**Принцип аналогии в морфологии**

В работе Белоногова Г. Г. и Зеленкова Ю. Г. описывается принцип построения алгоритма морфологического анализа текстов на основе принципа аналогии. Данный принцип используется в системах орфографического контроля русских текстов, системах автоматического индексирования документов и системах машинного перевода текстов с русского языка на английский и с английского языка на русский. Производительность программы на компьютере с процессором от 386 и выше составляет около 400 слов/с.

При автоматической обработке текста возникает проблема “новых“ слов. Для синтаксического анализа и синтеза необходимо знать грамматические характеристики слов. Если слова в словаре нет, то морфологический анализ не может быть выполнен, а следовательно не могут быть определены грамматические характеристики слова.

Для того, чтобы определить грамматические характеристики слов без словаря, Белоногов предложил принцип аналогии. Он основан на том, что существует сильная корреляционная связь между грамматическими характеристиками слов и буквенным составом их концов. Например: организация, приватизация, концентрация имеют ж. р., им. п. и ед. ч.; работают, понимают, привлекают - это глаголы в 3-ем лице мн. ч. и т. д.

Принцип аналогии проверялся на ряде индоевропейских языков: (русский, болгарский, латышский, испанский, английский) и оказался эффективным. Сначала он применялся для определения грамматических характеристик слов, не включенных в машинный словарь. Затем возникла идея при проведении морфологического анализа отказаться от машинного словаря.

Если по текстам большого объема составить словарь словоформ и назначить каждой словоформе некоторые грамматические признаки, а затем преобразовать данный словарь в обратный словарь словоформ, то можно обнаружить, что многие участки словаря имеют одинаковые наборы признаков.

Обратный словарь словоформ представляет собой список словоформ с такими характеристиками как признак длинны грамматического окончания, номер флективного класса (типа словоизменения) и числовой индекс, характеризующий такие признаки как “глагольность“, “местоименность”, “сравнительная степень”. Например:

масштаба 01/001/01

служба 01/056/01

возникшие 02/105/10

батальон 00/021/01

рассчитывая 00/152/10

Обратный словарь используется 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.

Объем обратного словаря можно сократить, если на всех его участках оставить по две словоформы: начальную и конечную. Более того из этих двух словоформ можно оставить только одну, и если словоформа текста не совпадет ни с одной словоформой обратного словаря, то ей приписывается информация непосредственно предшествующей словоформы этого словаря.

Данный сокращенный словарь можно еще сократить, если исключить из него начальные буквы словоформ, не оказывающие влияние 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. Например:

аба 01/001/01

еба 01/044/01

неба 01/071/01

авшие 02/105/10

тальон 00/021/01

тывая 00/152/10

После выполнения всех операций объем словаря сокращается в 8 раз. На точность первоначально включенных в словарь словоформ это не повлияет, а точность анализа остальных словоформ русского языка будет достаточно высокой.

Для морфологического анализа текстов на основе метода аналогии достаточно располагать обратным словарем концов слов. Но авторы разработки сделали еще “Словарь служебных и коротких слов”. В этот словарь были включены сначала предлоги, местоимения, частицы, союзы и короткие слова до 5 букв. Затем в него вошли также словоформы, которые по методу аналогии анализировались неверно. В результате этот словарь увеличился до 11 тысяч словоформ.

Таким образом, в процессе морфологического анализа словоформы ищутся в словаре “Служебных и коротких слов”, а затем в словаре концов словоформ. Результаты анализа, полученные по первому словарю, считаются более надежными, и словоформы, найденные в этом словаре, дальнейшей обработке не подвергаются.

В настоящее время вероятность правильного анализа слов при обработке текстов любой тематики превышает 99%.

В разработке данной системы наряду с авторами данной 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, Е. Г. Дружинина.