АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1985 год Выпуск 113

УДК 414:0.093

МАШИННАЯ ОБРАБОТКА СЛОВАРНЫХ ДАННЫХ КАК СПОСОБ ИСПОЛЬЗОВАНИЯ ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ ПРИ АВТОМАТИЧЕСКОМ РАСПОЗНАВАНИИ РЕЧИ

С.К.Егоров, М.Д.Люблинская, Т.В.Шарыгина

Современная постановка проблемы автоматического распознавания речи, трактуемая как понимание речи, характеризуется стремлением к использованию лингвистической информации различных уровней. Это в свою очередь требует тщательного изучения фонетических, грамматических и семантических свойств речи, которые должны способствовать порождению оптимальных правил распознавания. Алгоритмы распознавания слов могут быть более эффективными, если они используют знания о фонемно-морфологической структуре слов. В частности, базой для статистических исследований фонетико-морфологической структуры слова могут служить словари различного типа.

I. Работы по реализации машинных версий словарей, в особенности русского языка, приобретают в настоящее время важное значение [1]. На филологическом факультете Ленинградского университета ведутся работы с машинными версиями Деривационного словаря русского языка [2] и Грамматического словаря [3] (в дальнейшем RDD и ГС)*). Перенесение словарей на магнитную ленту (МЛ) дало колоссальное увеличение скорости работы, что в свою очередь позволяет проводить исследования и получать информацию на различных

*) Работа по перенесению RDD на МЛ и по получению статистики по отдельным параметрам была проведена в ИМ СО АН СССР в Новосибирске. В дальнейшей работе с этой версией мы внесли в нее некоторые изменения, в основном коснувшиеся определения частей речи. Приводимые в статье данные частично пересекаются с результатами, полученными в ИМ. Подготовка и создание машинной версии ГС было проведено в НИВЦ МГУ.

подмножествах словарей в пределах реальных сроков. Например, для получения списка суффиксов в словах словаря требуется полчаса машинного времени, и это время может быть значительно сокращено.

На материале указанных словарей было интересно получить следующие данные, полезные при распознавании речи:

1) количественное распределение слов по частям речи;

2) статистика распределения префиксов по позициям в слове, корреляция появления префиксов в разных частях речи;

3) такая же информация о суффиксах;

4) частота встречаемости экспонентов (единиц, различных в плане выражения) на материале словаря;

5) встречаемость бифонемных сочетаний согласных и их дистрибуция внутри и на границах морфем разного типа (корень, префиксы, суффиксы).

Данные по п.1-4 были получены на материале RDD, по п.5 - на материале ГС. Использование этих сведений позволит сузить границы поиска "правильных" последовательностей фонем и морфем.

2. Недостатки RDD отмечались неоднократно (см.например, [4, 5]), и в задачи нашей статьи не входит просто увеличить перечень критических замечаний. ЭВМ позволяет достаточно творчески обращаться с материалом словаря, корректировать и дополнять его в зависимости от целей исследования.

При перенесении RDD на МЛ каждое слово было снабжено пометкой о его частеречной принадлежности. Слово может иметь более одного частеречного маркера,например,"посреди" - наречие и предлог. Таким образом мы получили следующие классы: существительное, глагол,прилагательное,субстантивированное прилагательное, причастие, наречие, деепричастие, числительное, включающее количественные и порядковые, союз, предлог, междометие, частица.

Мы стремились работать со словами-цепочками "от пробела до пробела". Поэтому все единицы словаря, имеющие более одной составляющей, были отнесены в группу с общим названием "фразеологизмы", противопоставленную всему остальному словарю по указанному формальному признаку. Распределение лексики по частям речи представлено в табл.1.

3. При получении статистики по аффиксам слова, имеющие более одного корня, не рассматривались (это обусловлено специфичностью набора и представления такой лексики в словаре, например, "погрузочно-разгрузочный" или "контрольно-товароведческий"; сложные сло-

Таблица I

Часть речи	Количество	Процент
существительное	43736	42,79
глагол	27435	26,84
прилагательное	22984	22,49
субстантивированное прилагательное	98	0,10
причастие	5010	4,90
наречие	1938	1,89
деепричастие	117	0,11
местоимение	92	0,09
числительное	101	0,10
союз	88	0,09
предлог	75	0,07
междометие	130	0,13
частица	102	0,10
фразеологизмы	246	0,24

ва представляют собой отдельную тему исследования). Мы получили таблицу (табл.2) распределения префиксов по частям речи и ввели "коэффициент аффиксированности" части речи - отношение числа аффиксов в словах данной части речи к общему количеству слов этой части речи (появление междометий с префиксами объясняется особенностью представления их в словаре).

Таблица 2

Часть речи	Количество	Процент	Коэффициент аффиксиров.
существительное	10221	27,42	0,23
глагол	15026	40,30	0,55
прилагательное	6741	18,08	0,29
субстантивированное прилагательное	9	0,02	0,09
причастие	4075	10,93	0,81
наречие	1001	2,68	0,52
деепричастие	96	0,26	0,82
местоимение	25	0,07	0,27
числительное	1	0,00	0,00
союз	18	0,05	0,20
предлог	23	0,06	0,31
междометие	11	0,03	0,08
частица	15	0,04	0,13

Специальные программы в табличной форме представили список префиксов с указанием первого слова в словаре, где встретился этот префикс, и его частоты. Число префиксов в слове не превосходило 4. Самыми частотными оказались за-, по- и на-, которые составляли соответственно 8,4%, 7,6% и 6,0% от общего количества. Треть всего списка составили "уникальные", т.е. встретившиеся только I раз, экспоненты. И только 33 префикса (II% списка) встретились более I00 раз. Следует заметить, что ударные и безударные варианты различались.

4. Действия над суффиксами были почти аналогичны действиям над префиксами. В общем длина суффиксальной цепочки не превышала 7. Исключениями оказались два слова: бол-ь-ш-ев-из-ир-ов-а-нн=ый и субстанц-и-он-ал-из-ир-ов-а-нн=ый. Длина полученного списка - 700 единиц (ударные и безударные алломорфы, как и у префиксов, различались). Были получены данные о частоте встречаемости суффиксов, самыми частотными оказались -а́-, -а- и -ов-. Уникальные единицы занимают треть списка; I05 суффиксов (I5%) встретилось более I00 раз.

5. Вообще, словарь заставил нас посмотреть на некоторые слова с новой, неожиданной точки зрения. Приведем соответствующие примеры. Так, списки префиксов и суффиксов содержали единицы, не знакомые русскому читателю - префиксы: м- в слове по-м-бух (подчеркнут корень), г- в г-айда, жа- в жа-ворон-ок, вож- в вож-дел-ени-е; суффиксы: -рел- в аква-рел-ист, -бан- в чур-бан, -чих- в крол-ь-чих-а и др. В словах, чаще иностранного происхождения, не всегда мотивированным оказывалось членение и отнесение составляющих к тому или иному классу морфем. Локализация слов, где впервые появился аффикс, позволила "выловить" случаи типа: пара-докс, бис-сект-р-ис-а, рапс-од, авто-медон. Задачи корректировки словаря потребовали проверки списка корней, выявив еще ряд неточностей, например, включение слова "гросс-бух" в гнездо корня галтер, по аналогии с "глав-бух".

6. При получении сведений о консонантных сочетаниях слова, содержавшие внутри или на границе морфем "невозможные" пары согласных, образовали список исключений. Он позволил уточнить как перечень "невозможных" сочетаний, так и принципы морфемного членения, проведенного в машинной версии ГС. Что касается положительных результатов, представленных в виде таблицы сочетаний согласных для каждого типа морфем и их границ, то можно сделать общий вывод: в

принципе, для каждого типа морфем, за исключением корневых, и их стыков характерен весьма ограниченный набор частотных сочетаний. Например, внутри префиксов более 100 раз на материале 130000 слов словаря встретилось всего 6 сочетаний: гн, вн, нс, жн, рн, рс. Сами сочетания, естественно, делятся на частотные, нечастотные и "почти невероятные". Частотные могут быть универсальными или привязанными к определенному типу морфемы или границы. Например, цепочка пр/пр' внутри префикса встречается в 6 раз чаще, чем в любом другом месте словоформы.

7. Хотелось бы отметить взаимосвязь программ, с помощью которых сканировался текст словаря. Например, отделение флексий в RDD, основанное на частеречной принадлежности слова, было включено в программу анализа суффиксов, и это помогло выявить спорные случаи частеречной разметки. Просмотр списка корней позволил выявить случаи слияния корневых и аффиксальных морфем, например, бенедиктин-, элефантиаз-.

Полученные результаты являются предварительными и будут уточнены после проведения корректуры. Созданное программное обеспечение позволяет решать аналогичные задачи на материале обоих словарей, хотя отсутствие единой базы данных при работе с разными словарями заметно усложняет обработку. Накопленный опыт работы со словарями показал, что морфолого-семантическую проверку желательно осуществлять параллельно с подготовкой словаря на магнитных носителях. Это позволит гораздо быстрее покончить с трудоемким этапом подготовки данных.

Результаты обработки словарей регулярно обсуждались на семинаре при кафедре фонетики ЛГУ, всем членам семинара авторы выражают искреннюю благодарность. Анализ списка корней был проведен Т.Алексеевой в дипломной работе.

Литература

1. АНДРЮЩЕНКО В.М. Машинный фонд русского языка: постановка задачи и практические шаги. -Вопросы языкознания, 1985, №2, с.54-64.

2. WORTH D.S., KOZAK A.S., JOHNSON D.B. Russian Derivational Dictionary. - New York, 1970.

3. ЗАЛИЗНЯК А.А. Грамматический словарь русского языка. -М.: Русский язык. 1977. -879 с.

4. ТИХОНОВ А.Н. Проблемы составления гнездового словообразовательного словаря русского языка. -Самарканд, 1971.

5. ЕФРЕМОВА Т.Ф. Рецензия на работу: D.S.Worth, A.S.Kozak, D.B.Johnson. Russian Derivational Dictionary.– Вопросы языкознания, 1971, №4.

Поступила в ред.-изд.отд.
30 октября 1985 года