Человеку же в будущей системе УВД должны быть отведены функции постановки задач для автоматических подсистем УВД, контроль их исполнения, заботы о совершенствовании системы УВД в целом и разумеется, выполнение тех сложных функций управления, которые на текущий момент времени не включены либо по техническим причинам либо по принципиальным соображениям в автоматические блоки УВД. Кроме того, за пользователем такой системы могут быть оставлены посильные для человека функции эксперта, сводящиеся в основном к пополнению базы знаний информацией, которую не удалось еще формализовать, в том числе основанной на индивидуальном опыте и индивидуальных предпочтениях (скажем, при заходе солнца в такой-то местности удобнее подлетать к аэропорту по определенному направлению; или следует учесть, что в некотором районе потоки воздуха неустойчивы и т.д.). Учет такой информации может помочь выработать не только формально эффективное, но и наиболее понятное и приемлемое интуитивно человеком решение.

СПИСОК ЛИТЕРАТУРЫ

1. Величенко В.В. Технический интеллект. (В настоящем журнале.)
2. Технологии работы диспетчеров службы движения гражданской авиации. — М.: Воздушный транспорт, 1982. — 284 с.
3. Справочник пилота и штурмана гражданской авиации / Русол В.А., Киселев В.Ф., Крылов Г.О. и др. — М.: Транспорт, 1988. — 319 с.
4. Wilber G.F. Strategic route planning using informed best-first search // Proc. IEEE nat. aerosp. and electron. conf., Dayuton, May 23–27, 1988. — V. 3. — New York, 1988. — P. 1137–1141.
5. Искусственный интеллект: Справочник. Кн. 2: Модели и методы. — М.: Радио и связь, 1990. — 303 с.
6. Хейли А. Аэропорт: Роман. — М.: 1990. — 457 с.

О проблемах автоматической обработки текстов на естественных языках

О.С. Кулагина

В данной работе рассматривается аспект меры в лингвистическом знании и его использование в системах автоматической обработки текстов (АОТ) на естественных языках (ЕЯ).

Традиционные лингвистические описания различных ЕЯ имеют качественный характер. Однако в компьютерной, вычислительной лингвистике при решении задач АОТ важно знать и учитывать не только качественные, но и количественные характеристики лингвистических феноменов. В дальнейшем изложении среди систем АОТ нас будут интересовать в основном системы автоматического анализа текстов, причем такие, в которых ставится задача по-возможности полного учета и отображения морфологических, синтаксических и семантических характеристик анализируемого текста. Такого рода постановка задачи возникает, например, естественным образом в системах машинного перевода текстов с одних ЕЯ на другие, поскольку перевод должен и правильно передать смысл входного текста, и иметь синтаксическое подобие, без чего он превращается в перевод. В отличие от перевода, в других классах задач АОТ достаточно анализа, дающего лишь частичное, вырожденное преобразование текста, как, например, поисковый образ в виде списка дескрипторов в системах информационного поиска и т.п. Проблемы частичного анализа, также как и проблемы синтеза текстов здесь затрагиваться не будут, хотя и для них количественный аспект описания ЕЯ имеет существенное значение.

Машинные системы АОТ, имеющие целью невырожденные преобразования текстов, часто называемые лингвистическими процессорами (ЛП), моделируют в определенном ограниченном смысле человеческое владение ЕЯ. Для человека владение ЕЯ включает возможность понимания и создания текстов на этом языке. Анализ текста системой (переход от текста к его внутреннему представлению) подобен пониманию, а синтез (переход от внутреннего представления к тексту) — изложению мысли. Говоря о подобии, мы совершенно не имеем в виду воспроизведение каких бы то ни было механизмов мышления, а только сходство результатов функционирования.

Трудность построения эффективных ЛП определяется самой природой ЕЯ, характерной для них сложностью, многоуровневостью, неоднороднос-

тью, нечеткостью, недетерминированностью переходов. Сложность ЕЯ проявляется в наличии очень большого числа элементов, разнообразии их свойств и отношений, способности образовывать различные сочетания с новыми свойствами и со сложными иерархическими соотношениями. Неоднородность проявляется, например, в том, что при любых классификациях элементов ЕЯ наблюдается очень большой разброс по величине классов. В качестве простого примера приведем склонение русских существительных. Если в качестве класса взять существительные, имеющие один и тот же набор окончаний в 12-местной парадигме, то среди таких классов окажутся как такие, которые содержат несколько тысяч слов, так и классы, состоящие из одного слова. Очень сильно различаются также частоты употребления различных элементов ЕЯ в текстах. Нечеткостью характеризуются границы различных множеств: и области значений слов и словосочетаний, и область синтаксической правильности выражений на ЕЯ и др. Недетерминированность переходов, например, от смысла к тексту и от текста к смыслу, проявляется как на уровне отдельных слов (омонимия, синонимия), так и, в еще большей степени, на уровне предложений и текстов.

Перечисленные свойства ЕЯ обеспечивают широту их возможностей, гибкость, но делают чрезвычайно трудной задачу автоматизации работы с текстами.

ЛП естественно базируются на тех описаниях ЕЯ, которые существуют в виде словарей и грамматик. С самого начала работ по автоматизации обработки текстов было видно, что грамматики, написанные в свое время для человека, неудовлетворительны для создания ЛП. При этом на первых порах казалось, что их основной недостаток состоит в недостаточной формализации, в апелляции к пониманию при формулировке правил. Со временем прояснилось, что созданию эффективных ЛП мешает сугубо качественный характер описания ЕЯ, традиционный для лингвистики, в то время как знание человека о ЕЯ включает и аспект, который естественно назвать аспектом меры, или количественным. Качественный аспект знаний о ЕЯ обеспечивает возможность получить в той или иной ситуации некоторый набор альтернатив, тогда как аспект, названный количественным, обеспечивает упорядочение этих альтернатив по определенным предпочтениям. Иными словами, носитель языка знает не только, как можно понять или выразить нечто на этом языке, но также и то, какое из этих пониманий или выражений является наиболее естественным, регулярным, наилучшим.

Как было сказано выше, в данном изложении мы будем в основном ориентироваться на ЛП, анализирующие текст. В настоящее время существует большое число ЛП-анализаторов, реализующих морфологический и синтаксический анализ для разных языков и обеспечивающих достаточно полный охват соответствующих ЕЯ. В отличие от них, семантический анализ в ЛП обычно делается с ориентацией на определенную ограниченную предметную область.

При описании синтаксиса ЕЯ в целях АОТ и неоднородность, и недетерминированность проявляются в еще большей степени. Неоднородность

проявляется, например, в разбросе по величине классов слов с одинаковыми валентностями на наличие определенных подчиненных или определенных управляющих.

Большой разброс можно наблюдать и с точки зрения частоты употребления тех или иных синтаксических конструкций, где под синтаксической конструкцией понимается сочетание представителей определенных синтаксических классов, связанных определенными синтаксическими отношениями. Здесь, обычно, имеет место следующая ситуация. ЕЯ предоставляет некоторый набор альтернатив, который перечисляется в обычной, т.е. созданной для человека, а не для машины, грамматике. Все перечисленные возможности являются допустимыми, правильными, но с точки зрения естественности, регулярности, они неравноправны. Носитель языка обладает определенным знанием о предпочтительности одних перед другими, но это знание в обычных грамматиках не отражается.

В качестве известного примера можно привести способ выражения подлежащего в русском языке. В грамматиках можно прочесть, что подлежащее выражается именем (существительным, местоимением и т.п.) в именительном падеже, глаголом в неопределенной форме или целым простым предложением. Носителю языка очевидно преобладание первого способа, но это знание приобретено им не из грамматических руководств.

Другой пример касается свободы порядка слов в русском языке. Предположим, что требуется проанализировать следующее входное предложение: "Возрастание S вызвало увеличение Q". Формально оно может быть проанализировано двумя способами: оба существительных ("возрастание" и "увеличение") могут быть как подлежащим, так и дополнением при глаголе "вызвало", поскольку русский язык допускает постпозицию подлежащего. Так, например, если бы анализируемое предложение звучало "Прирост S вызвало увеличение Q" слово "увеличение" определялось бы как подлежащее однозначно благодаря согласованию с глаголом по роду. Однако два возможных анализа первого предложения неравноправны, несмотря на формальную правильность обоих. Тот, при котором, предшествующее глаголу, т.е. слово "возрастание" будет признано подлежащим, а "увеличение" — дополнением является более естественным. Подчеркнем, что выбор между альтернативными анализами делается именно в терминах предпочтения, а не деления на правильно/неправильно, т.е. не на качественном уровне.

Еще один пример выбора по предпочтению относится к выявлению наиболее предпочтительного из набора возможных управляющих для некоторого слова при установлении синтаксических отношений между словами. При построении синтаксической структуры анализируемого предложения в виде дерева зависимостей ставится задача установить между словами анализируемого предложения синтаксические отношения, или, в иных терминах, синтаксические связи, которые являются бинарными и несимметричными: из двух слов, связанными некоторым отношением одно является главным, или управляющим, а другое — подчиненным, или управляемым. Возможно-

сти слова вступать в синтаксические отношения с другими словами определяются его валентностями, которые также являются направленными: одни определяют способность слова выступать в качестве подчиненного, другие — в качестве управляющего. Последние образуют так называемую модель управления, которая характеризуется числом мест, способами заполнения этих мест, степенью необходимости их заполнения, сочетаемостью различных способов заполнения и другими сведениями. Пусть, например, анализируется словосочетание: "Такое решение Иванова". Тут естественно считать, что слово "Иванова" заполняет место субъекта в слове "решение". Однако, если данное словосочетание продолжено: "Такое решение Иванова устраивает", то в этом случае мы подчиним слова "Иванова" глаголу "устраивает", а соответствующая валентность слова "решение" останется незаполненной. Здесь опять нельзя сделать выбор на уровне деления на правильно/неправильно, а приходится из двух альтернатив выбирать ту, которая предпочтительнее. Иными словами, мы отдаем предпочтение тому из двух возможных управляющих, валентность которого сильнее, т.е. тому, которое предъявляет более сильное требование на заполнение соответствующего места.

Подавляющее число ЛП, осуществляющих синтаксический анализ предложений, базируется в той или иной форме на идее синтаксической правильности. Такой подход вполне удовлетворителен для формальных языков, например, в трансляторах для языков программирования. Однако в применении к ЕЯ опора только на деление "правильно/неправильно" приводит к следующим нежелательным последствиям. Если в грамматику ЛП включаются "на равных правах" все допустимые альтернативы, то растет число получаемых этим ЛП вариантов анализа для одного предложения, и, особенно, возрастают переборы на промежуточных стадиях работы. Если же некоторые альтернативы, допустимые в языке, но редкие, исключаются из грамматики ЛП, то он лишается возможности анализировать те входные предложения, в которых они употреблены. В силу сказанного выше представляется естественным искать выход на пути дифференцированного подхода к разным альтернативным возможностям, при котором учитывалась бы степень допустимости, степень естественности того или иного выбора и тем самым, более адекватным образом учитывалась бы природа ЕЯ.

Такой подход использован в системе анализа русских текстов (система АРТ), разрабатываемой в Институте прикладной математики им. М.В. Келдыша РАН и описанной в работах [1–4]. Система АРТ приводит морфологический и синтаксический анализ входных предложений, причем она рассчитана на предложения из области так называемой деловой прозы, т.е. на тексты того типа, каким пишутся научно-технические статьи.

Начальные, т.е. досинтаксические, этапы обработки предложения проходят по обычной для ЛП схеме. Обращение к словарю системы позволяет снабдить каждую входную словоформу наборами признаков (по числу лексем), на которых базируется вся остальная обработка. Морфологический анализ выдает для каждой входной словоформы все возможные морфологические представления. Например, для словоформы "стекло" будет получено, что либо это существительное в именительном падеже или винительном падеже единственного числа, либо форма среднего рода единственного числа прошедшего времени от глагола "стекать", причем, как сказано выше, словарь системы снабдит каждый из вариантов соответствующими синтаксическими признаками (синтаксический и семантический классы, подклассы, валентности и др.). С помощью словаря оборотов в предложении выявляются обороты: словосочетания, выступающие в синтаксическом анализе как единое целое, например, сложные предлоги и союзы, вводные, идиоматические обороты и т.п. Приводится также некоторое предварительное снятие омонимии для тех случаев, когда это легко сделать по линейному контексту. Например, если словоформе "стекло" непосредственно предшествует неомонимичный предлог, омоним, у которого синтаксический класс "глагола", можно отбросить.

Метод синтаксического анализа, использованный в системе АРТ, является обобщением фильтрового подхода, причем обобщением в первую очередь за счет последовательного учета предпочтений, о которых говорилось выше. Целью синтаксического анализа (САн) в системе АРТ является построение синтаксического представления входного предложения в виде двух структур: дерева зависимостей и дерева фрагментов. Дерево зависимостей представляет собой ориентированное корневое дерево в смысле теории графов, в котором узлами являются текстовые единицы (ТЕ) входного предложения, а дугами, или ветвями, — синтаксические связи (ССв) между ТЕ, характеризуемые определенными типами. ТЕ — это отдельные словоформы, знаки препинания, а также обороты, о которых сказано выше. ССв, устанавливаемые в системе АРТ, правильнее назвать не синтаксическими, а семантико-синтаксическими, поскольку их деление на типы более тонкое, чем чисто синтаксическое (различается несколько сот типов ССв). В дереве фрагментов узлами являются более сложные единицы: подцепочки входного предложения, являющиеся простыми предложениями, причастными и деепричастными оборотами и др. Между ними также устанавливаются связи определенных типов.

При этом в отличие от ЛП, использующих чисто фильтровый подход, в системе АРТ ставится задача нахождения не просто какого-то одного такого представления и не всех возможных правильных синтаксических представлений указанного вида (которых для некоторых предложений бывает несколько сот и даже тысяч), а одного представления, в некотором смысле наилучшего.

Аппаратом учета предпочтений является система оценок, которыми характеризуются как валентности слов, так и гипотетические ССв, возникающие в процессе анализа.

Синтаксический анализ в системе АРТ происходит по следующей схеме. Начальная стадия построения дерева зависимостей — это построение исходного набора гипотетических ССв между ТЕ анализируемого предложения. По способу их построения ССв распадаются на три класса. Одни из

них строятся в соответствии с моделями управления, упомянутыми выше, т.е. на основе того, какие валентности на наличие определенных подчиненных имеются у ТЕ с моделями управления. Эти гипотетические ССв получают при этом некоторые исходные оценки, учитывающие, в первую очередь, степень необходимости заполнения того или иного места в модели управления. Другие связи строятся по таблице синтагм, в которой учтены валентности слов на наличие у них определенных управляющих. Эти ССв получают предварительные оценки, учитывающие, в первую очередь, степень близости ТЕ в предложении. Например, для прилагательного будут построены определительные ССв со всеми теми существительными, с которыми у него есть согласование, причем оценка будет тем больше, чем ближе определение и определяемое. Наконец, связи третьего вида устанавливаются для сложных или кратных союзов. В результате начальной стадии САн получается некоторый исходный, вообще говоря, избыточный, набор гипотетических ССв, снабженных исходными оценками.

Наиболее существенная стадия САн — это выбор из исходного набора того поднабора, который даст искомое дерево зависимостей. Если на подготовительной стадии учитывались свойства связываемых ТЕ и , иногда, линейный контекст, то на этой основной стадии учитывается структурный контекст, сочетаемость, совместимость ССв. Этот учет выражается в пересчете оценок, при котором оценки могут как увеличиваться, так и уменьшаться. Уменьшение оценки некоторой гипотезы до нуля означает ее отбрасывание. Пересчет оценок учитывает количество претендентов на ту или иную роль, их расположение, в том числе проективность, сочетаемость различных подчиненных при общем управляющем и т.п. Например, если имеется только один претендент на заполнение некоторого места в модели управления какой-то ТЕ, причем места с высоким требованием на заполнение, то соответствующая ССв получает значительное повышение оценки. Поскольку ставится цель построить связное дерево, для каждого узла (т.е. каждой ТЕ) должна присутствовать входящая в него ССв. Предположим, что в исходном наборе появляется ТЕ, для которой имеется только один претендент на роль управляющего, т.е. для узла имеется только одна входящая в него ССв (ССв считаются направленными от управляющей ТЕ к подчиненной, а такая ССв, которая является единственной, входящей в соответствующий узел, называется уникальной). Ясно, что уникальные ССв должны войти в окончательное дерево зависимостей, поэтому оценки тех ССв, которые несовместимы или плохо совместимы с уникальными, соответственно понижаются. Уникальными ССв могут быть не только в исходном наборе, но становиться таковыми при отбрасывании каких-то гипотез.

Пересчет оценок может сопровождаться и снятием омонимии. Например, если ССв уникальна для омонимичного узла, можно отбросить все его омонимы, кроме того, который участвует в уникальной ССв.

В результате пересчета оценок получается некоторый, обычно сокращенный по сравнению с исходным, набор ССв с новыми оценками. Можно

сказать, что исходные оценки ССв отражают априорные закономерности ЕЯ, а оценки, получившиеся в результате пересчета, отражают конкретную ситуацию, имеющую место в анализируемом предложении.

Следующий шаг состоит в выборе для каждого узла тех ССв, которые получили максимальные оценки, и формирования из них согласованного (с точки зрения участия в ССв омонимов и совместимости ССв) набора, образующего дерево. Эксперименты, проведенные на многочисленных предложениях, показали, что обычно для большинства ТЕ устанавливается одна ССв с максимальной оценкой и только для нескольких ТЕ (трех-четырех при предложениях длиной около 20 слов) остается несколько ССв с максимальной оценкой. Для таких ТЕ окончательный выбор делается из соображений "естественности", т.е. учета типов ССв, расположения альтернативных управляющих относительно данной ТЕ, а также их удаленности от нее.

Построение дерева фрагментов также происходит в несколько этапов. На начальном этапе устанавливаются границы фрагментов, в качестве которых выступают союзы и знаки препинания. Эти границы, являющиеся границами первого уровня, характеризуются типами и получают оценки. Например, очевидно, что неомонимичный подчинительный союз, подкрепленный знаком препинания, это более сильная граница, чем запятая. Подцепочки ТЕ, оказавшиеся между соседними границами первого уровня, являются фрагментами первого уровня. Фрагменты также характеризуются определенными типами, в зависимости от того, какая ТЕ являются во фрагменте наиболее "весомой". Такая ТЕ называется главой фрагмента. Так, наибольший вес имеет личный глагол, несколько меньший — предикативное наречие или краткое прилагательное и т.д. Фрагмент, в котором нет ТЕ более весомой, чем деепричастие, это деепричастный оборот и т.п. Фрагменты первого уровня по определенным правилам объединяются во фрагменты второго уровня. Те, в свою очередь, могут объединяться во фрагменты третьего уровня. Правила объединения фрагментов сформулированы в терминах типов фрагментов и границ.

При пересчете оценок ССв учитывается деление на фрагменты. ССв, связывающие две ТЕ, находящиеся в одном фрагменте, получают подкрепление. С другой стороны, оценки ССв, входящих, например, во фрагмент, содержащий личный глагол, от ТЕ других фрагментов, понижаются, причем величина понижения зависит от того, какую границу переходит данная ССв.

Связи между фрагментами частично индуцируются связями между ТЕ, отобранными после пересчета оценок, когда имеются две ТЕ разных фрагментов, связанные ССв. В других случаях связи между фрагментами устанавливаются по определенным правилам, которые учитывают типы фрагментов и их границ, расположение фрагментов. Учитывается также заполненность мест предикатов и, если некоторое место не заполнено никакой ТЕ, то в качестве подчиненного может быть взят целый фрагмент. Например, в предложениях "Иванов знает биологию" и "Иванов знает, что биология ин-

тересна" одно и то же место предиката "знает" заполнено, в одном случае одной ТЕ ("биология"), а в другом — целым фрагментом ("что биология интересна").

При установлении связей между фрагментами проверяется согласованность обеих структур. Недопустимо, чтобы один фрагмент оказался подчинен двум разным. Нормально, чтобы ССв тех ТЕ, которые находятся в одном фрагменте, образовывали поддерево дерева зависимостей и т.п.

Возвращаясь к аспекту меры, о котором шла речь выше, хочется предостеречь от чересчур прямолинейного использования количественных характеристик в задачах АОТ. Например, соблазнительно простым является выбор из альтернативных ССв всегда самой короткой. Однако при таком упрощенном подходе, не учитывающем всей совокупности факторов, трудно ждать удовлетворительного результата. Успеха при построении ЛП можно ожидать только при широком и серьезном учете как качественных, так и количественных характеристик ЕЯ.

Некоторые другие проявления аспекта меры лингвистического знания, например, на уровне семантики, описаны в работе [5].

СПИСОК ЛИТЕРАТУРЫ

1. Кулагина О.С. Морфологический анализ русских глаголов / Препринт ИМП им. М.В. Келдыша АН СССР, № 195, М., 1985.
2. Кулагина О.С. Морфологический анализ русских именных словоформ / Препринт ИМП им. М.В. Келдыша АН СССР, № 10, М., 1986.
3. Кулагина О.С. Об автоматическом синтаксическом анализе русских текстов / Препринт ИМП им. М.В. Келдыша АН СССР, № 205, М., 1987.
4. Кулагина О.С. О синтаксическом анализе на основе предпочтений / Препринт ИМП им. М.В. Келдыша АН СССР, № 3, М., 1990.
5. Кулагина О.С. Об аспекте меры в лингвистическом знании // Вопросы языкознания — 1991. — № 1. — С. 48–60.

Компьютерная информационная система распознавания динамических образов в потоке изображений на основе алгоритмической технологии

Ц.Г. Литовченко

Рассматривается новая информационная технология, позволяющая выделять и распознавать "зашумленные" объекты в их развитии и динамике. Предлагается схема транспьютерного комплекса для реализации соответствующих процедур.

В последовательном потоке изображений (кадров) присутствует/отсутствует объект, характеристики которого изменяются от кадра к кадру на интервале "жизни" объекта. При отображении объекта в плоскость изображения через аппаратуру наблюдения имеют место обычные для подобных отображений искажения, которые вызваны ограниченной разрешающей способностью, неточностью фокусировки, смазами и другими дефектами. Вместе с объектом в изображение (кадр) попадают окружающие предметы (фон), которые его маскируют. Маскирующее действие фона таково, что в отдельно взятом изображении обнаружение и распознавание искомого объекта либо невозможно, либо возможно с ограниченными характеристиками качества распознавания. Это вызвано тем, что интенсивность излучения искомого объекта, регистрируемая аппаратурой наблюдения в области изображений, близка к интенсивности фоновых объектов. К тому же искомые и фоновые объекты близки по форме и потому плохо различимы.

Однако динамические признаки искомого объекта, которые проявляются в закономерностях изменений на интервале его жизни, могут значительно отличаться от аналогичных признаков фоновых объектов, обладающих меньшей изменчивостью и большей статичностью. Эти признаки могут быть выявлены и использованы для распознавания только в потоке изображений, образующих кинофильм. Отсюда следует, что методы и алгоритмы распознавания образов, имеющих исторические закономерности развития их "жизни", должны опираться на информационную технологию, содержащую средство регистрации и обработки потоков изображений.

Изображение динамического объекта представлено в виде совокупности массивов данных $X_{t_0,T}(t) = \{A_i, \alpha_i, \beta_i, t\}$; $t = t_0, \ldots, t_0 + T$; $i = 1, \ldots, I$, об уровнях яркости излучения A_i, пеленгах α_i, β_i каждой i-ой точки изображения объекта на кадре номера t на интервале "жизни" объекта $t_0, t_0 + T$.