УДК 62.506.2.

М. Ф. БОНДАРЕНКО, канд. техн. наук, *А. И. ЧУГУН*

ЗАДАЧИ ПРИВЕДЕНИЯ ЕДИНИЦ ТЕКСТА ЕСТЕСТВЕННОГО ЯЗЫКА К КАНОНИЧЕСКОМУ ВИДУ И ЕЕ ИСПОЛЬЗОВАНИЕ ПРИ ОБРАБОТКЕ БОЛЬШИХ ИНФОРМАЦИОННЫХ МАССИВОВ

1. Формальным результатом работы информационно-логических систем самого различного назначения являются разнообразные по форме и содержанию тексты, которые формируются в самой системе и поступают на ее выходное устройство в форме, доступной для непосредственного понимания пользователем. Опыт построения и эксплуатации таких систем показал, что наиболее отвечает этим требованиям естественный язык. Использование его выдвигает на первый план требование о машинных методах его переработки. Автоматическая переработка текстовой информации, представленной на естественном языке, включает в себя различные этапы анализа и синтеза текстов, в том числе, грамматическую обработку предложений и отдельных единиц текста, под которой понимаются процессы автоматического анализа и синтеза, приведения словоформ к каноническому виду, идентификации словоформ их лексическим эквивалентам и т. п.

В первых работах по машинному переводу в нашей стране и за рубежом задачи подобного типа решались «насильственным» методом. В память ЭВМ помещали все словоформы (или их основы), необходимые для обработки текстов определенной

тематической направленности, к ним априори приписывалась требуемая информация, а грамматическая обработка входных единиц текста производилась путем их сравнения со словоформами из машинных словарей с помощью сравнительно простых алгоритмов (см., например, [1, 2]). Как показывает примерный расчет [3], при таком подходе требуются ЭВМ с очень большим объемом памяти, к тому же использовать машины второго поколения затруднительно даже для грамматической обработки текстовой информации. Машины третьего поколения с большим объемом памяти хотя и позволяют грамматически обрабатывать единицы текста описанным методом, но не лишены основных его недостатков. Во-первых, фиксирование в машинном словаре определенного количества словоформ резко обнажает основное противоречие между открытым характером естественного языка и закрытым его представлением на алгоритмическом языке. Во-вторых, помещение в память ЭВМ огромных массивов информации и последовательный перебор всех элементов этих массивов на совпадение с обрабатываемой словоформой значительно снижает производительность машины. Из практики машинного перевода известно, что половина времени работы машины тратится на поиск информации в словарях.

Преодолеть эти недостатки можно при помощи универсальных систем грамматической обработки словоформ, которые не были бы привязаны к определенному тематическому словарю и позволили получать необходимую информацию алгоритмически.

2. Независимо от подхода к грамматической обработке словоформ на морфологическом уровне (обработка отдельных словоформ вне контекста), в конечном итоге, предусматривается алгоритмизация трех основных процессов: анализ словоформ для получения информации о их грамматических признаках; синтез различных словоформ слова по одному его представителю (словоформе); приведение словоформ к каноническому виду и идентификация их лексическим эквивалентам. Автоматизация этих процессов требует извлечения максимальной информации из самой словоформы.

ЭВМ не имеет комплексного мышления, которым обладает человек, и для нее признаком конкретной словоформы служит только комбинаторика букв (порядок следования, их количество и т. п.) определенного алфавита. Решая задачи морфологической обработки, оперируют одними и теми же формальными признаками, присущими словоформе. Поэтому при построении независимых алгоритмов для каждой из задач многократно проверяют эти признаки и последовательно обрабатывают однотипные массивы информации.

Исключить эти процессы позволяет комплексный подход к грамматической обработке словоформ на морфологическом уровне. Заключается он в следующем: 1) выявляют для каж-

дой конкретной задачи тот минимально необходимый набор предварительных данных, который приводит к ее однозначному решению; 2) анализируют всю совокупность решаемых задач, устанавливая иерархию их по признаку независимости одной от другой, т. е. очередность решения; 3) начинают построение блока алгоритмов решения морфологической обработки с алгоритма первоочередной и наиболее независимой задачи; 4) учитывают при построении последующих алгоритмов соподчиненность их друг другу и возможность максимального использования информации о словоформе, полученной на предыдущих этапах их обработки.

3. При помощи предлагаемого подхода строится блок алгоритмов для автоматической морфологической обработки спрягаемых форм глаголов русского языка. Первоначальным и основным алгоритмом блока является алгоритм приведения словоформ к каноническому виду, поскольку, во-первых, для решения данной задачи не привлекаются результаты решения задачи синтеза словоформ, во-вторых, решение ориентировано на извлечение максимальной информации из анализа формальных признаков, присущих словоформе. Это практически исключает влияние на точность решения грамматической информации, т. е. результатов решения задачи грамматического анализа. Использование результатов грамматического анализа для этих целей возможно только при установлении всех парадигматических связей слова, что предварительно требует классификации всех словоформ по типам формообразования. А решение второй задачи очень тесно связано с задачей идентификации словоформ их основным лексическим эквивалентам, в качестве которых могут выступать, например, словарные формы или основы слов, в свою очередь, задача идентификации аналогична задаче приведения словоформ к каноническому виду. Наиболее удобной для выявления основного лексического значения глагольной словоформы является форма инфинитива, ее мы и примем за канонический вид.

Детерминистский алгоритм приведения текстовых глагольных словоформ к каноническому виду построен на исследовании комбинаторики нескольких последних букв или (в небольшом числе случаев) комбинаторики всех букв словоформы. Анализ основ и флексий словоформ из множества всех спрягаемых синтетических форм невозвратных глаголов русского языка позволил разбить это множество на пять подмножеств следующего типа: I — содержит словоформы, оканчивающиеся на *-ть*, *-чь*, *-ти*; II — *-ю* (*-у*), *-ют* (*-ут*), *-ать* (*-ять*); III — *-ешь* (*-ишь*), *-ет* (*-ит*), *-ем* (*-им*), *-ете* (*-ите*); IV *-й* (*-йте*), *-ь* (*-и*), *-ьте*; V *-л*, *-ла*, *-ло*, *-ли*. Особую группу составляют словоформы, которые в прошедшем времени имеют особое окончание, например, МЕРЗ, ТЕР и т. п. В процессе решения задачи отнесение словоформы к тому или иному подмножеству производится

алгоритмически. Схему поиска, осуществляемого справа налево, можно представить в виде иерархии, вначале проверяется последняя буква словоформы, на втором этапе предпоследняя буква и т. д. до тех пор, пока словоформа не будет однозначно отнесена к одному из подмножеств.

Каждое подмножество имеет в общем алгоритме отдельную ветку, набор окончаний и особых признаков, на основе которых словоформы этого подмножества приводятся к каноническому виду.

С выхода этого алгоритма словоформа, уже в каноническом виде, направляется на вход алгоритма классификации глаголов в форме инфинитива по типам формообразования, который является вторым основным алгоритмом блока грамматической обработки словоформ на морфологическом уровне. В результате анализа парадигм спряжения глаголов обнаружено 126 типов формообразования. Формальным отличием одного типа от другого является наличие хотя бы одного особого правила образования какой-либо словоформы из парадигмы, состоящей из 13 словоформ (включая инфинитив). Данный алгоритм построен также на исследовании комбинаторики букв словоформ, но только в форме инфинитива. В процессе обработки на этом этапе словоформе приписывается номер типа формообразования и из нее выделяется машинная основа, необходимая для синтеза остальных словоформ из данной парадигмы.

Задачи синтеза словоформ и морфологического анализа сводятся в данном случае к очень простым процедурам. Синтез словоформы осуществляется путем добавления к выделенной из формы инфинитива машинной основе нужного окончания из набора окончаний данного типа формообразования, а анализ — путем сравнения окончания исследуемой словоформы на совпадение с одним из 13 окончаний приписанного ей типа формообразования, хранящегося в таблице, которая состоит из 126 строк и 13 столбцов.

4. Рассмотренный комплексный подход к грамматической обработке словоформ на морфологическом уровне почти полностью освобождает ЭВМ на этом этапе от оперирования громадными массивами информации. Использование небольших массивов в каждом из названных алгоритмов практически не влияет на скорость обработки словоформ. Например, ветка алгоритма приведения словоформ к каноническому виду, обрабатывающая подмассив 1, обращается к массиву словоформ в 55 единиц и небольшому массиву окончаний, а весь алгоритм классификации глаголов в форме инфинитива по типам формообразования содержит массив словоформ в 257 единиц, массив окончаний на 126 единиц и массив приставок на 90 единиц. Алгоритмы анализа и синтеза используют массив эталонных окончаний типов формообразования.

Все алгоритмы и внутренние процедуры записаны на языке

PL/I и отлаживались на ЭВМ ЕС 1020. Скорость обработки одной словоформы в режиме синтеза в среднем составляет одну секунду.

Список литературы: 1. *Кулагина О. С.* О машинном переводе с французского языка на русский. — В кн.: Проблемы кибернетики. Вып. 3. М., 1960, с. 181—208. 2. *J. M. Daniel.* Translation by computer. — «Electronics weekly», 304, 1966, p. 6—11. 3. *Пиотровский Р. Г.* Текст, машина, человек. Л., «Наука», 1975. 327 с.