SAM основан на использовании механизма скрипта, т. е. схематического описания некоторой ситуации. Часто описываемый эпизод не относится к разряду стандартных, поэтому трудно предвидеть, о чем дальше может быть сказано в тексте. В этих случаях для понимания рассказа, логики его событий может быть использован механизм понимания целей и планов действующих лиц рассказа, как это делается в программе PAM [2, 8]. Эта программа предназначена для «понимания» рассказов, состоящих из двух — девяти предложений. Рассказы классифицируются на 16 типов в зависимости от наличия одной или нескольких целей у действующих лиц и взаимного влияния целей на происходящие события. В начале обработки текста PAM переводит его на язык КЗ, затем применяется механизм определения целей и соответствующих планов. В результате применения этих механизмов получается новое распространение, дополненное представление значения текста. Затем выполняется ответ на вопрос относительно содержания рассказа либо синтез перефразированного текста. PAM может отвечать на вопросы о целях поступков, причинах событий и т. д., а также пересказывать текст от лица различных персонажей. PAM — программа экспериментальная. Ее авторы не стремились к максимальной компактности и быстродействию программы.

Построение эффективного лингвистического интерфейса БД является одной из важнейших задач в области АОЕЯ. Система ROBOT может рассматриваться в качестве примера такого

интерфейса [5, 61]. Этот интерфейс имеет возможности для работы с большим словарем терминов и других английских слов. При построении словаря используется инвертированный список названий элементов данных в БД. Такой список может строиться автоматически. Это облегчает применение данного интерфейса для других БД, а также своевременное пополнение словаря новыми терминами. В словарь включены также синонимы и обобщающие понятия по отношению к названиям элементов данных в БД, что позволяет получать ответы на запросы, описывающие данные не так, как это сделали разработчики БД. Для анализа запроса и снятия его возможной неоднозначности используется прежде всего словарь, в котором словам сопоставлена морфологическая, синтаксическая и семантическая информация. Затем выполняется синтаксический анализ запроса. В БД вводятся все варианты синтаксического анализа данного запроса. Те варианты анализа, для которых не нашлось соответствующих структур информации в БД, считаются ошибочными. Если в БД имеются ответы на несколько вариантов запроса, то эти варианты и соответствующие им ответы представляются заказчику для выбора. Опубликованы сообщения об использовании данного лингвистического интерфейса в 12 коммерческих БД [24].

Машинный перевод (МП) был первым видом АОЕЯ, который привлек внимание исследователей. Работы по МП ведутся с начала 50-х годов. Его основные проблемы: многозначность слов и предложений, использование контекста при переводе, эллипсис, метафора и т. д. — еще далеки от своего окончательного решения. В США было создано несколько действующих систем технического МП с различных языков на английский и наоборот. Вследствие невысокого качества переводов и сравнительно невысокого быстродействия эти системы не пользуются спросом у заказчиков. Но они являются потенциальным товаром на рынке программного обеспечения ЭВМ и иногда находят сбыт. Например, в еженедельнике деловых кругов «Computer World» помещено сообщение о том, что для ускорения перевода документации на продукцию иностранных заказчиков фирма «Solar Turbines Inc.» в Сан Диего купила пакет программ МП — Weidner Communications, Inc's Translation System. Применение системы МП ускорило оформление документации на 70 %. Переводчики этой фирмы просматривают на экране дисплея переведенный текст, внося, где необходимо, стилистические и смысловые коррективы. Пакет программ работает на мини-ЭВМ [62].

К наиболее крупным центрам МП, располагающим действующими системами, относятся Джорджтаунский университет в Вашингтоне и национальная Лаборатория Комиссии по Атомной Энергии в Оук Ридже, штат Теннеси [9]. Система МП Джорджтаунского университета используется в практически

неизмененном виде с 1964 г. Важная часть системы — словари русских и английских слов. Например, для русского слова, представленного в виде основы, в словаре помещена такая информация: сведения о морфологии, различные значения слова и условия реализации этих значений, данные о переводном эквиваленте и т. п. Процесс анализа русского текста проходит в два этапа. На первом этапе выполняется сегментация предложений, обработка идиоматических выражений, исключение из анализа некоторых слов (формулы, специальные знаки, списки и т. д), учет влияния исключенных слов на соседние, обработка неопознанных слов. На втором этапе производятся синтагматический и синтаксический анализы.

Процесс синтеза также состоит из двух стадий: 1) выбор и морфологическое оформление английских слов, соответствующих единицам русского текста; 2) взаимное расположение английских слов в предложении, введение артиклей и т. д. Система МП Джорджтаунского университета относится к типу эмпирических систем. Она работает автоматически, без участия пред-, интер- и постредакторов [9].

Указанные системы МП в Вашингтоне и Оук Ридже работают на ЭВМ устаревших типов и используются не на полную мощность в связи с нехваткой заказов. Стоимость МП по сравнению с ручным невелика. Основные затраты связаны с перфорацией исходного текста для ввода в ЭВМ. Некоторое время назад произведено обследование заказчиков, пользующихся системами МП. При обследовании ставилась цель выяснить: кто нуждается в МП, какого перевода ждет заказчик, чем неудовлетворены они в существующем положении дел и т. п. Эти данные учтены при дальнейшем совершенствовании систем МП [9].

В рамках одной статьи невозможно дать сколько-нибудь полный перечень систем АОЕЯ и охарактеризовать их. Отметим лишь, что большинство из них имеет экспериментальный характер. Но разработчики стараются, как правило, придать системе товарный вид, чтобы найти заказчиков и продолжить финансирование работ по совершенствованию данной системы. Программы АОЕЯ для практических целей используются сравнительно редко. Но несомненная потенциальная практическая ценность подобных исследований и масштабы работ в этой области свидетельствуют о том, что с разработкой систем АОЕЯ в значительной мере связывается будущее вычислительной техники во многих областях человеческой деятельности. Отношение большинства американских ученых к данной проблеме можно сформулировать следующим образом: чтобы практическое использование систем АОЕЯ стало реальностью в будущем, над этим нужно работать сейчас.

Список литературы: 1. *Odell J.* Are Natural Language Interfases Possible? IBM Systems Research Institute.— Technical Report TR 73—24. August 1981.—40 p. 2. *Schank R. C., Abelson R. P.,* Scripts, Plans, Goals and Understanding.— Lawrence Erlbaum Associated Publishers, Hillsdale, 1977.—246 p. 3. *Wilensky R.* Understanding Goal — Based Stories. Garland Publishers, New York, 1980.—317 p. 4. *Hendrix G. G.* Developing a Natural Language Interface to Complex Data.— ACM Transactions on Data Base Systems, 1978, 3, p. 105—147. 5. *Harris L. R.* Use Oriental Data Base Query with ROBOT Natural Language Query System.— International Journal of Man-Machine Studies. 1977, No 9, p. 697—713. 6. *Kradeloh K. D., Lockeman A. C.* Access to Data Base Systems via Natural Language.— Natural Language Communication with Computers. Springer Verlag, Berlin, 1978, p. 49—86. 7. *Lehnert W. G.* The Process of Question Answering.— Lawrence Erlbaum Associated Publishers, Hillsdale. 1978.—278 p. 8. *Schank R. C., Riesbeck Ch.* Inside Computer Understanding. Five Programs Plus Miniatures.— Lawrence Erlbaum Associated Publishers. Hillsdale. 1981.— 386 p. 9. *Henisz B., MacDonald R., Zarechnak M.* Machine Translation.— Mouton Publishers, The Hague, 1979.—265 p. 10. *Translating* and the Computer. Proceedings of a Seminar. Ed. by B. M. Shell, North-Holland Publishing Company, Amsterdam, 1979.—189 p. 11. *Brown G. P.* Some Problems in German to English Mashine Translation.— Cambridge, MIT, Project MAC, 1974.—189 p. 12. *Selfridge M.* A Process Model of Language Acquisition. Ph. D. dissertation, Yale University, 1980.—280 p. 13. *Jones M. A.* Natural Language Acquisition by Procedurally Extended Grammars. Ph. D. dissertation, University of Kansas, 1980.—120 p. 14. *Orgen P. J.* The Induction of the Syntax of Natural Language by Computer. Ph. D. dissertation. The University of Wiskonsin-Madison, 1979.—120 p. 15. *Winograd T.* Beyond Programming Languages.— Communications of the ACM, 1979, 22, p. 391—401. 16. *Ballard B., Biermann A.* Programming in Natural Language.— Natural Language Conference Proceedings. 1979, p. 288—297. 17. *Reeker Z. H.* Natural Language Programming and Natural Programming Language.— Australian Computer Journal, 1980, 12, August, 1980, p. 89—92 18. *Woods W. A.* The Luna Science Language Information System: Final Report. Report 2378, Bolt Beranek and Newnam Co., Cambridge, Mass., 1972.—198 p. 19. *Woods W. A.* Semantics for a Question — Answering Systems, Garland Publishers, New York 1979.—346 p. 20. *Sager N.* The String Parser for Scientific Literature.— Natural Language Processing. Ed. by R. Ruston. Algorithm Press, New York, 1973, p. 82—99. 21. *Berry-Rogghe G. L., Wuls H.* An Overview of PLIDIS. A Problem Solving Information System with German as Query Language.— Natural Language Communication with Computers. Ed. by G. Goos and J. Hartmanis. Springer Verlag, Berlin. 1978, p. 87—132. 22. *Green B. F.* BASEBALL: An Automatic Question Answer.— Computer and Thought. Ed. by E. A. Fiegenbaum, J. Feldman, McGrau Hill, New York, 1963, p. 35—61. 23. *Lindsay R. K.* Inferencial Memory as the Basis of Machines which Understand Natural Language.— Computers and Thought Ed. by E. A. Feigenbaum, J. Feldman. McGraw Hill, New York, 1963, p. 85—107. 24. *Bates M.* The Theory and Practice of Augmented Transition Network Grammars.— Natural Language Communications with Computers. Ed. by G. Goos and J. Hartmanis. Springer — Verlag, Berlin, 1978, p. 191—260. 25. *Markus P. M.* A Theory of Syntactic Recognition for Natural Languages. Ph. D. Thesis, MIT, 1978.—198 p. 26. *Damerau F.* Advantages of a Transformational Grammar for a Question Answering. — Proceedings of the 5-th IJCAI-77, MIT, Cambridge, Mass, 1977.—192 p. 27. *Schank R. C.* An Intergrated Understander.— Computational Linguistics, 1980, 6, No 1, p. 13—30. 28. *Schank R. C.* Conceptual Information Processing. Noth-Holland Publishing Company, Amsterdam, 1975.— 374 p. 29. *Brown J. S., Burton R. R.* A Paradigmatic Example of an Artificially Intelegent Instructional System.— International Journal of Man-Mashine Studies. 1978, No 10, p. 323—339. 30. *Codd E. F.* RENDEZVOUS Version 1: An Experimental English—Language Query Formation System for Casual User. Report RJ 2144 (29407). San Jose, Calif.: IBM Research Laboratory, 1978.— 173 p. 31. *Walts D. L.* The PLANES System: Natural Language Acces to a Large Data Base. Technical Report T-34, Coordinated Science Laboratory, University of Illinois, Urbana, 1976.—189 p.