АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 002

Г. Г. Белоногов, Ю. Г. Зеленков

ЕЩЕ РАЗ О ПРИНЦИПЕ АНАЛОГИИ В МОРФОЛОГИИ

Описывается идея построения алгоритма морфологического анализа текстов на основе принципа аналогии. Алгоритм реализован на ПЭВМ типа IBM PC/AT и в течение ряда лет эксплуатируется в системах орфографического контроля русских текстов, системах автоматического индексирования документов и системах машинного перевода текстов с русского языка на английский и с английского на русский. Производительность программы морфологического анализа на ПЭВМ с процессором типа Intel 386 — более 400 слов/с.

При автоматической обработке текстов возникает проблема «новых» слов. Дело в том, что для синтаксического анализа и синтеза текстов необходимо знать грамматические характеристики всех входящих в них слов. Такие характеристики обычно определяются в результате морфологического анализа с использованием машинных словарей. Но если какое-либо слово текста не содержится в словаре, то не может быть выполнен и его морфологический анализ.

Возможность определения грамматических характеристик слов без словаря интересовала многих ученых, но ее обычно связывали с суффиксами и окончаниями. А этого недостаточно, так как многие формы слов не имеют суффиксов и окончаний (например, *стол, переход, полос, нес, пригоден* и др.).

Более общий подход был предложен Г. Г. Белоноговым [1, 2, 3] и реализован в ряде версий процедур морфологического анализа [3, 4, 5]. Этот подход основан на использовании принципа аналогии, который может быть сформулирован следующим образом. Существует сильная корреляционная связь между грамматическими характеристиками слов и буквенным составом их концов. Поэтому слова, имеющие аналогичные концы, с высокой вероятностью имеют и одинаковые грамматические характеристики. Например, слова *организация, приватизация* и *концентрация* являются существительными женского рода в форме именительного падежа единственного числа; слова *работают, принимают* и *привлекают* — глаголами, имеющими форму третьего лица множественного числа; слова *главных, сильных, серых* — прилагательными в форме родительного или винительного падежа множественного числа, слова *сделанному, выданному, проданному* — страдательными причастиями в форме дательного падежа единственного числа мужского или среднего рода.

Принцип аналогии проверялся на ряде индоевропейских языков (русский, болгарский, латышский, испанский, английский) и оказался весьма эффективным. В процедурах морфологического анализа русских текстов он сначала применялся только для определения грамматических характеристик «новых» слов (слов, не включенных в машинные словарь). При этом использовались таблицы конечных буквосочетаний фиксированной длины [1, 4, 5]. Далее возникла мысль отказаться при морфологическом анализе от машинного словаря слов и построить всю процедуру анализа на принципе аналогии. Этот подход излагается ниже.

Если по текстам достаточно большого объема (например, в несколько миллионов или, еще лучше, в несколько десятков миллионов слов) составить словарь словоформ, назначить каждой словоформе грамматические признаки (например, признаки части речи, типа словоизменения, рода, числа, падежа, лица и др.) и преобразовать полученный таким образом словарь в обратный словарь словоформ, то можно обнаружить, что многие участки словаря (иногда довольно значительного размера) имеют совершенно одинаковые наборы признаков. В этом случае вывод о наличии сильной корреляционной связи между буквенным составом концов словоформ и наборами грамматических признаков, характеризующими эти словоформы, напрашивается сам собой.

В *Приложении 1* представлены фрагменты обратного словаря словоформ, в котором каждой словоформе поставлен в соответствие признак длины грамматического окончания, номер флективного класса (типа словоизменения) и числовой индекс, характеризующий такие признаки, как «глагольность», «местоименность», «сравнительная степень» и др. Признаки грамматического рода, числа и лица в явном виде не указаны (они могут быть легко определены по номеру флективного класса и грамматическому окончанию словоформы).

Обратный словарь может использоваться для автоматического морфологического анализа текстов, если составляющие их словоформы отождествлять со словоформами словаря и приписывать им грамматическую информацию, указанную в словаре. Словоформам текста, которые не находятся в словаре, можно приписывать грамматическую информацию тех словоформ словаря, концы которых в максимальной степени совпадают с концами этих «новых» словоформ текста. Технически это удобно делать, если инвертировать словоформы словаря и словоформы текста перед их поиском в словаре (последние буквы поставить на первые места, предпоследние — на вторые и т. д.). Тогда можно применить один из методов ускоренного поиска (например, метод «деления пополам»).

Объем словаря, представленного в *Приложении 1*, можно существенно сократить, если на всех его участках с одинаковой грамматической информацией оставить только по две словоформы (начальную и конечную), а остальные исключить. Это никак не повлияет на точность морфологического анализа. Более того, можно в каждой паре словоформ с одинаковой грамма-

тической информацией оставить только по одной, например, начальной словоформе, условившись, что если словоформа текста не совпадает ни с одной словоформой обратного словаря, то ей, по окончании дихотомического поиска, приписывается информация непосредственно предшествующей словоформы этого словаря. Фрагменты сокращенного обратного словаря словоформ представлены в *Приложении 2*

Словарь, представленный в *Приложении 2*, можно еще сократить, если исключить из него начальные части словоформ, не оказывающие влияния на результаты морфологического анализа. При этом у каждой пары рядом стоящих словоформ оставляются справа совпадающие конечные буквосочетания и еще по одной букве, которые не совпадают. Результаты такой обработки словаря словоформ представлены в *Приложении 3*.

После выполнения описанных выше операций исходный обратный грамматический словарь словоформ сокращается в восемь раз. Тем не менее, на точность морфологического анализа всех первоначально включенных в него словоформ это не повлияет, а точность анализа всех остальных словоформ русского языка будет достаточно высокой.

Для морфологического анализа текстов на основе метода аналогии достаточно располагать обратным словарем концов слов (см. *Приложение 3*). Однако авторы сочли полезным сформировать еще один машинный словарь — «Словарь служебных и коротких слов». В этот словарь первоначально были включены предлоги, союзы, местоимения, частицы и короткие слова длиной до пяти букв. Однако в дальнейшем в него вошли и другие словоформы, которые по методу аналогии анализировались неправильно. В результате «словарь служебных и коротких слов» увеличился до 11 тыс. словоформ. В процессе морфологического анализа текстов словоформы сначала ищутся в словаре «Служебных и коротких слов», а затем — в словаре концов словоформ. Результаты анализа, полученные в процессе поиска по первому словарю, считаются более надежными, и словоформы, найденные в этом словаре, последующей обработке не подвергаются.

Из предыдущих рассуждений следует, что точность работы алгоритма морфологического анализа, построенного на основе принципа аналогии, может повышаться в процессе его эксплуатации. В настоящее время вероятность правильного анализа слов при обработке текстов любой тематики превышает 99%.

В *Приложении 4* приведены результаты морфологического анализа небольшого фрагмента текста. Слева по вертикали расположены слова исходного текста. За ними следуют двузначные индексы длин грамматических окончаний слов, затем через косую черту — номера их флективных классов, далее — двузначные дополнительные грамматические признаки основ слов («местоименность», «глагольность», «сравнительная степень» и др.) и, наконец, — наборы цифровых индексов, обозначающих грамматический род, число, падеж и лицо. Например, словоформа *компьютерная* имеет грамматическое окончание, состоящее из двух букв, флективный класс 103, двузначный признак «собственно прилагательное» (в отличие, например, от местоименных и отглагольных прилагательных) и двузначный признак «женский род, именительный падеж единственного числа», а словоформа *область* — окончание, состоящее из одной буквы, флективный класс 055, признак «собственно существительное» и набор из двух двузначных признаков: «именительный падеж единственного числа» и «винительный падеж единственного числа».

Рассмотренная система морфологического анализа реализована на ПЭВМ типа IBM PC/AT и работает со скоростью более 400 слов/с. Она используется в различных системах автоматической обработки текстов (обнаружение и исправление орфографических ошибок, автоматическое индексирование, машинный перевод текстов с русского языка на английский и с английского на русский, автоматизированное составление словарей различного назначения и др.). В разработке системы морфологического анализа наряду с авторами статьи принимали участие научные сотрудники Отдела лингвистических исследований ВИНИТИ А. П. Новоселов, Е. Ю. Рыжова, С. А. Самоделкина, Ал-др А. Хорошилов, Ал-сей А. Хорошилов и Е. Г. Дружинина.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г. Г., Давыдова И. М. О возможности определения грамматических классов по буквенным кодам слов // НТИ. Сер. 2.— 1967.— № 8.
2. Белоногов Г. Г. Об использовании принципа аналогии при автоматической обработке текстовой информации // Проблемы кибернетики.— 1974.— № 28.
3. Белоногов Г. Г., Новоселов А. П., Губарь Н. Т. Морфологический анализ слов на основе словаря словоформ // НТИ. Сер. 2.— 1975.— № 9.
4. Белоногов Г. Г., Зеленков Ю. Г. Алгоритм морфологического анализа русских слов // Вопр информ. теории и практики.— 1985.— № 53.
5. Белоногов Г. Г., Загика Е. А., Новоселов А. П. Автоматизация лингвистической обработки словарей в системе научно-технической информации // Вопр. кибернетики. Прикладные аспекты лингвистической теории / Под ред. акад. А. Л. Ершова.— М.. ВИНИТИ, 1987.

ПРИЛОЖЕНИЕ 1

Фрагменты обратного грамматического словаря словоформ
(словарь содержит около 170 тыс. лексических единиц)

масштаба 01/001/01
хлеба 01/001/01
амеба 01/044/01
неба 01/071/01
погреба 01/001/01
небоскреба 01/001/01
техучеба 01/056/01
служба 01/056/01
разведслужба 01/056/01
телеслужба 01/056/01
гидрометеослужба 01/056/01
дружба 01/056/01
изба 01/056/01
перелиба 01/001/01

противоизгиба 01/001/01
самоизгиба 01/001/01
ушиба 01/001/01
шайба 01/056/01
полба 01/056/01
столба 01/001/01
дамба 01/056/01
лимба 01/001/01
бомба 01/056/01
клумба 01/056/01
желоба 01/001/01
пневможелоба 01/001/01
виброжелоба 01/001/01
аэрожелоба 01/001/01

полужелоба 01/001/01
короба 01/001/01
полукороба 01/001/01
телепроба 01/056/01
иммунопроба 01/056/01
ферментоиммунопроба 01/056/01
цветопроба 01/056/01
строба 01/001/01
способа 01/001/01
хозспособа 01/001/01
арба 01/056/01
верба 01/056/01
ущерба 01/001/01
губа 01/056/01

дуба 01/001/01
куба 01/001/01
луба 01/001/01
палуба 01/056/01
труба 01/056/01
ахтуба 01/056/01
шуба 01/056/01
ходьба 01/056/01
судьба 01/056/01
резьба 01/056/01
мельба 01/056/01
стрельба 01/056/01
эльба 01/056/01
борьба 01/056/01
снегоборьба 01/056/01
косьба 01/056/01
пастьба 01/056/01
. . . . . .
отсутствовавшие 02/105/10
свидетельствовавшие 02/105/10
преследовавшие 02/105/10
прореагировавшие 02/105/10
контролировавшие 02/105/10
регистрировавшие 02/105/10
продемонстрировавшие 02/105/10
вызывавшие 02/105/10
выпадавшие 02/105/10
поедавшие 02/105/10
опоздавшие 02/105/10
продолжавшие 02/105/10
многорожавшие 02/105/10
возникавшие 02/105/10
принимавшие 02/105/10
выпавшие 02/105/10
работавшие 02/105/10
обработавшие 02/105/10
. . . . . .
медальон 00/001/01
батальон 00/001/01
почтальон 00/021/01
павильон 00/001/01
кинопавильон 00/001/01
шатильон 00/001/01
бульон 00/001/01
каньон 00/001/01
виньон 00/001/01
авиньон 00/001/01
реюньон 00/001/01
гардсьон 00/001/01
асунсьон 00/001/01
лосьон 00/001/01
берн 00/001/01
дерн 00/001/01
серн 00/044/01
цистерн 00/056/01
автомолцистерн 00/056/01
автоцистерн 00/056/01
. . . . . .
анонсирует 02/116/10
массирует 02/116/10
прогрессирует 02/116/10
вегетирует 02/116/10
интерпретирует 02/116/10
имитирует 02/116/10
лимитирует 02/116/10
эмитирует 02/116/10
эжектирует 02/116/10
спроектирует 02/116/10
гарантирует 02/116/10
ориентирует 02/116/10
дезориентирует 02/116/10
документирует 02/116/10
шунтирует 02/116/10
зашунтирует 02/116/10
экспортирует 02/116/10
контрастирует 02/116/10
диагностирует 02/116/10
компостирует 02/116/10
. . . . . .
протестует 02/116/10
лафет 00/001/01
пинцет 00/001/01
фальцет 00/001/01
зачет 00/001/01
влечет 02/120/10
привлечет 02/120/10
повлечет 02/120/10
течет 02/120/10
истечет 02/120/10
вытечет 02/120/10
счет 00/001/10
расчет 00/001/10
отчет 00/001/10
учет 00/001/10
. . . . . .
влияет 02/116/10
повлияет 02/116/10
ослабляет 02/116/10
потребляет 02/116/10
избавляет 02/116/10
возглавляет 02/116/10
сплавляет 02/116/10
направляет 02/116/10
исправляет 02/116/10
заставляет 02/116/10
представляет 02/116/10
оставляет 02/116/10
составляет 02/116/10
выставляет 02/116/10
удешевляет 02/116/10
подготовляет 02/116/10
осуществляет 02/116/10
проявляет 02/116/10
объявляет 02/116/10
предъявляет 02/116/10
. . . . . .
соединены 01/126/10
подсоединены 01/126/10
объединены 01/126/10
подчинены 01/126/10
пополнены 01/126/10
выполнены 01/126/10
осолонены 01/126/10
захоронены 01/126/10
прояснены 01/126/10
упрочнены 01/126/10
уточнены 01/126/10
присвоены 01/126/10
освоены 01/126/10
построены 01/126/10
выстроены 01/126/10
удостоены 01/126/10
пены 01/056/01
арены 01/056/01
наварены 01/126/10
сварены 01/126/10
спарены 01/126/10
оребрены 01/126/10
одобрены 01/126/10
удобрены 01/126/10
. . . . . .
рассчитывая 00/152/10
учитывая 00/152/10
перелистывая 00/152/10
испытывая 00/152/10
тонкоклювая 02/103/01
черноклювая 02/103/01
толстоклювая 02/103/01
полагая 00/152/10
предполагая 00/152/10
располагая 00/152/10
избегая 00/152/10
прибегая 00/152/10
облегая 00/152/10
прилегая 00/152/10
пренебрегая 00/152/10
передвигая 00/152/10
раздвигая 00/152/10
продвигая 00/152/10
отодвигая 00/152/10
сдвигая 00/152/10
выдвигая 00/152/10
отжигая 00/152/10
достигая 00/152/10
. . . . . .
извлекая 00/152/10
сопоставляя 00/152/10
составляя 00/152/10
изготовляя 00/152/10
разветвляя 00/152/10
осуществляя 00/152/10
проявляя 00/152/10
выявляя 00/152/10
замедляя 00/152/10
определяя 00/152/10
распределяя 00/152/10
перераспределяя 00/152/10
разделяя 00/152/10
отделяя 00/152/10
уделяя 00/152/10
выделяя 00/152/10
позволяя 00/152/10
закрепляя 00/152/10
прикрепляя 00/152/10
скрепляя 00/152/10
. . . . . .

ПРИЛОЖЕНИЕ 2

Фрагменты сокращенного обратного словаря словоформ
(объем словаря — около 28 тыс. лексических единиц)

масштаба 01/001/01
амеба 01/044/01
неба 01/071/01
погреба 01/001/01
техучеба 01/056/01
перегиба 01/001/01
шайба 01/056/01
столба 01/001/01
дамба 01/056/01
лимба 01/001/01
бомба 01/056/01
желоба 01/001/01
телепроба 01/056/01
строба 01/001/01
арба 01/056/01
ущерба 01/001/01

губа 01/056/01
дуба 01/001/01
палуба 01/056/01
. . . . . .
отсутствовавшие 02/105/10
. . . . . .
медальон 00/001/01

почтальон 00/021/01
серн 00/044/01
цистерн 00/056/01
. . . . . .
анонсирует 02/116/10
. . . . . .
протестует 02/116/10

лафет 00/001/01
влечет 02/120/10
счет 00/001/10
. . . . . .
влияет 02/116/10
. . . . . .
соединены 01/126/10

пены 01/056/01
наварены 01/126/10
. . . . . .
рассчитывая 00/152/10
. . . . . .
извлекая 00/152/10
. . . . . .

ПРИЛОЖЕНИЕ 3

Фрагменты обратного словаря концов словоформ
(объем словаря — около 28 тыс. концов слов)

аба 01/001/01
еба 01/044/01
неба 01/071/01
реба 01/001/01
чеба 01/056/01
иба 01/001/01
йба 01/056/01
лба 01/001/01
мба 01/056/01
имба 01/001/01
омба 01/056/01

лоба 01/001/01
роба 01/056/01
троба 01/001/01
арба 01/056/01
. . . . . .
ерба 01/001/01
губа 01/056/01
. . . . . .
дуба 01/001/01
луба 01/056/01
. . . . . .
авшие 02/105/10
. . . . . .
дальон 00/001/01

тальон 00/021/01
серн 00/044/01
терн 00/056/01
. . . . . .
ирует 02/116/10
. . . . . .
тует 02/116/10
фет 00/001/01
ечет 02/120/10
счет 00/001/10
. . . . . .

ияет 02/116/10
. . . . . .
нены 01/126/10
пены 01/056/01
рены 01/126/10
. . . . . .
тывая 00/152/10
. . . . . .
влекая 00/152/10
. . . . . .

ПРИЛОЖЕНИЕ 4

Результаты морфологического анализа

Компьютерная 02/103/01/31
лингвистика 01/060/01/11
—
это 01/112/02/2124
область 01/055/01/1114
знаний 01/073/01/22
,
связанная 02/103/10/31
с 00/162/025
решением 02/073/01/15
задач 00/057/01/22
автоматической 02/106/01/32333536
обработки 01/060/01/122124
информации 01/061/01/1213162124
,
представленной 02/103/01/32333536
на 00/164/046
естественном 02/103/01/1626
языке 01/006/01/16
.

Центральными 03/103/01/45
научными 03/103/01/45
проблемами 03/056/01/25
компьютерной 02/103/01/32333536
лингвистики 01/060/01/122124
являются 04/116/00/6/*
проблема 01/056/01/11
моделирования 01/073/01/122124
процесса 01/001/01/12
понимания 01/073/01/122124
смысла 01/001/01/12
текстов 02/001/01/22
(
перехода 01/001/01/12
от 00/155/02
текста 01/001/01/12
к 00/156/03
формализованному 03/103/10/1323
представлению 01/073/01/13

его 00/145/02
смысла /01/001/01/12
)
и 00/153/01
проблема 01/056/01/11
синтеза 01/001/01/12
речи 01/054/01/1213162124
(
перехода 01/001/01/12
от 00/155/02
формализованного 03/103/10/121422
представления 01/073/01/122124
смысла 01/001/01/12
к 00/156/03
текстам 02/001/01/23
на 00/164/046
естественном 02/103/01/1626
языке 01/006/01/16
)

Материал поступил в редакцию 23.02.95.