Лингвистический процессор в составе системы распознавания речи

М.Г. Мальковский, И.А. Волкова, П.В. Благовещенский

Существует достаточно много задач обработки текстов на естественном языке, требующих наличия лингвистического процессора (ЛП) — программы, осуществляющей анализ и/или синтез текста и выполняющей некоторые дополнительные действия, диктуемые конкретной задачей. В состав ЛП входит лингвистическая база знаний (различные словари и таблицы), опирающаяся на выбранную формальную модель языка, и исполнительные модули, реализующие отдельные методы (этапы) анализа и/или синтеза текста.

Как правило, в ЛП выделяют модули, работающие с языковыми данными на различных уровнях: морфологическом, синтаксическом, семантическом, а иногда и прагматическом.

В данной работе рассматривается задача получения текстового представления устной английской речи, начиная с момента, когда звуковые сигналы уже преобразованы в фонемы, а те, в свою очередь, сгруппированы в слова. Тем самым ЛП является частью интегрированной системы распознавания речи. Такая система имеет в своем составе несколько компонентов. Задача преобразования звуковых сигналов в слова решается частью системы, которую можно назвать фонетическим распознавателем (ФР). ЛП корректирует результаты работы ФР, исходя из предположения, что произнесено связное высказывание, а не случайный набор слов. Каждое очередное слово фразы, вообще говоря, представляется множеством возможных (по некоторым критериям) словоформ из словаря системы, которые могли бы оказаться текстовой записью "расслышанной" группы фонем.

Таким образом, в нашем случае на вход лингвистического процессора поступает не цепочка слов, а так называемая "гирлянда", где на каждом словоместе указывается некоторый набор словоформ, претендующих занять данную позицию в предложении и определенных ФР с той или иной степенью достоверности, которая задается числовым штрафом, приписанным каждой словоформе в гирлянде. При распознавании слитной речи для каждой произнесенной фразы на вход ЛП могут подаваться гирлянды разной длины, из которых могут быть составлены предложения различной длины и структуры.

Лингвистическая база знаний рассматриваемого лингвистического процессора состоит из грамматического словаря английского языка, словаря словосочетаний, словаря моделей управления и грамматических таблиц.

Для малофлективных языков, каковым является английский язык, удобно

хранить в грамматическом словаре все формы всех слов. Словарная статья содержит необходимую морфосинтаксическую информацию, сведения о частотности и семантических свойствах словоформы.

Морфосинтаксические признаки словоформы включают: индекс грамматического класса (существительное в форме общего падежа, существительное в форме притяжательного падежа, глагол в личной форме, артикль и др.), значения присущих грамматическому классу грамматических переменных (одушевленность существительного, степень сравнения прилагательного, лицо личного местоимения, переходность глагола и др.). Сведения о частотности отражают употребительность словоформы и стилистические ограничения (высокоупотребительное, устарелое, разговорное и т.п.). Семантика словоформы описывается с помощью отсылки к семантическому классу (живое существо, признак-цвет, процесс информационного обмена и т.п.).

Словарь словосочетаний имеет более сложную организацию, т.к. существуют различные виды словосочетаний и фразеологических оборотов.

Учитываются, в частности, возможность разрыва словосочетания (сравнить *and so on* и *cut* [SOMETHING] *out*), степень его идиоматичности. Как и для словоформ, описываются морфосинтаксические, семантические, стилистические и частотные характеристики.

Большинство неразрывных словосочетаний с фиксированным порядком слов могут быть представлены одной неделимой единицей какого-либо грамматического класса. Для их хранения подходит словарь такой же структуры, как и для хранения словоформ. Для разрывных словосочетаний требуются более гибкие способы представления. Так например, описание фразовых глаголов (*cut out, look down on, look for, take off* и др.), которые часто могут быть разрывными словосочетаниями, дается в словаре моделей управления. В нем указывается, какие грамматические конструкции (синтаксические и семантические свойства, длина) и насколько часто могут разрывать фразовый глагол.

Словарь моделей управления в настоящее время содержит информацию только для глагольных форм. В нем указано, сколько и какие актанты (подлежащее, дополнения) могут быть у данного предиката (глагольной формы). Для каждого актанта указываются его конкретный вид (например, прямое дополнение в виде именной группы, инфинитивной группы, THAT-клаузы и т.п.) и способ реализации (именная группа, выраженная герундием, группой существительного или местоимением; инфинитивная группа с частицей *to* или без нее и т.п.), семантические ограничения (подлежащее должно быть представлено именной группой, обозначающей некоторое живое существо и т.п.) и вероятность появления при данном предикате (прямое дополнение должно быть обязательно, косвенное дополнение встречается крайне редко и т.п.).

Синтаксические правила представлены с помощью грамматической таблицы, где в виде сетевой грамматики описаны различные синтаксические конструкции (группы) английского языка. Например, отдельные узлы

ти соответствуют группе существительного, группе количественного числительного, причастным оборотам, сложным глагольным формам. Операторы на дугах сети проверяют синтаксическую и семантическую сочетаемость слов и групп. Так, есть оператор, проверяющий согласование подлежащего и сказуемого (по форме и семантическим классам), неопределенного артикля или указательного местоимения и ключевого слова группы существительного (по числу). Такой же сетью описаны и возможные комбинации синтаксических групп и предложений в составе фразы. Есть отдельные узлы для простого повествовательного предложения, вопросительного предложения, придаточного предложения, оборота типа *there is* и некоторые другие.

Отметим, что особенности объемлющей задачи предполагают решение ряда нетрадиционных проблем.

Во-первых, на вход ЛП подается гирлянда, на каждом словоместе которой может стоять несколько словоформ (в том числе, несколько реальных омофонов: *to — two — too, boys — boy's — boys'*), а каждая словоформа может иметь несколько грамматических значений (*to* — предлог или глагольная частица). В результате число вариантов цепочек, обрабатываемых ЛП, становится непомерно большим (до нескольких миллионов). Поэтому приходится применять быстрые методы анализа и эвристики, сокращающие перебор, например, заранее отбрасывать предложения с заведомо недопустимой сочетаемостью грамматических слов (предлог + глагол, артикль + глагол, артикль + артикль и т.п.) или предложения, суммарный штраф словоформ которых превышает некоторое пороговое значение.

Во-вторых, возможна ситуация, когда из слов гирлянды нельзя составить ни одного синтаксически верного предложения. Причин тут две:

1) На вход системы распознавания могут поступать произвольные речевые фрагменты, например, неполные реплики диалога, заголовки, а также неграмматичные (т.е. не подчиняющиеся используемому формальному описанию) предложения. Так, в одной достаточно популярной американской шутке есть такая фраза *I is what I is.*, в которой нарушены условия согласования личного местоимения и формы глагола *to be*. Неграмматичность такого типа можно назвать естественной.

2) Возможны и ситуации, когда речевые фрагменты, поступающие на вход системы, грамматичны, но из-за ошибок ФР (в позиции, в которой должна стоять третья форма глагола, находятся только существительные в форме множественного числа и т.п.) все получаемые по гирлянде предложения могут оказаться неграмматичными. Допустим, что была произнесена фраза *John loves Mary*, а ФР выдал следующие варианты: для первого словоместа *John*; для второго — *gloves* и *doves*; для третьего — *Mary* и *very*. Очевидно, что из этих слов нельзя составить ни одного синтаксически верного предложения. В этом случае мы имеем дело с так называемой наведенной неграмматичностью.

В-третьих, изменяется сама основная функция лингвистического процессора. Обычно задача ЛП состоит в анализе входной цепочки словоформ и

генерации одного или нескольких вариантов разбора (или же в фиксации неграмматичности предложения). В нашем случае ЛП должен упорядочить допустимые цепочки и выбрать один вариант фразы (который и предъявляется как результат распознавания) из нескольких цепочек-кандидатов, любая из которых может быть неграмматична. Один из возможных механизмов решения этой проблемы — "островной анализ". Под "островом" понимается грамматически содержательный фрагмент предложения (именная группа, глагол-предикат с частично заполненными валентностями и др.).

Если цепочка почти полностью покрывается островами, а набор островов и "разрывов" удовлетворяет определенным эвристическим критериям, то такую цепочку можно признать допустимой и приписать ей соответствующий грамматический штраф.

Решение многокритериальной задачи выбора единственного варианта распознавания осуществляется на основе этих штрафов, а также следующей информации: штрафов, поступивших от ФР; информации о результатах локальных проверок грамматической структуры и семантической сочетаемости слов; статистических данных о лексической сочетаемости слов и частотности отдельных слов; статистических данных о частотности синтаксической структуры островов, предложений или фразы в целом.

Описанные разработки ведутся в рамках договора о сотрудничестве между факультетом вычислительной математики и кибернетики МГУ им. М.В. Ломоносова и фирмой Accent, Inc. (США). В работе принимают участие более десяти специалистов (инженеры знаний, эксперты-филологи, программисты) из России (МГУ, Институт Востоковедения РАН), США и Великобритании.

Реализация ЛП и необходимых инструментальных средств (формирование, тестирование и сопровождение лингвистической базы знаний; планирование и анализ результатов экспериментов) выполнена сотрудниками факультета ВМиК МГУ В.Г. Абрамовым, И.Г. Головиным и В.Н. Пильщиковым. Язык реализации — Visual C++, версия 2.0, операционная среда — Windows NT, версия 3.5.

О разработке процедур автоматического решения задач

А.С. Подколзин

Разработана компьютерная система, имитирующая поведение человека при решении математических задач. Обучение системы проводилось для таких разделов, как элементарная алгебра, тригонометрия и дифференциальное исчисление: при этом был достигнут достаточно высокий процент решаемых задач средней сложности. В основе решателя лежит новый принцип организации процессов решения задач, обеспечивающий практически квазиградиентное его функционирование и позволяющий преодолеть эффект перебора.

Разработка компьютерных решателей задач является важным направлением в математической кибернетике и теории интеллектуальных систем. Такие решатели составляют основу самых различных экспертных и интеллектуальных систем (см., например, [7, 11, 12, 13]), используемых для автоматизации инженерных расчетов в технике, управления сложными технологическими и динамическими процессами, обработки информации при проведении научных исследований, а также систем компьютерного обучения. Создание решателей стимулировало проведение как исследований в конкретных предметных областях (компьютерная алгебра, вычислительная геометрия, дискретная оптимизация и др.) ориентированных на развитие используемого при решении задач математического аппарата, так и исследований математических моделей процесса решения задач в целом [1, 3, 5, 9].

Можно выделить два основных подхода, использовавшихся при разработке систем автоматического решения задач. Первый из этих подходов, связанный с созданием экспертных систем в конкретных предметных областях, основан на применении древовидной классификации задач из рассматриваемой области и накоплении библиотеки алгоритмических процедур, решающих задачи различных типов согласно их классификации. Каждая такая процедура определяет целенаправленный процесс преобразования входных данных ("постановка задачи") при помощи некоторого списка извлеченных из теории рассматриваемой предметной области утверждений ("знаний"), дающих возможность видоизменить описание задачи с уменьшением его сложностных характеристик. Пользователь, работающий с экспертной системой такого рода, по существу находится в программной среде, из которой извлекает те или иные процедуры и выполняет с их помощью очередной шаг обработки информации. Процесс обработки информации при этом имеет ярко выраженный диалоговый характер, и система играет роль